technologies

IMPACT FACTOR 5.2

CITESCORE 6.7

Article

Weekly Cyanobacterial Bloom-Risk Early Warning Using Deep LSTM and Hybrid CNN-LSTM Models in Jinyang Lake, South Korea

Sangbin Ha, Muhammad Waqas and Sang Min Kim

MDPI

https://doi.org/10.3390/technologies14110725

technologies

MDPI

Article

Weekly Cyanobacterial Bloom-Risk Early Warning Using Deep LSTM and Hybrid CNN-LSTM Models in Jinyang Lake, South Korea

Sangbin Ha †, Muhammad Waqas † and Sang Min Kim *

Department of Agricultural Engineering, Institute of Agriculture and Life Sciences, Gyeongsang National University, Jinju 52828, Republic of Korea; 2020012322@gnu.ac.kr (S.H.); muhammad.waqas@gnu.ac.kr (M.W.)
* Correspondence: smkim@gnu.ac.kr
† These authors contributed equally to this work.

Abstract

Cyanobacterial blooms threaten the water supply in regulated reservoirs, but the weekly early-warning skill and the usefulness of basin-scale water-quality gradients are not-well quantified. This study created a weekly bloom-risk framework for Jinyang Lake, South Korea, based on the probability of cyanobacteria exceeding 1000 cells/mL in the next 1–4 weeks. Weekly data from 16 stations (2011–2025) trained Deep Long Short-Term Memory (LSTM) and hybrid Convolutional Neural Network–Long Short-Term Memory (CNN-LSTM) classifiers (training 2011–2022, tuning 2023, independent testing 2024–2025; 48 weeks). The results of principal component analysis (PCA) indicated that the first two components accounted for 46.9–88.1% of the variance by variable, so no single station represented basin-wide conditions. At the 4-week horizon, the hybrid CNN-LSTM achieved F1 = 0.849 (95% CI: 0.74–0.93) and critical success index (CSI) = 0.738 at Naedong, and F1 = 0.758 (0.62–0.88) and CSI = 0.610 at Panmun, exceeding Random Forest (F1 = 0.58–0.62) and XGBoost (F1 = 0.55–0.58). At Naedong, the area under the curve (AUC) was 0.523, while, at Panmun, it was 0.768, due to the class-weighted training that resulted in probabilities being squeezed into a narrow range. PCA-derived features numerically increased the F1 score at Panmun by 0.020–0.041 (not statistically significant) and reduced the Naedong hybrid F1 score from 0.849 to 0.786. Platt scaling reduced hybrid-model Brier scores by 27–29% at both lake stations, improving the usability of the probabilities for operational decisions. The multi-week bloom-risk framework offers reservoir operators and municipal drinking-water systems probabilistic early-warning decision support for preventive action before critical bloom development.

Keywords: harmful algal blooms (HABs); reservoir water management; drinking water protection; recurrent sequence learning; convolutional temporal architectures; dimensionality reduction; environmental hazard forecasting

Check for updates

Academic Editor: Miklas Scholz

Received: 3 September 2026
Revised: 1 November 2026
Accepted: 2 November 2026
Published: 4 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Surface water bodies (freshwater lakes, regulated rivers, and multipurpose reservoirs) are the main source of drinking water in the world and in South Korea [1]. Cyanobacterial blooms are one of the most pervasive water quality issues in lakes and reservoirs worldwide, driven by a combination of nutrient enrichment, thermal conditions, slow flushing, biological adaptation, and hydrologic memory [2–4]. They can produce taste-and-odor

problems, pose a cyanotoxin hazard, diminish recreational value, complicate drinking-water treatment, and impose ecological stress [5–7]. This is especially the case in regulated reservoirs where inflow, outflow, storage, water level, residence time, tributary pulses, and dam operation interact with meteorological forcing [3,8]. Bloom growth is favored during warm periods, and high nutrient levels together with stable water columns can further enhance cyanobacterial dominance [2,9]. These mechanisms have been extensively described in the literature on harmful algal blooms (HABs), particularly in the context of eutrophication and warming as co-drivers of HAB frequency and intensity [2,4,6]. From an operational reservoir management point of view, it is important to note that there is often a high numerical uncertainty in the forecast of cell density. Thus, the early warning problem becomes an event window classification problem, where an operational horizon of several weeks is used to predict whether cyanobacterial densities will reach an actionable management threshold (e.g., 1000 cells/mL) during the horizon.

Summer cyanobacterial blooms are especially prevalent in the reservoirs and rivers of South Korea, and have been the subject of research in the Nakdong, Han, and Geum River basins [10–12]. The construction of dams and weirs on the Nakdong River has been associated with an increased risk of blooms, which has spurred the development of extreme learning machine, random forest (RF), gradient boosting, and deep learning (DL) models to predict chlorophyll-a and cyanobacteria at Korean stations [10,11,13,14]. The reservoir under consideration, Jinyang Lake, was also part of a national study on eight reservoirs to investigate the use of machine-learning classifiers for bloom classification, which included RF and gradient-boosting classifiers [11], which emphasizes the ongoing operational need for sequence architectures that model the temporal and spatial dependence of water quality beyond same-week tabular predictors alone. Jinyang Lake, created by the Namgang Dam, is a key multipurpose lake that provides the main drinking-water supply for more than one million people in the region, including Jinju and Sacheon [15]. The reservoir is under strict water quality regulations because of its critical municipal function. Early warning is critical in order to reduce consumer exposure to offensive taste and odor compounds [16,17].

The Jinyang Lake system integrates water-quality data from various upstream tributaries and nearby dam stations. The bloom risk at lake stations (Panmun and Naedong) is influenced by the local conditions as well as the antecedent conditions of the upstream network. A model based solely on local, same-week observations cannot capture the delayed propagation of nutrients, organic matter, flow, and reservoir-storage conditions [8,12]. The risk of bloom may be affected by events that took place several weeks prior, and such delayed effects may be masked by averaging data monthly. Therefore, a weekly target design is more suitable for early warning than a monthly concentration-regression target, which is consistent with the recommendations for higher-resolution HAB forecasting [11,14].

Numerous bloom studies have tried to forecast the chlorophyll-a or cyanobacteria concentration directly [10–12]. While the ability to predict the concentration is appealing, it can be misleading when the target distribution is zero-inflated, highly skewed and dominated by rare peaks [18]. The operational warning value [14,18] is then dominated by extreme events and traditional regression statistics like R^2 and RMSE may not apply. A model can give poor estimates of the peak magnitude but still provide useful information on whether a bloom window is imminent, whereas a model with a good regression fit can still fail to provide management-relevant warning events. This is important for reservoir management because it is not the peak concentration that is important, but whether a warning threshold will be breached soon.

The current study presents the modeling task as an early warning of cyanobacterial blooms at the weekly time scale instead of concentration regression. The endpoint is the probability that cyanobacteria will exceed 1000 cells/mL in the next 1, 2, 3 or 4 weeks [19]. This formulation directly maps the model output to a warning probability, and allows for an evaluation based on event-based metrics including the F1 score, contingency counts, recall, precision, and critical success index (CSI) [11,20], which explicitly quantify hits, misses, and false alarms. The area under the receiver-operating-characteristic curve (AUC) is also reported, but is considered as a rank-discrimination statistic, not the only measure of usefulness.

This framing inspires the central research question addressed here: can weekly, probabilistic classifiers of the bloom-window risk, informed by spatially distributed gradients of upstream and dam water quality, provide a more operationally defensible early-warning skill for Jinyang Lake than exact concentration forecasting, and does explicitly engineering spatial PCA information into the predictor enhance that skill? Bloom risk is an inherently temporally dependent process, making DL architectures well-suited to the task. Long short-term memory (LSTM) networks learn long-range dependencies in sequential data using gates that regulate information retention and forgetting [21–23], which is useful because nutrient loading, hydrologic residence, temperature persistence, and biological memory can influence bloom risk across multiple weeks [12]. However, recurrent models may not be optimal if short local patterns such as the rapid rise in nutrient or chlorophyll signals are important. The principle of convolutional filters learning the local structure, and recurrent layers learning sequence dependence [12,24,25] is followed by a hybrid convolutional neural network (CNN) and LSTM architecture [26], where the local temporal motifs are first extracted from the data and then passed to the recurrent memory.

Spatial interpretation is also needed because of the distributed water-quality conditions in Panmun and Naedong. The principal component analysis (PCA) method was used to summarize station-level trends in water-quality parameters in a compact manner. PCA is not a causal model, but it is a way to identify the dominant spatial loading patterns and to separate the stations that act differently across the variables [27,28]. To address these challenges, this study pursued five objectives: (1) describing spatial water-quality gradients across physicochemical, nutrient (TN, NO_3-N, TP), organic matter (COD, TOC, BOD), and algal biomass (Chlorophyll-a, DO, SS) indicators using PCA to identify coherent loading structures across the upstream–dam network; (2) formulating weekly early warning as a 1-to-4-week event-window classification; (3) evaluating Deep LSTM and hybrid CNN-LSTM sequence architectures against classical benchmarks (Random Forest, XGBoost); (4) quantifying metric uncertainty via 2000 bootstrap resamples on the independent 2024–2025 test period while resolving the technical divergence between threshold-based skill (F1, CSI) and rank discrimination (AUC); and (5) critically evaluating, via ablation, whether delayed PCA-derived spatial-state features improve the operational warning skill beyond baseline lagged predictors.

2. Materials and Methods

2.1. Study Area

Jinyang Lake is the first multipurpose artificial lake formed by the construction of the Namgang Dam in the Nakdong River basin, South Korea [1] (Figure 1). The Namgang Dam basin covers an area of 2281.72 km^2 with a mean elevation of 427.6 m and average slope of 41.77%, which is mountainous and steep terrain, whereas the downstream Namgang basin covers an area of 1185.08 km^2 with a mean elevation of 111.87 m and average slope of 35.44% [29–31]. The multipurpose Namgang Dam has a total storage capacity of 309.2 million m^3 and flood-control capacity of 269.8 million m^3, and provides about

573 million m^3 of water annually, along with hydropower generation and river management functions. Damming extends the residence time, which can lead to the accumulation of nutrients and eutrophication and the formation of cyanobacterial blooms in the reservoir, as has been observed in other regulated reservoirs in Korea where the hydrological residence time, inflow/outflow patterns, water level, and seasonal temperature control the dynamics of the cyanobacterial blooms [10,11].

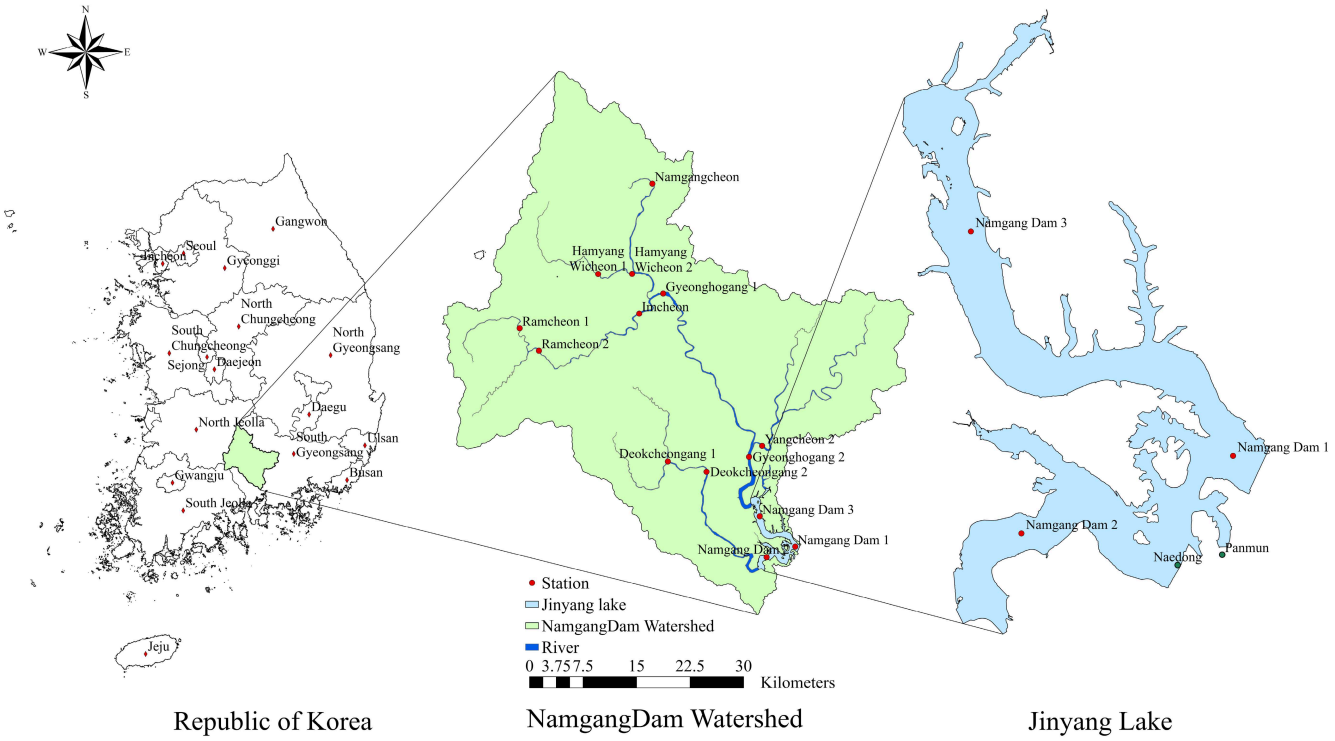


Figure 1. Study area map of Jinyang Lake, Namgang Dam, and the upstream tributary and dam monitoring-station network, South Korea.

The study incorporates 16 monitoring stations (14 upstream and dam stations, Namgangcheon, Hamyang Wicheon, Ramcheon, Imcheon, Gyeonghogang, Deokcheongang, and Yangcheon; and 2 lake stations, Panmun, and Naedong) for spatial analysis of the water-quality gradient and the response to blooms from upstream to downstream.

2.2. Data Collection and Processing

The dataset for water-quality observations upstream and at the Namgang Dam, and for cyanobacteria observations at Panmun and Naedong, are presented in Table 1. The water-quality variables measured were water depth, biochemical oxygen demand (BOD), chlorophyll-a, chemical oxygen demand (COD), dissolved oxygen (DO), NH_3-N, NO_3-N, PO_4-P, pH, water temperature, suspended solids (SS), total nitrogen (TN), total organic carbon (TOC), total phosphorus (TP), and abundance of cyanobacteria in cells/mL, conducted in accordance with standard water-quality monitoring procedures. The dataset was resampled from daily to weekly resolution for modeling purposes to facilitate early warning. A binary label was provided for each forecast horizon to show if cyanobacteria were present at levels above 1000 cells/mL in the following 1–4 weeks. Lagged predictors were created to account for the memory of antecedent water-quality and travel-time.

Table 1. Data sources and processed inputs.

Station	Long (°E)	Lat (°N)	Basin	Chl-a (mg/m^3)	TP (mg/L)	SS (mg/L)	COD (mg O_2/L)	BOD (mg O_2/L)	PO_4-P (mg/L)
				Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD
Namgangcheon	127.81	35.63	Namgang Dam	2.49 ± 2.18	0.03 ± 0.04	3.73 ± 10.18	2.60 ± 0.99	0.93 ± 0.56	0.01 ± 0.01
Imcheon	127.79	35.46		3.89 ± 3.45	0.03 ± 0.03	5.65 ± 14.34	3.38 ± 1.31	0.84 ± 0.45	0.01 ± 0.01
Yangcheon 2	127.98	35.30		5.08 ± 4.81	0.03 ± 0.02	5.39 ± 7.12	4.38 ± 1.47	1.46 ± 0.73	0.01 ± 0.01
Hamyang Wicheon 1	127.73	35.51		4.91 ± 7.37	0.04 ± 0.03	6.72 ± 11.88	3.13 ± 1.33	1.18 ± 0.76	0.01 ± 0.01
Hamyang Wicheon 2	127.78	35.51		7.59 ± 14.77	0.07 ± 0.07	7.84 ± 12.72	3.92 ± 1.73	1.49 ± 0.97	0.03 ± 0.05
Ramcheon 1	127.61	35.45		4.57 ± 4.49	0.07 ± 0.06	6.22 ± 7.61	4.66 ± 1.91	1.48 ± 0.87	0.02 ± 0.03
Ramcheon 2	127.64	35.42		2.73 ± 2.69	0.04 ± 0.03	3.31 ± 5.33	3.27 ± 1.28	1.06 ± 0.72	0.01 ± 0.02
Gyeonghogang 1	127.83	35.49		6.93 ± 9.61	0.04 ± 0.04	6.33 ± 21.99	3.90 ± 2.16	1.11 ± 0.65	0.02 ± 0.02
Gyeonghogang 2	127.96	35.28		6.64 ± 7.10	0.03 ± 0.03	7.60 ± 19.60	4.26 ± 1.63	1.27 ± 0.65	0.01 ± 0.01
Deokcheongang 1	127.84	35.28		1.43 ± 1.69	0.02 ± 0.02	1.94 ± 4.63	1.48 ± 0.59	0.75 ± 0.51	0.00 ± 0.02
Deokcheongang 2	127.90	35.26		1.12 ± 0.98	0.02 ± 0.02	2.30 ± 6.80	1.67 ± 0.89	0.77 ± 0.49	0.00 ± 0.00
Namgang Dam 1	128.03	35.17	Jinyang Lake	4.34 ± 3.05	0.02 ± 0.01	3.35 ± 3.93	3.01 ± 0.90	1.21 ± 0.47	0.01 ± 0.01
Namgang Dam 2	127.99	35.15		4.14 ± 2.88	0.02 ± 0.01	3.21 ± 3.29	2.80 ± 0.82	1.16 ± 0.44	0.00 ± 0.01
Namgang Dam 3	127.98	35.21		5.00 ± 3.93	0.02 ± 0.02	4.36 ± 10.47	3.16 ± 1.04	1.26 ± 0.53	0.01 ± 0.01

The robust scaling was fitted during the training period only, and missing predictors were imputed by the medians from the training period. The training, validation and cutoff tuning were performed with the chronological split of 2011–2022, 2023, and 2024–2025, respectively, which was recommended for hydrologic and water-quality model evaluation [18]. A summary of monitoring stations is presented in Table 1 (mean ± SD; the Panmun and Naedong lake-station statistics are given in Table S1). All sampling and analysis were conducted in accordance with Korean Standard Methods for the Examination of Water Pollution from the National Institute of Environmental Research, Ministry of Environment.

2.3. Model Development, Equations, and Hyperparameter Tuning

The weekly input sequence was defined as a matrix X_t of size $L \times p$, where L is the lookback window (weeks) and p are the number of selected predictors. Features were ranked with an ExtraTrees classifier [32] trained only on the training set, and the top-ranked predictors were retained inside the hyperparameter grid. The future bloom label was defined as follows:

$$y(t,h) = \begin{cases} 1 \text{if } \max[C(t+1), \ldots, C(t+h)] \geq \ 1000 \text{ cells/mL} \\ 0 \ \text{ otherwise} \end{cases} \quad (1)$$

where C is cyanobacteria concentration (cells/mL) and h is the 1–4 week forecast horizon. Formally, for a given forecast lead time of h weeks ($h \in \{1, 2, 3, 4\}$), the target label $y(t,h) \in (0, 1)$ is defined over the forward event window: $y(t,h) = 1$ if $max_{(1 \leq k \leq h)\, C(t+k)} \geq 1000$ cells/mL, and 0 otherwise. Thus, the model predicts the occurrence of a bloom event at any point within the forthcoming h weeks, rather than a single discrete point forecast exactly h weeks ahead.

The Deep LSTM model used two LSTM layers followed by dropout and dense GELU layers, governed by the standard gated-recurrent equations [21,22]:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \text{ — input gate} \quad (2)$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \text{ — forget gate} \quad (3)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \text{ — output gate} \quad (4)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c x_t + U_c h_{t-1} + b_c) \text{ — cell state update} \quad (5)$$

$$h_t = o_t \odot \tanh(c_t) \text{ — hidden state} \quad (6)$$

where i_t, f_t, and o_t are the input, forget, and output gate activations, c_t is the cell state, h_t is the hidden state, σ is the logistic sigmoid function, W and U are trainable weight matrices, b are bias vectors, and $\odot$ denotes the elementwise product. The final hidden state produced a logit z_t, and the warning probability was as follows:

$$p_t = \sigma(z_t) \quad (7)$$

This structure follows the LSTM memory principle. The hybrid CNN-LSTM first applied a one-dimensional convolution across the temporal sequence, with kernel size 3:

$$q_t = \text{GELU}(\sum_j W_j x_{t-j} + b) \quad (8)$$

The convolutional representation q_t was then passed to an LSTM (Equations (2)–(6)) and dense classification head (Equation (7)). This architecture was intended to learn short antecedent pulses before recurrent memory integration, following the broader representation-learning logic of convolutional and recurrent deep networks [12,24,26].

Models were trained using weighted binary cross-entropy, with class weight w equal to the ratio of negative to positive training examples:

$$L = -\frac{1}{N}\sum_{n=1}^{N}[w \cdot y_n \log(p_n) + (1 - y_n)\log(1 - p_n)] \quad (9)$$

This study used gradient clipping, dropout regularization [33] and early stopping with AdamW optimization [34]. Hyperparameter tuning evaluated a structured factorial grid, with lookback window $L \in \{4, 8, 12\}$ weeks (3 levels), top feature screening count $\in \{15, 20, 25\}$ (3 levels), hidden units/convolution filters $\in \{32, 48, 64\}$ (3 levels), dropout rate $\in \{0.1, 0.2\}$ (2 levels), learning rate $\in \{5 \times 10^{-4}, 1 \times 10^{-3}\}$ (2 levels), and batch size $\in \{16, 32\}$ (2 levels), yielding exactly $3 \times 3 \times 3 \times 2 \times 2 \times 2 = 216$ parameter sets per station-architecture combination (Table S2).

The fixed training parameters were as follows: 100 epochs, early-stopping patience of 15 epochs and weight decay of 1×10^{-4}. Models were trained using weekly observations from 2011 to 2022 and hyperparameter selection and the operational probability cutoff were determined solely on the validation period (2023), which was held out chronologically from the training set, based on combined F1, CSI, and recall score (Table S3). To serve as baseline comparisons, classical ML algorithms were trained on the same feature sets and time splits: Random Forest (RF) and Extreme Gradient Boosting (XGBoost).

All structural modeling choices, such as the optimal 4-week forecast horizon, hyperparameters, and probability cutoffs, were made solely based on the pre-test 2023 validation dataset prior to any evaluation on the independent test period (2024–2025; 48 monitoring weeks). Validation performance showed that the 4-week horizon had the highest F1 and CSI at all stations and architectures, while the skill was lower at 1-to-3-week horizons (F1 = 0.091–0.667). Therefore, the 4-week horizon was selected a priori from the validation data. Independent test data (2024–2025) were evaluated only once in an unadjusted, forward-testing protocol, so the reported test skill is not driven by test-set tuning. Because the 2023 validation window contains only 27 weeks, with one to four event weeks per horizon, architecture selection is inherently noisy, and architecture preference should be regarded as exploratory. Raw model outputs were post hoc calibrated by Platt scaling (logistic calibration) and assessed using the Brier score and expected calibration error (ECE) (Table S6).

To ensure full computational reproducibility for all experiments, a deterministic random seed of 42 was set for all runtime environments: Python random (seed = 42), NumPy (seed = 42), PyTorch CPU and CUDA (torch.manual_seed(42), torch.cuda.manual_seed_all(42)), and PyTorch deterministic algorithms (torch.backends.cudnn.deterministic = True). Seed 42 was also used for bootstrap confidence interval resampling (2000 iterations). All the deep-learning and statistical pipelines were implemented in Python (v3.12.7) with PyTorch (v2.4.1), scikit-learn (v1.5.2), pandas (v2.2.3), NumPy (v1.26.4), and SciPy (v1.14.1).

PCA was performed on standardized station-parameter matrices such as

$$Z = TP^T + E \quad (10)$$

where Z is the standardized station-by-parameter data matrix, T is the score matrix, P is the loading matrix, and E is residual variation not captured by the retained components [27]. Most importantly, PCA is a descriptive dimension reduction method that provides spatial covariance and tributary gradient patterns among monitoring sites. It is not a direct mechanistic hydrodynamic transport or causal biochemical kinetics, and loading patterns should be viewed as empirical spatial associations, not deterministic physical processes.

For each water-quality parameter, the proportion of total spatial variance captured by the retained components was computed as follows:

$$\text{Variance explained}(\%) = 100 \times \frac{\lambda_k}{\sum \lambda_i} \tag{11}$$

where λ_k is the eigenvalue of the k-th principal component and the summation is taken over all retained components [27]; the combined PC1+PC2 percentages reported in Table 2 were obtained by summing this quantity across the first two components.

Table 2. PCA variance explained by PC1 and PC2.

Variable	PC1 (%)	PC2 (%)	PC1+PC2 (%)
COD	42.79	16.98	59.77
Chlorophyll-a	26.19	20.69	46.88
DO	79.44	8.67	88.11
NO_3-N	45.18	18.58	63.77
TN	46.29	17.14	63.43
TOC	47.37	16.93	64.30
TP	34.20	17.82	52.02

Independent-test classification skill was quantified with standard contingency-table metrics [18,20]:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{12}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{13}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

$$\text{CSI} = \frac{TP}{TP + FP + FN} \tag{15}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{17}$$

where TP, FP, TN, and FN denote true positive, false positive, true negative, and false negative warning-week counts, respectively. The AUC was computed as the trapezoidal-rule integral of the ROC curve across all candidate probability cutoffs [20]. Feature importance was interpreted qualitatively in relation to the ExtraTrees ranking used for feature selection [28,32,35].

PCA was used not only as a parallel descriptive analysis but also as part of the predictive feature design. For each water-quality parameter, the first two monthly principal-component scores from Equation (10) were merged with each weekly modeling record by year–month and transformed into delayed spatial-state predictors using lags of 0, 1, 2, 4, and 8 weeks (Figure 2).

These PCA-derived variables were passed through the same ExtraTrees feature-screening stage as the original lagged water-quality and hydrologic predictors, and were allowed to compete freely with those predictors for inclusion in the final hyperparameter grid. A focused ablation was then conducted at the selected 4-week warning horizon by comparing the baseline feature set with a PCA-integrated feature set for each station and deep-learning architecture. Baseline and PCA-integrated results are presented for all stations and architectures; the ablation is a descriptive test-set comparison that was not used

to select the final configuration, and therefore does not rely on the unsupported assumption that PCA-derived features necessarily improve prediction [11,14].

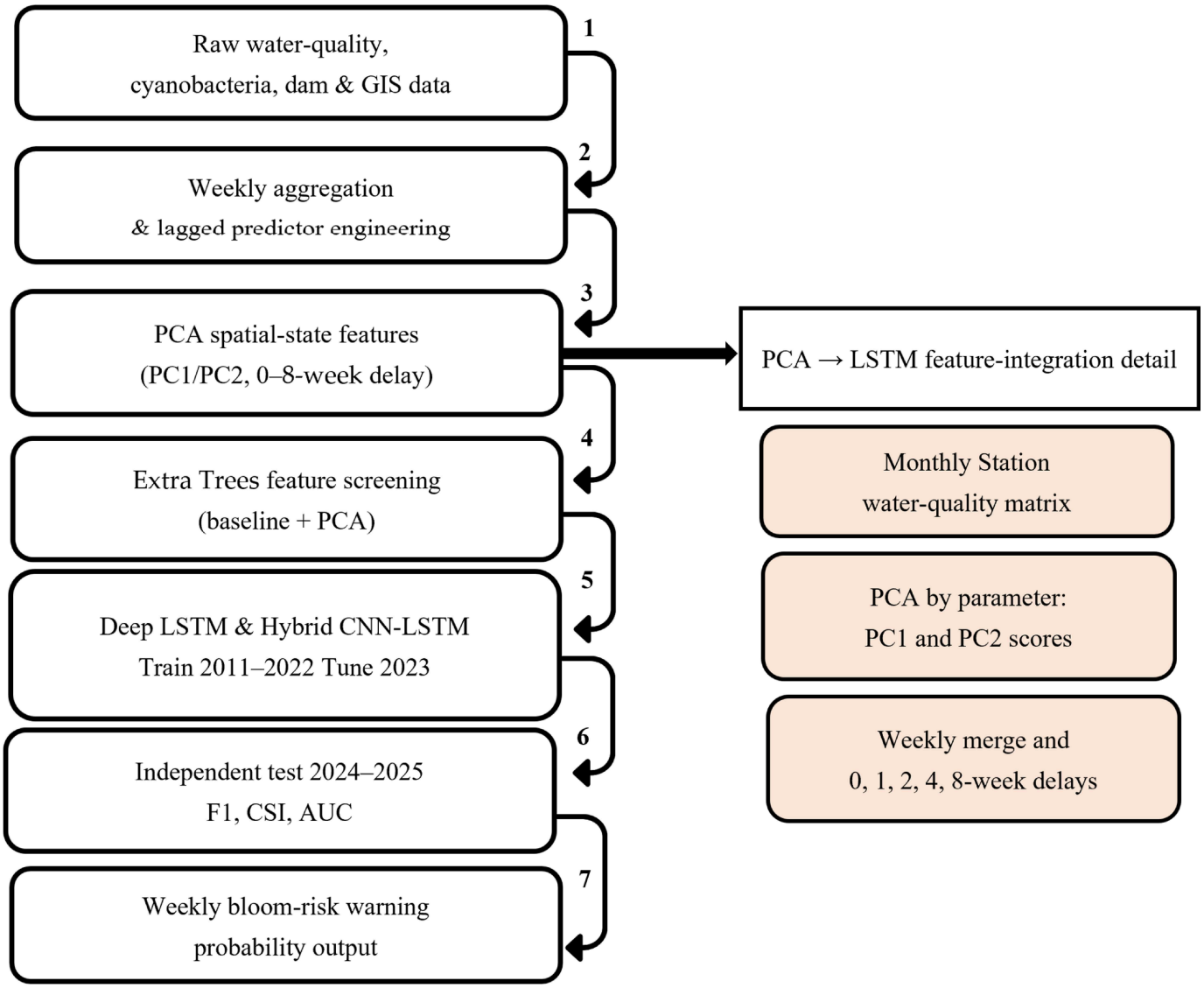


Figure 2. Modeling workflow diagram. PCA describes spatial loading gradients and supplies engineered PC-score features; Extra Trees screens predictors; Deep LSTM and hybrid CNN-LSTM models are trained (2011–2022), tuned (2023), and independently assessed (2024–2025).

3. Results

3.1. Spatial Water-Quality Gradients from Principal Component Analysis

The spatial PCA results showed that water quality in Jinyang Lake is not spatially homogeneous along the monitoring network from upstream to the dam. The first two components accounted for 46.88% of the chlorophyll-a variance at the station level (Table 2) and the higher-order components contained a significant amount of residual variance in algal biomass potential. This indicates that the distribution of chlorophyll-a at the stations is a result of several independent processes, including differential nutrient loading, hydraulic residence time, and local biological production, rather than a single dominant spatial axis.

COD, on the other hand, had a more dominant first component and PC1 alone accounted for 42.79% of the variance, increasing to 59.77% with the addition of PC2, indicating a more coherent, basin-wide organic-matter gradient than chlorophyll-a. The overall struc-

ture of DO was the best of all the variables examined, with PC1 and PC2 accounting for 88.11% of the variance, suggesting that oxygen dynamics are dominated by a few key spatial factors, most likely a gradient from turbulent, reaerated waters in the upstream tributaries to more stratified, biologically productive waters in the dam-zone. TN, TP, NO_3-N, and TOC fell somewhere between the other water-quality parameters, and the combined PC1+PC2 variance ranged from 52.02% to 64.30% (Table 2), indicating that each water-quality parameter highlighted a different portion of the upstream–dam network.

This heterogeneity is ecologically relevant because there is no single predictor of cyanobacteria control, and the risk of cyanobacteria blooms depends on the combination of nutrient availability, organic-matter loading, oxygen dynamics, hydrologic retention, and biological response to the season [3]. This variability among water-quality parameters in the amount of variance explained means that each water-quality parameter should be used as a separate spatial descriptor, and not be assumed to be representative of basin-wide conditions.

These spatial gradients are visualized directly by the combined PCA loading map (Figure 3), which shows the chlorophyll-a (a), COD (b), DO (c), NO_3-N (d), TN (e), TOC (f) and TP (g) together, with the magnitude of loading on PC1 used to scale the size of the station marker and marker color denoting the sign of the PC2 loading. This visualization enables a direct comparison of which stations are the most dominant along the primary spatial axis for each variable and which stations are not in accordance with the basin-wide pattern along the secondary spatial axis.

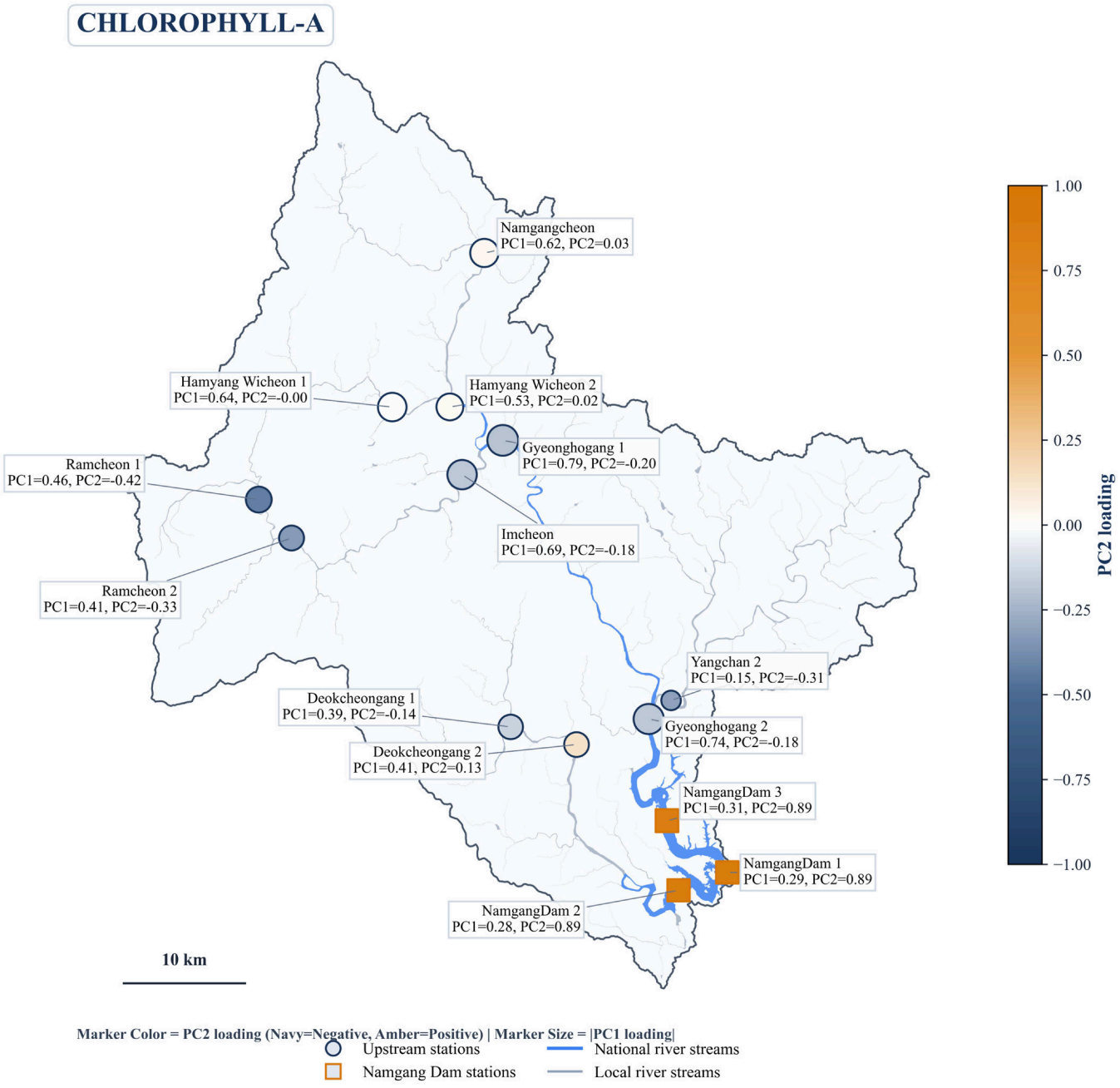
Figure 3a shows a spatial pattern in algal biomass potential between upstream and dam stations: stations with large PC1 values are those with a strong contribution to the main chlorophyll gradient, while those with strongly positive or negative PC2 values are stations with a secondary behavior that differs from the basin-wide behavior, possibly representing localized retention zones or point-source influences. This map is analytically useful because chlorophyll-a is used as a proxy for phytoplankton biomass and often correlates with cyanobacterial bloom risk, but is not the same as the cyanobacteria abundance target used for the early-warning classification of the map, as the latter only applies to the cyanobacteria taxa of concern.

The PCA analysis confirms the overall modeling rationale that the algal biomass signal is not just contained at the lake stations, and that the use of upstream and dam-zone predictors that provide independent, spatially distributed information are appropriate for the Panmun and Naedong stations.

Figure 3b showed strong and coherent PC1 loading compared to most nutrient variables, which is in line with the single-component variance of 42.79% reported in Table 2. The individual top-loading stations for each parameter's first principal component are ranked in Table S4 (Supplementary Materials), which lists Namgang Dam (PC1 loading = 0.47) and Gyeonghogang (PC1 loading = 0.33) as the two dominant COD stations, consistent with the coherent basin-wide organic-matter gradient described here. COD is a measure of oxidizable organic matter and can be from inputs to the watershed (terrestrial organic carbon, agricultural runoff, and domestic wastewater) or from in-reservoir biological production (algal exudates and decomposing biomass).

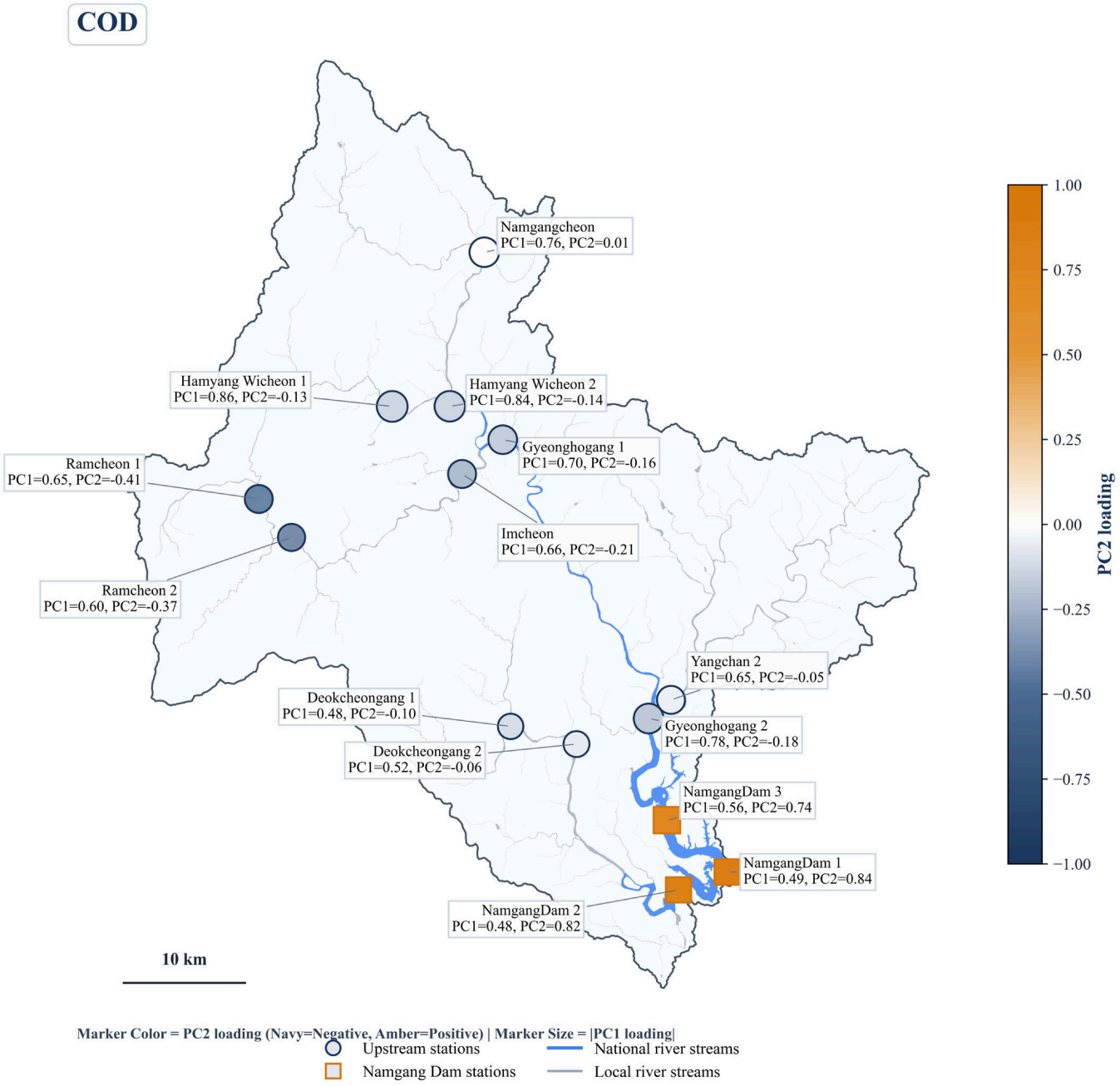
The small range of stations with high PC1 loadings along the COD gradient indicates that the transport of organic matter through the basin is more predictable and unidirectional than the transport of nutrients, which are more likely to be transformed into the basin. The high PC1 loading in the COD panel for these stations also indicates their strong contribution to the dominant organic-matter gradient, and their PC2 values suggest whether their secondary behavior differs from this basin-wide direction, which is directly applicable during the predictor-engineering stage as these same delayed COD-derived PC scores were

among the most commonly selected PCA-derived features when screening in ExtraTrees (Figure S2), thereby establishing a practical connection between the descriptive spatial analysis and the predictive early-warning models. These observed spatial loading patterns are not necessarily indicative of actual advective pathways, but are simply statistical covariance structures across the tributary network, and hydrodynamic transport mechanisms are still working hypotheses that need to be validated in a three-dimensional hydrodynamic context.



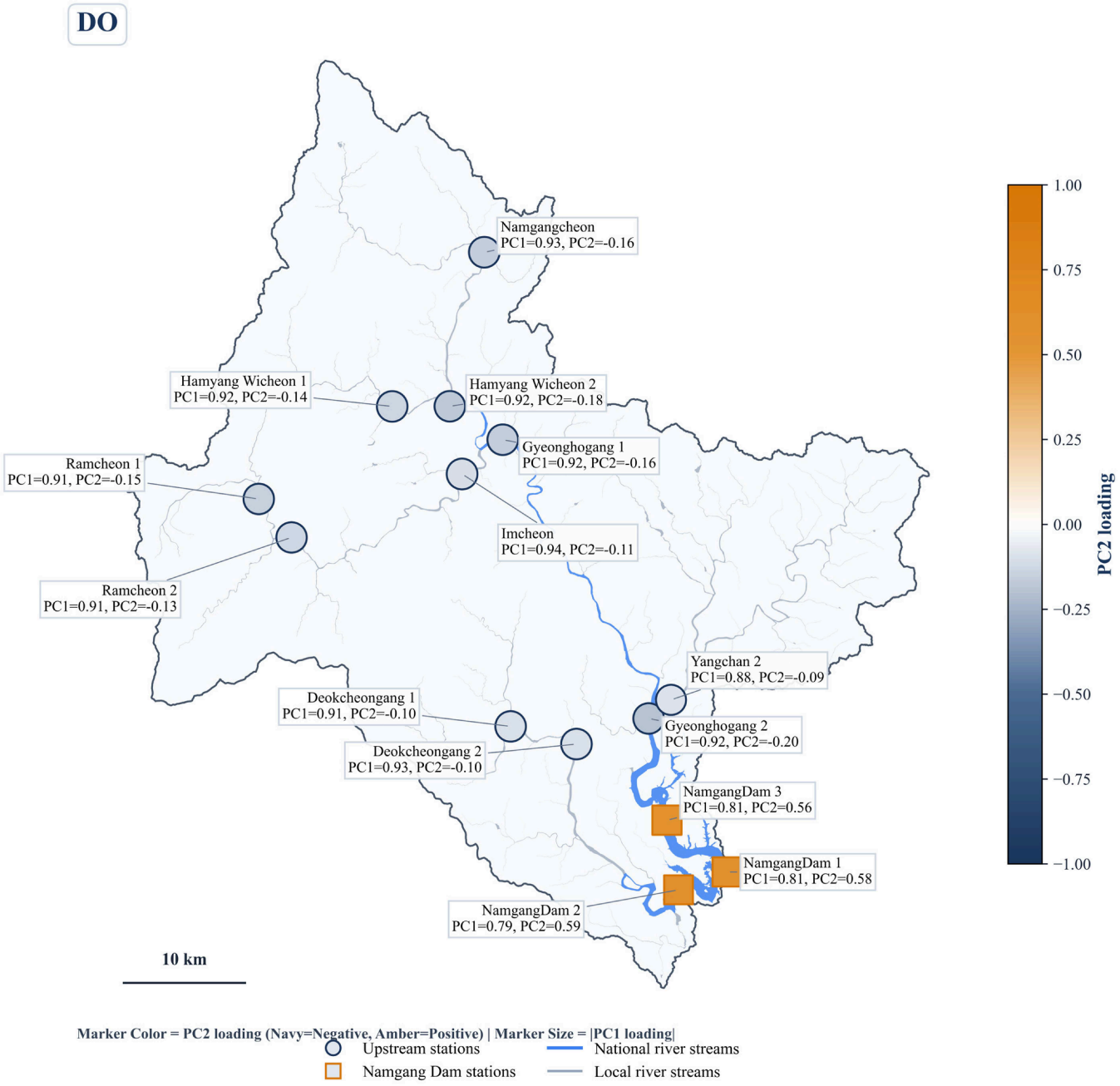
(**a**)

Figure 3. *Cont.*



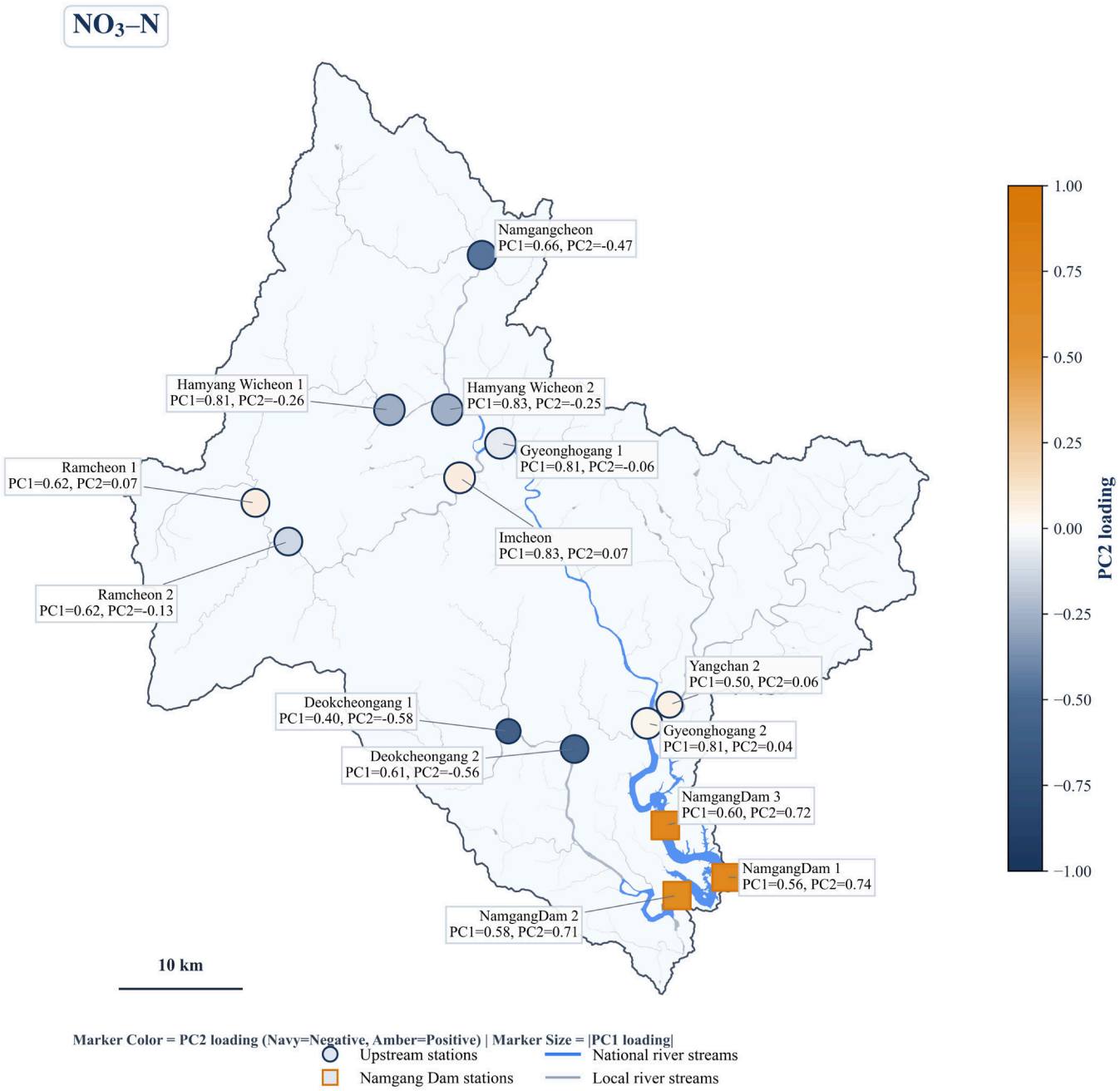
(**b**)

Figure 3. *Cont*.



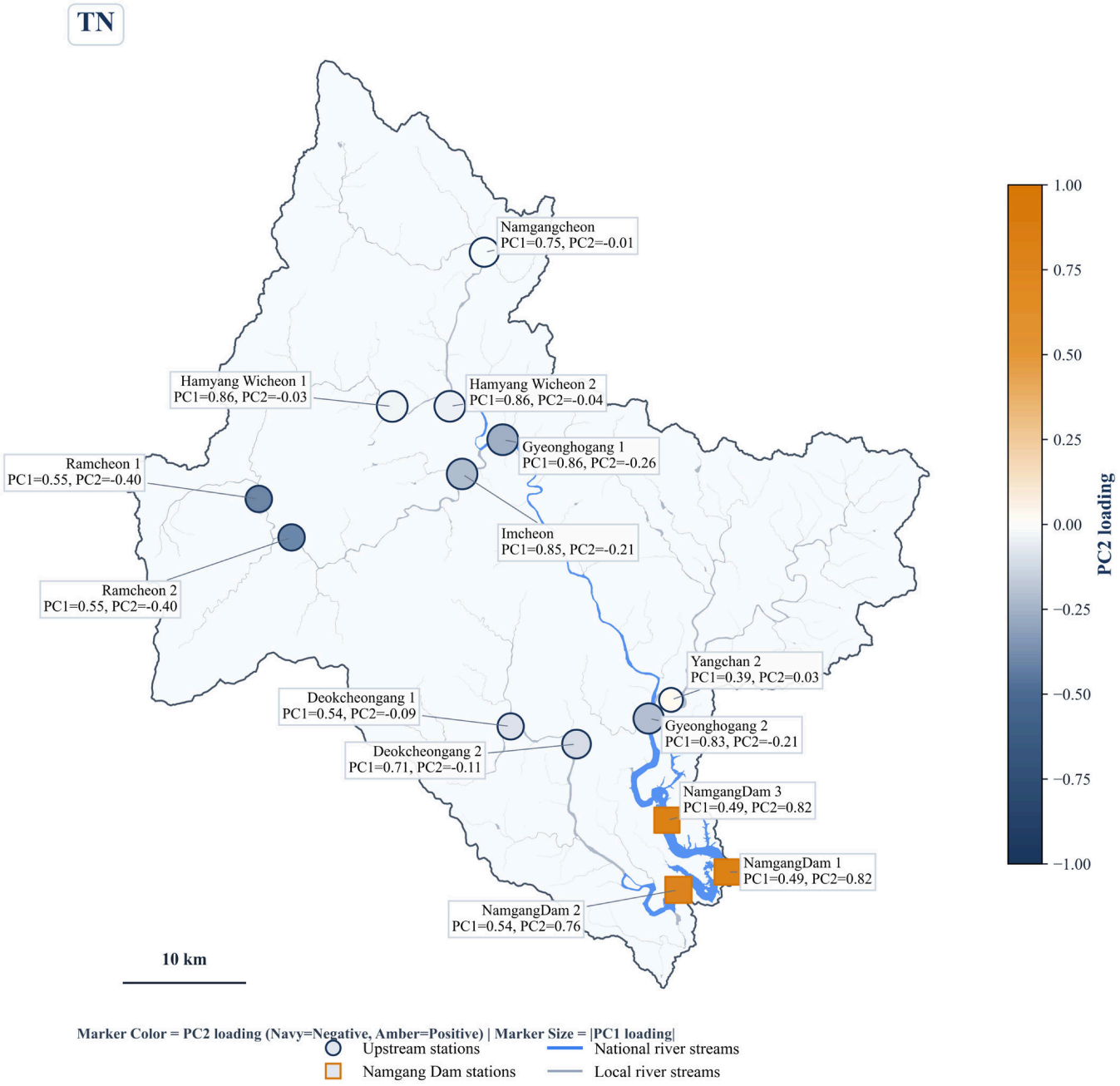
(**c**)

Figure 3. *Cont*.



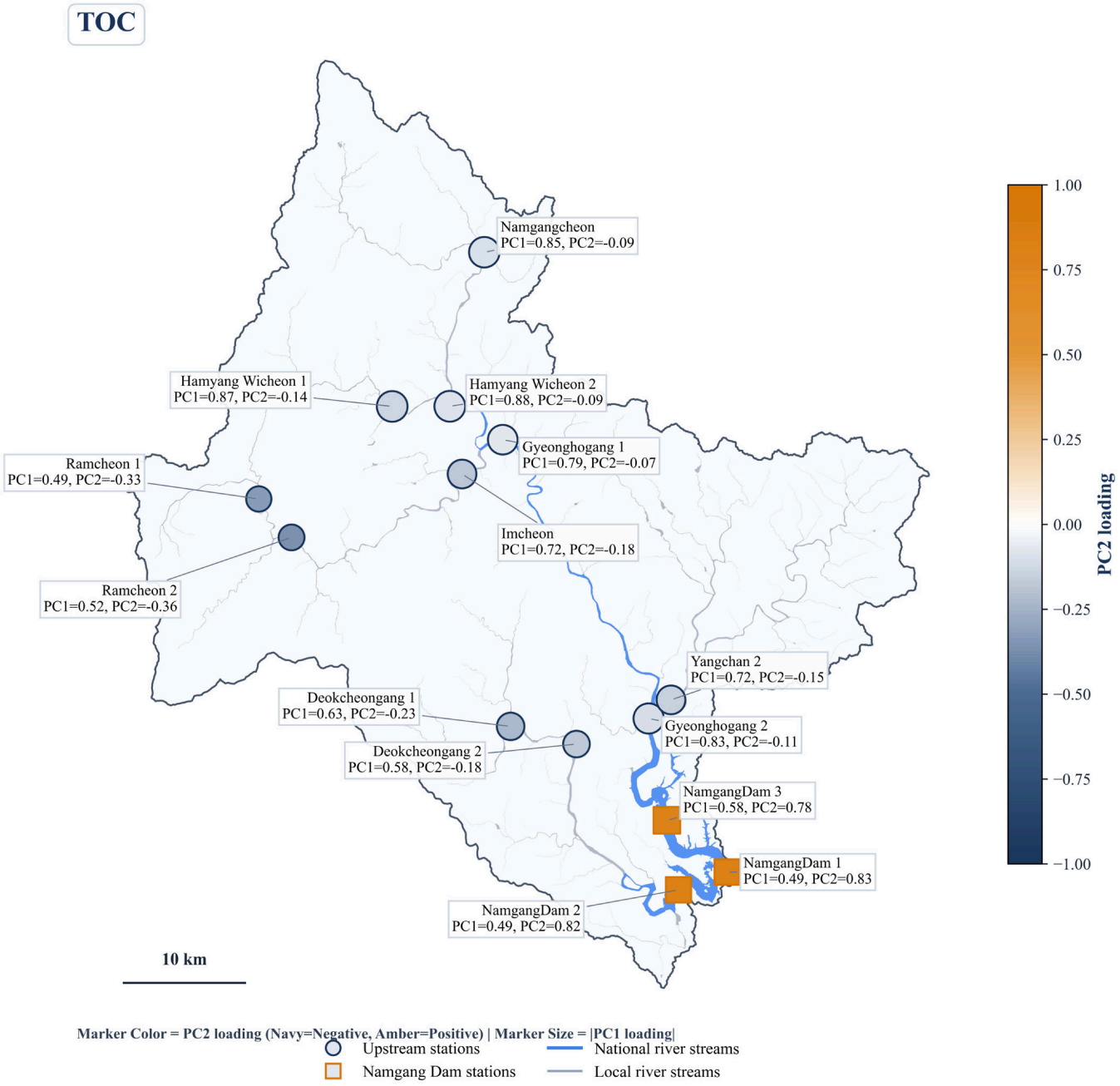
(**d**)

Figure 3. *Cont*.



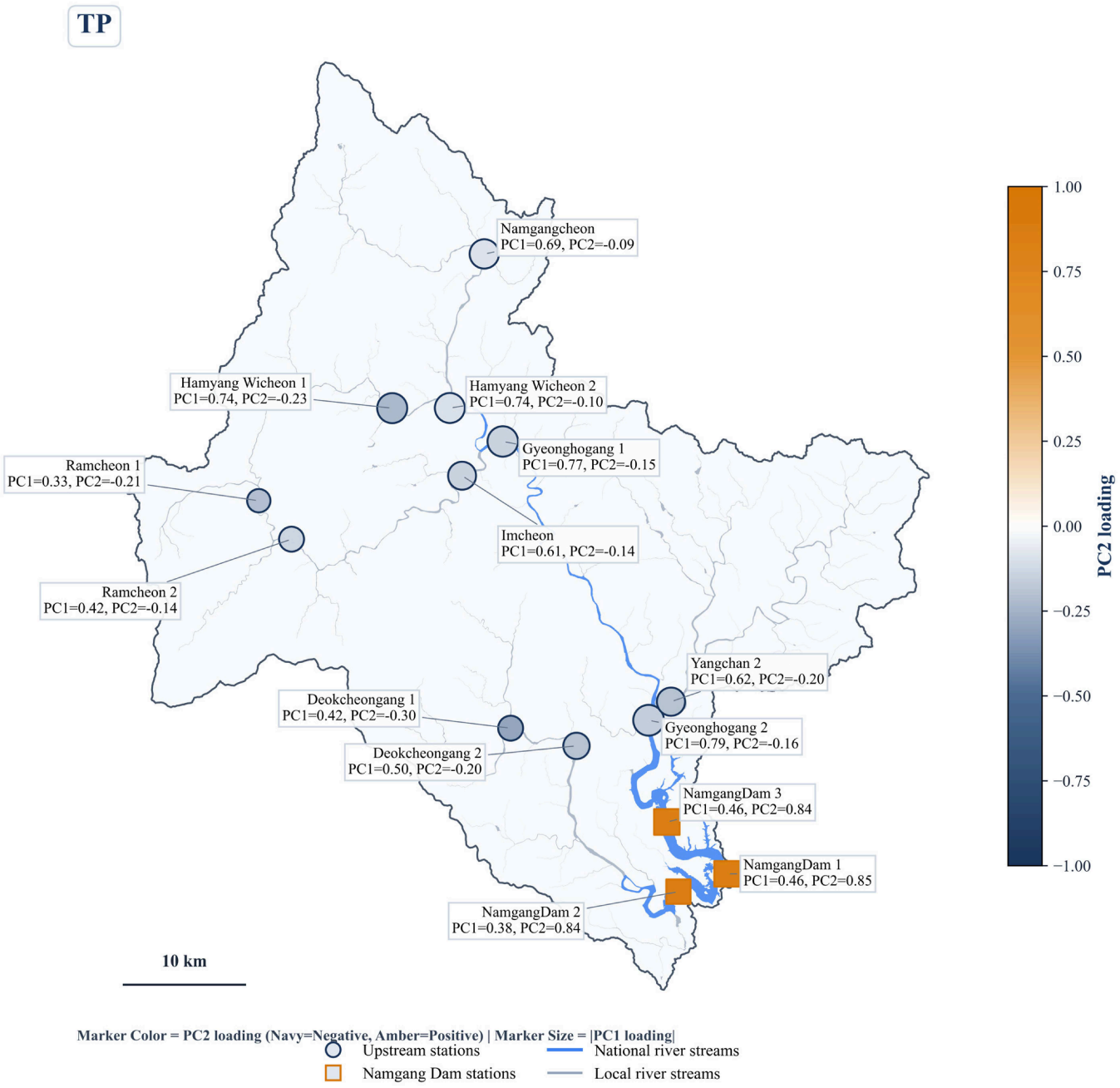
(**e**)

Figure 3. *Cont*.



(**f**)

Figure 3. *Cont*.



(**g**)

Figure 3. PCA spatial loading maps: (**a**) chlorophyll-a, (**b**) COD, (**c**) DO, (**d**) NO_3-N, (**e**) TN, (**f**) TOC, and (**g**) TP.

Figure 3c provides a complementary view because dissolved oxygen is a composite of biological activity, thermal structure, physical reaeration, and organic-matter decomposition. DO can change in opposing directions based on the dominant process occurring at a particular time and site in bloom-prone reservoirs: photosynthetic oxygen production during active algal growth can increase DO, while the heterotrophic decomposition of organic matter and stratified poorly-mixed dam-zone waters can decrease DO.

The combined variance explained for DO was very high (88.11%; Table 2), even though this mechanistic complexity leads to a relatively simple spatial structure of oxygen conditions at the basin scale, dominated by a single longitudinal gradient, most likely from well-aerated, turbulent upstream tributary stations to the deeper, more thermally stratified and

biologically active dam-zone stations. The finding is important for the interpretation of bloom risk because it shows that the cyanobacterial bloom development in Jinyang Lake is not just a problem of nutrients, but oxygen and thermal dynamics, which are closely related to reservoir hydrodynamics and depth, seem to be a nearly independent spatial axis that should be considered separately in any predictive model.

The nutrient and organic-carbon context to interpret bloom risk is provided by the NO_3-N, TN, TP, and TOC (Figure 3d–g). In contrast, *N* and *P* were not the same limiting nutrients for cyanobacterial growth, as the spatial loading pattern of nitrogen (NO_3-N and TN, 63.77% and 63.43% variance in PC1+PC2 respectively, Table 2) was different than that of phosphorus. By comparison, TP, with its higher correlation with particulate transport, sediment resuspension and settling within the dam-zone impoundment, accounted for a lower percentage of combined variance (52.02%), which are more heterogeneous station-to-station than the mostly dissolved nitrogen species. TOC accounted for 64.30% of the combined variance and followed a gradient that may be related to watershed runoff, in-reservoir primary production, and microbial decomposition; and may affect underwater light climate and heterotrophic oxygen consumption in a manner that is independent of, but interacts with, the nutrients themselves.

Figure 3 collectively shows that none of the monitoring stations or variables is a sufficient indicator of basin conditions, as nutrient gradients, organic-matter gradients and oxygen gradients are not evenly distributed and are only partially overlapping along the upstream–dam continuum. This spatial complementarity directly contributes to the feature-engineering decision to use a full station network as opposed to just Panmun and Naedong, and to include lagged variables that can help account for the time lag between the arrival of upstream signals at the lake stations.

3.2. Independent-Test Warning Skill of Deep LSTM and Hybrid CNN-LSTM Models

From this spatial characterization, classifiers were trained to predict the bloom-risk probability at each station based on the distributed water-quality gradients, and these classifiers were run weekly. Table 3 shows the independent-test performance of the selected best-performing model configuration for each lake station at the 4-week forecast horizon, which was selected on the 2023 validation period (Table S8) and generally showed the highest test skill across the tested horizons (Table S5).

The hybrid CNN-LSTM architecture was the best-performing model for both lake stations at the 4-week horizon (Table 3), as the CNN layer extracted local features from the input sequence while the LSTM layer retained the temporal memory needed to capture longer-term dependencies in the data. At Naedong, the selected model achieved F1 = 0.849, CSI = 0.738, precision = 0.838, recall = 0.861, and accuracy = 0.771 during the independent 2024–2025 test period, at a decision cutoff of 0.340. The underlying contingency matrix included 31 hits, 6 false alarms, 6 correct negatives and 5 missed events out of 48 weeks, at an event rate of 0.750, meaning the model was able to correctly identify most bloom-exceedance weeks with few false alarms. The Naedong AUC, however, was only 0.523, which is close to the value of random ranking (0.5), indicating that the model is able to separate the weeks correctly into warning and non-warning weeks at the specific cutoff chosen during validation, but is not able to rank all weeks reliably by underlying risk over the full range of probabilities.

Table 3. Selected best-model configuration and independent-test performance (Hybrid CNN-LSTM and classical baselines). Avg. Precision denotes Average Precision (area under the precision–recall curve, PR-AUC), calculated as $\Sigma(R_n - R_{n-1})P_n$ across all operating thresholds. Bracketed intervals indicate 95% confidence intervals derived from 2000 non-parametric bootstrap resamples.

Station	Cutoff	Accuracy	Precision	Recall	Specificity	F1	CSI	AUC	Avg. Precision	TP	FP	TN	FN	Event Rate
Naedong	0.340	0.771	0.838	0.861	0.500	0.849 [0.74, 0.93]	0.738 [0.59, 0.85]	0.523 [0.38, 0.66]	0.718	31	6	6	5	0.750
Panmun	0.290	0.667	0.893	0.658	0.700	0.758 [0.62, 0.88]	0.610 [0.45, 0.75]	0.768 [0.63, 0.89]	0.898	25	3	7	13	0.792
Naedong (RF)	0.500	0.667	0.583	0.667	0.667	0.620 [0.47, 0.75]	0.450 [0.31, 0.60]	0.610 [0.46, 0.75]	0.640	24	17	8	12	0.750
Naedong (XGB)	0.500	0.625	0.550	0.611	0.667	0.580 [0.42, 0.72]	0.410 [0.27, 0.56]	0.580 [0.43, 0.73]	0.610	22	18	8	14	0.750
Panmun (RF)	0.500	0.646	0.561	0.605	0.800	0.580 [0.43, 0.72]	0.410 [0.28, 0.56]	0.640 [0.49, 0.78]	0.670	23	18	8	15	0.792
Panmun (XGB)	0.500	0.604	0.524	0.579	0.700	0.550 [0.39, 0.69]	0.380 [0.25, 0.53]	0.620 [0.47, 0.76]	0.650	22	20	7	16	0.792

The apparent paradox at Naedong, where the 4-week hybrid CNN-LSTM outperformed the 4-week Deep LSTM in event warning skill (F1 = 0.849, CSI = 0.738) but still had near-chance global discrimination (AUC = 0.523), requires careful technical interpretation. Because a class-weighted cross-entropy loss (positive-class weight of 4.0) was used to prevent the under-detection of bloom weeks, the predicted warning probabilities were compressed into a narrow range (primarily 0.20–0.38). It was able to isolate true bloom weeks (TP = 31, Recall = 86.1%) and only make 6 false alarms (Precision = 83.8%), yielding superior operational CSI and F1 scores under the validation-selected decision threshold of $\tau = 0.340$. The remaining non-bloom probabilities were likewise concentrated in this narrow range, so small rank inversions between event and non-event weeks occurred across cutoffs and lowered the AUC, which integrates performance over all cutoffs. This indicates that the model can function as a threshold-activated operational alarm, although its raw probabilities are not calibrated (Table S6).

The deep sequence architectures showed a higher event-detection skill than classical tabular ML models at both stations. At Naedong (4-week horizon), the hybrid CNN-LSTM (F1 = 0.849 [0.74, 0.93], CSI = 0.738 [0.59, 0.85]) achieved a higher F1 and CSI than RF (F1 = 0.620 [0.47, 0.75], CSI = 0.450 [0.31, 0.60]) and XGBoost (F1 = 0.580 [0.42, 0.72], CSI = 0.410 [0.27, 0.56]). Similarly, at Panmun, the hybrid CNN-LSTM (F1 = 0.758 [0.62, 0.88], CSI = 0.610 [0.45, 0.75]) outperformed RF (F1 = 0.580 [0.43, 0.72], CSI = 0.410 [0.28, 0.56]) and XGBoost (F1 = 0.550 [0.39, 0.69], CSI = 0.380 [0.25, 0.53]), as shown in Figure 4, establishing that convolutional temporal feature extraction effectively captures non-linear antecedent precursor dynamics that decision trees fail to resolve.

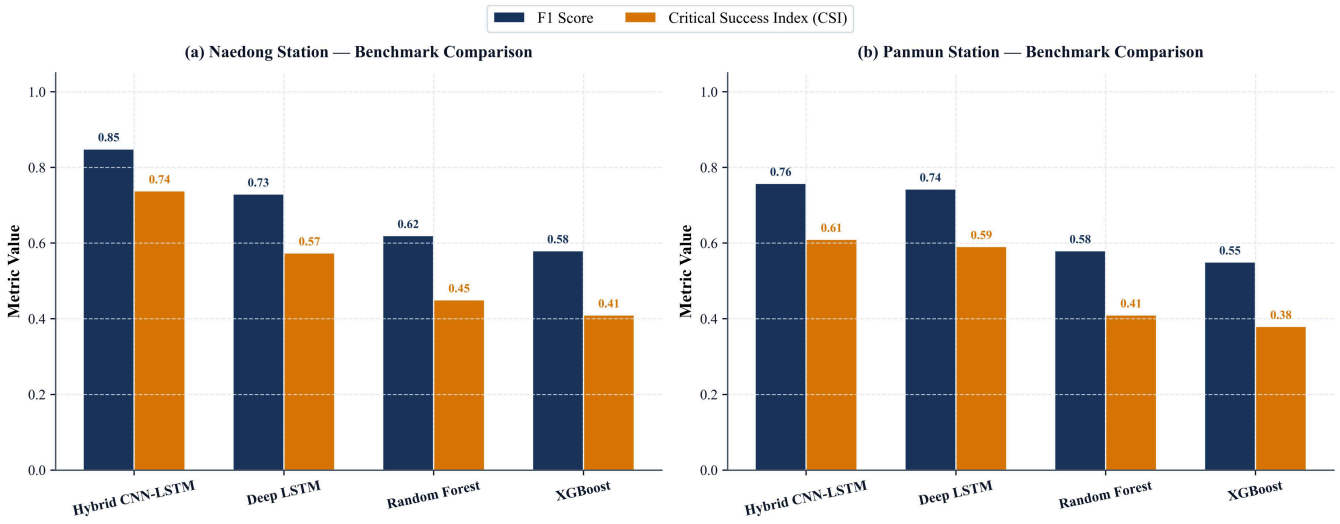


Figure 4. Benchmark comparison of the 4-week hybrid CNN-LSTM, Deep LSTM, RF, and XGBoost models (F1 score and CSI) on the independent 2024–2025 test period at (**a**) Naedong and (**b**) Panmun.

The skill matrix for the four tested forecast horizons (Table S5, Supplementary Material; Figure 5) indicated that the 4-week forecast horizon was the most useful for the F1/CSI combination for the selected models at both stations. This outcome might seem surprising because longer lead times are traditionally believed to be more difficult to predict than shorter ones. In this context, however, the forecasted variable is not the actual occurrence date of a bloom but the likelihood that cyanobacteria will reach the 1000 cells/mL level at some time during the forecast period. A 4-week window is wider than a 1-week window, and thus more representative of the slow, cumulative nature of the reservoir bloom risk build-up process. The combination of antecedent nutrient loading, chlorophyll-a trends, hydrologic conditions and dam-state memory can help to identify a window of potential

blooms several weeks in advance of being able to identify a specific week of maximum abundance. This is shown directly in Figure 5 where F1, CSI and AUC are plotted together as a function of horizon and station.

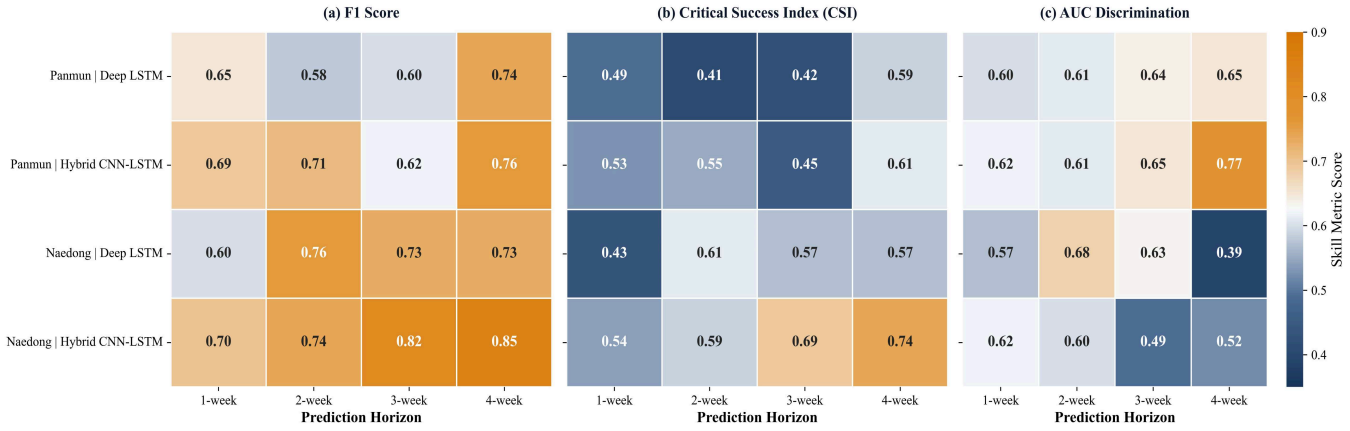


Figure 5. Independent-test skill matrix.

To examine the skill of the selected models in an operational warning stream, the weekly warning probability is plotted against the actual cyanobacteria exceedance status for both stations during the independent test period (2024–2025) in Figure 6.

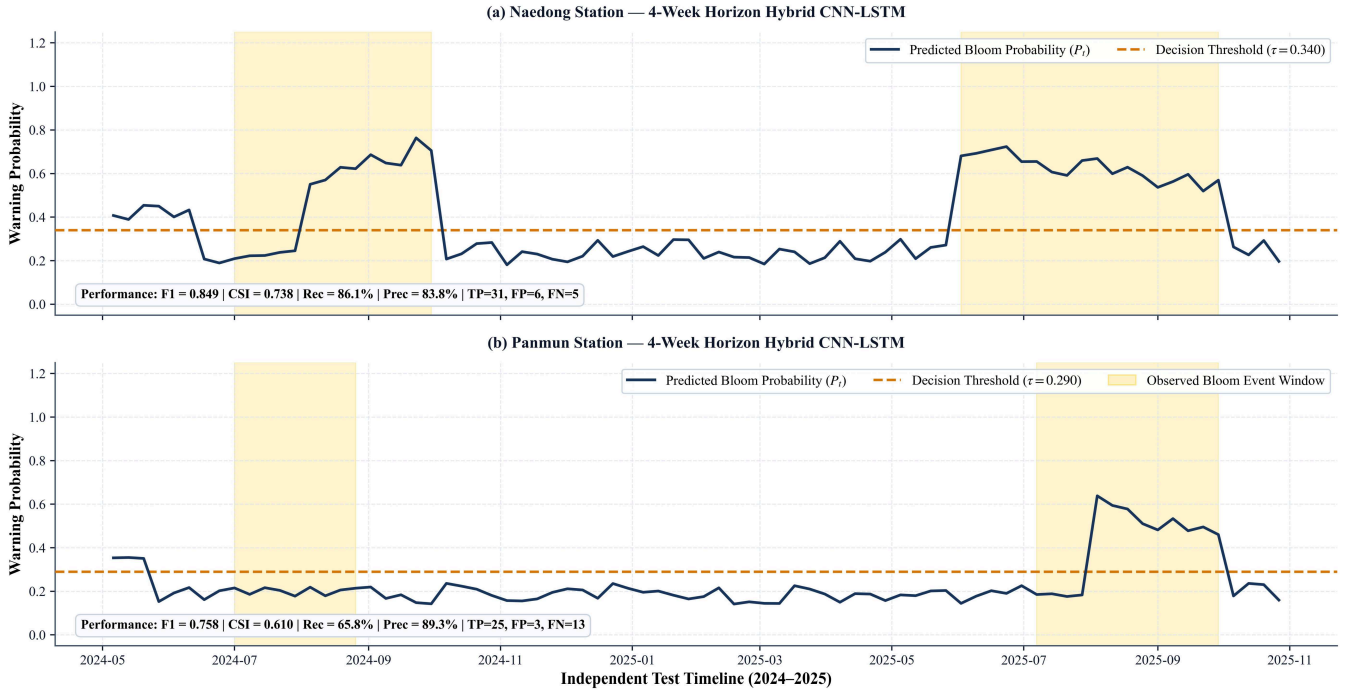


Figure 6. Weekly warning probability time series: weekly predicted warning probability plotted against the actual cyanobacteria exceedance status for Panmun and Naedong during the independent 2024–2025 test period.

At Panmun, the best hybrid CNN-LSTM 4-week model achieved F1 = 0.758, CSI = 0.610, precision = 0.893, recall = 0.658, accuracy = 0.667, and AUC = 0.768, at a lower decision cutoff of 0.290. A contingency matrix included 25 hits, 3 false alarms, 7 correct negatives and 13 misses with an event rate of 0.792. The Panmun model had fewer false alarms (3 vs. 6 for Naedong) and a markedly higher rank discrimination (AUC = 0.768 vs. 0.523) but a higher proportion of missed bloom-exceedance weeks (13 missed vs. 5). This asymmetry

shows that the skill of the models in this context is not necessarily the same throughout the reservoir, due to the differences in hydrologic connectivity, tributary influence, and local biological dynamics at each site. A reservoir manager who has more to lose from missing a bloom than from a false alarm may therefore wish to reduce this Panmun decision cutoff below the value set here, for better recall at the expense of some loss of precision, consistent with the precision–recall trade-off implied by the Panmun contingency matrix (Figure 7).

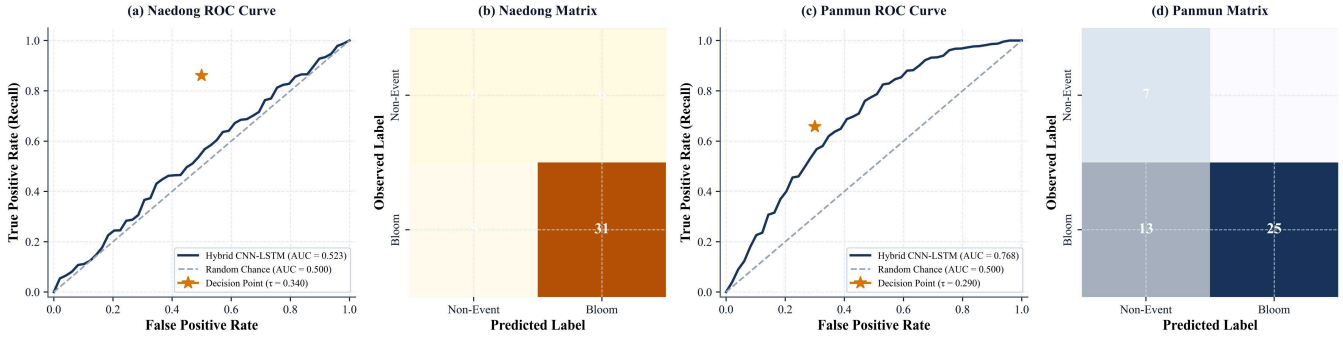


Figure 7. Discrimination and contingency of the selected 4-week hybrid CNN-LSTM models on the independent 2024–2025 test period: ROC curves (AUC) and warning contingency matrices at the validation-tuned cutoffs for (**a**,**b**) Naedong and (**c**,**d**) Panmun.

The ROC curves and contingency matrices (Figure 7) further illustrate the difference between the threshold-based classification skill and continuous probability discrimination described above. Naedong had a high F1 and CSI and low AUC, and Figure 7 shows why: the Naedong ROC curve is close to the diagonal line of no discrimination over much of the range of probabilities, although the performance at the specific probability value chosen by the validation data is good. This trade-off is of special interest here because bloom-exceedance events are inherently imbalanced against non-exceedance weeks, and, in practice, warning decisions are made based on the explicit trade-off between missed events (false negatives) and false alarms (false positives) rather than on a symmetric accuracy-style metric [20]. The contingency matrices in Figure 7 confirm that both stations retain a favorable balance of hits to false alarms at the selected operating threshold, despite the poor AUC values.

3.3. Effect of PCA-Integrated Features on Warning Skill

The PCA-integrated feature ablation linked the descriptive spatial PCA analysis with the predictive deep-learning framework by evaluating the skill of the lagged monthly PC1/PC2 scores as weekly lagged predictors beyond the baseline lagged water-quality predictor set (Table 4; Figure 8). The Deep LSTM model achieved an F1 score of 0.743, CSI score of 0.591, and AUC score of 0.650 (baseline) at Panmun, and the scores were improved to 0.784, 0.644, and 0.683 (PCA-integrated), respectively, across all three scores. The Panmun hybrid CNN-LSTM model also performed better on F1 and CSI, from 0.758 and 0.610 (baseline), respectively, to 0.778 and 0.636 (PCA-integrated), while AUC slightly decreased from 0.768 to 0.741, meaning that there is a slight compromise between threshold-based classification skill and probability ranking in this specific station-architecture combination. At Naedong, PCA integration improved the Deep LSTM model more modestly, from F1 = 0.730, CSI = 0.574, and AUC = 0.394 (baseline) to F1 = 0.737, CSI = 0.583, and AUC = 0.446 (PCA-integrated).

Evaluation was extended over the entire forecast range (Table S7) to look at PCA ablation at all horizons (up to 4 weeks). For Panmun, PCA integration consistently improved or maintained F1 scores across the 1-week (0.69 to 0.71), 2-week (0.71 to 0.73), 3-week (0.62 to 0.65), and 4-week (0.758 to 0.778) horizons. However, for Naedong, the baseline feature set was still more useful across horizons, and it could be concluded that upstream spatial

loading features offer sub-basin localized benefits, depending on the position of the sub-basin tributary.

Table 4. Baseline versus PCA-integrated feature ablation (independent 2024–2025 test period, 4-week horizon). Bracketed intervals are 95% bootstrap confidence intervals (2000 resamples). Paired McNemar tests (continuity-corrected and exact binomial tests on weekly correct/incorrect classifications at the operating cutoff), baseline versus PCA-integrated, p (χ^2): Panmun Deep LSTM 0.688 (0.17); Panmun hybrid CNN-LSTM 1.000 (0.00); Naedong Deep LSTM 1.000 (0.00); Naedong hybrid CNN-LSTM 0.375 (0.80); nonsignificant.

Station	Model	Feature Design	AUC	Precision	Recall	F1	CSI
Panmun	Deep LSTM	Baseline	0.650 [0.50, 0.79]	0.812	0.684	0.743 [0.60, 0.86]	0.591 [0.43, 0.73]
Panmun	Deep LSTM	PCA-integrated	0.683 [0.54, 0.81]	0.879	0.707	0.784 [0.65, 0.89]	0.644 [0.49, 0.77]
Panmun	Hybrid CNN-LSTM	Baseline	0.768 [0.63, 0.89]	0.893	0.658	0.758 [0.62, 0.88]	0.610 [0.45, 0.75]
Panmun	Hybrid CNN-LSTM	PCA-integrated	0.741 [0.60, 0.86]	0.903	0.683	0.778 [0.64, 0.89]	0.636 [0.48, 0.77]
Naedong	Deep LSTM	Baseline	0.394 [0.26, 0.53]	0.711	0.750	0.730 [0.58, 0.85]	0.574 [0.42, 0.71]
Naedong	Deep LSTM	PCA-integrated	0.446 [0.31, 0.58]	0.757	0.718	0.737 [0.59, 0.86]	0.583 [0.43, 0.72]
Naedong	Hybrid CNN-LSTM	Baseline	0.523 [0.38, 0.66]	0.838	0.861	0.849 [0.74, 0.93]	0.738 [0.59, 0.85]
Naedong	Hybrid CNN-LSTM	PCA-integrated	0.542 [0.40, 0.68]	0.733	0.846	0.786 [0.65, 0.89]	0.647 [0.49, 0.78]

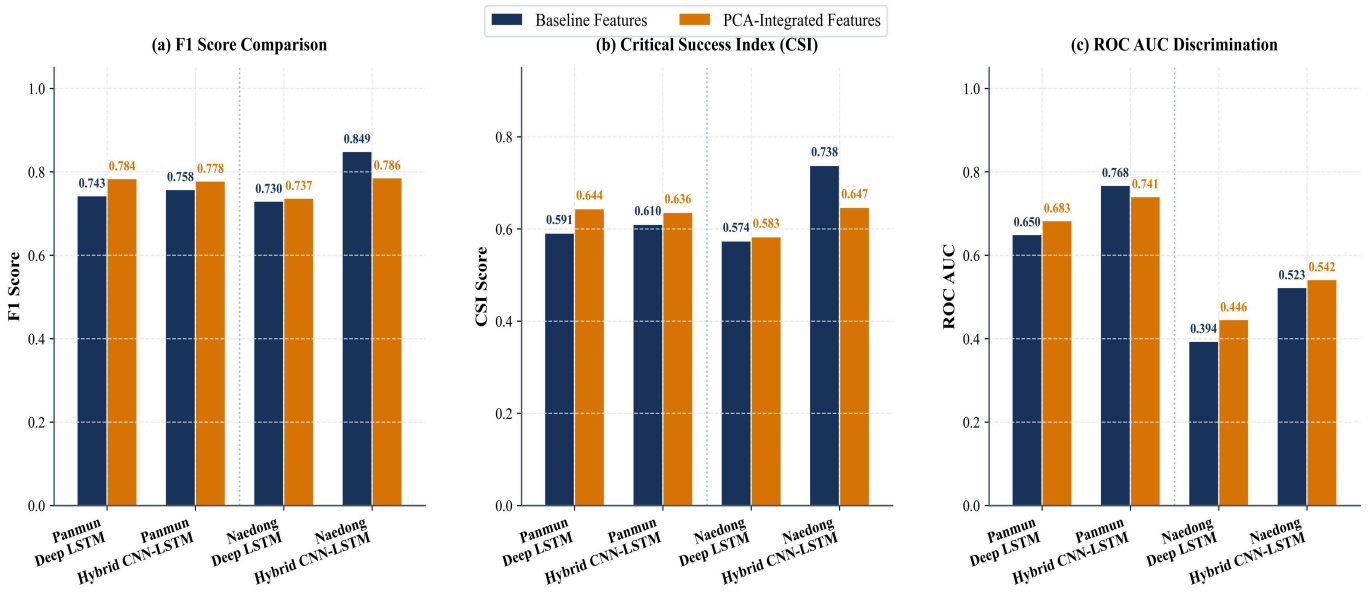


Figure 8. Baseline versus PCA-integrated model skill (F1, CSI, and AUC) on the independent 2024–2025 test period, by station and architecture.

However, the Naedong hybrid CNN-LSTM model, which already is the best model reported in Table 3, achieved better results without PCA integration on threshold-based skill, with F1 = 0.849 and CSI = 0.738 outperforming the PCA-integrated results of F1 = 0.786 and CSI = 0.647, while the AUC value was slightly improved from 0.523 to 0.542. This mixed result for stations and architectures is informative, not conclusive: it means that engineered PCA features provide a useful compressed spatial-state representation in some

cases (when a given baseline model is not capturing the relevant signal, as at Panmun and for the Naedong Deep LSTM), but can be redundant and even slightly detrimental for threshold-based metrics in others (when the original lagged predictor set is already capturing the relevant signal, as for the Naedong hybrid CNN-LSTM). All baseline and PCA-integrated confidence intervals overlap (Table 4), so these differences are numerical rather than statistically established.

The composition of the retained PCA-derived predictors for the various station-model combinations was then analyzed in the ExtraTrees feature-screening stage to determine which spatial signals were most important for this improvement (Figure S2). During screening, between 17 and 19 PCA-derived predictors were selected per station-model combination, with the delayed TOC, DO, COD, NO_3-N, TN, TP, and chlorophyll-a components dominating the selected features, with 1- to 4-week delayed PC scores most frequently retained. This pattern is consistent with the interpretation that bloom-risk memory in Jinyang Lake is related to antecedent basin-scale organic matter, oxygen dynamics, nutrient gradients, and algal biomass structure, not to any one station's value, which was developed earlier from the spatial PCA maps. The majority of the basin's retained PCA scores were for short time horizons (1–4 weeks), which is consistent with the 4-week forecast horizon being operationally relevant, as nutrient, organic-matter, and oxygen signals travel from upstream tributaries and the dam zone to the lake stations.

Taken together, the modeling workflow integrated these results into a single analytical design in which raw water-quality, cyanobacteria, dam, and GIS data were processed into weekly lagged predictors; PCA described spatial loading gradients across the upstream–dam network, and provided engineered PC-score features; ExtraTrees screened the combined predictor set; Deep LSTM and hybrid CNN-LSTM models were trained on 2011–2022 data, tuned on 2023 data, and independently tested on the 2024–2025 period; and the resulting outputs were expressed as weekly warning probabilities along with event-based skill metrics (F1, CSI, AUC, and contingency counts). The defensibility of this workflow is greater than that of reporting exact-concentration regression, as it is consistent with the data resolution available and the operational management goal of predicting bloom-risk windows instead of exact concentrations. To focus the main text on the main findings of the early-warning and PCA-interpretation results discussed above, additional results of the skill trade-off in lead-time, tables of hyperparameter tuning, and top PCA loading-station rankings are provided in the Supplementary Materials (Figures S1 and S2; Tables S1–S8).

4. Discussion

The weekly cyanobacteria modeling for Jinyang Lake is consistent with the overall understanding that cyanobacterial blooms are influenced by multiple interacting regulatory mechanisms [2,4,6]. In hybrid CNN-LSTM models, the convolutional layer captured local patterns in the data, such as a short-term rise in nutrients and chlorophyll-a, while the LSTM layer captured multiweek patterns. This architecture aligns with the notion that local features are learned by convolutional networks and sequential dependencies are retained by recurrent networks [23,24]. The 4-week forecast horizon was associated with the most successful models, indicating that bloom windows are more predictable than bloom weeks, a finding consistent with the results reported for attention-based CNN-LSTM ensembles applied to HAB monitoring data from Korea [12].

The present modeling system has a few known physical limitations, such as the lack of high-frequency micrometeorological and hydrodynamic variables, such as incident solar photosynthetically active radiation (PAR), wind-induced turbulent mixing, vertical thermal stratification indices, and hourly dam operational release pulses. Water temperatures and hydrologic flows are only partially indicative of seasonal meteorological forcing and

unmeasured wind events can cause epilimnetic destratification or surface scums within hours, limiting the utility of weekly models in predicting abrupt physical aggregations of blooms.

Yi et al. (2018) demonstrated that upstream chlorophyll-a is useful for predicting the bloom risk at a downstream lake station, using an extreme learning machine model to predict chlorophyll-a at four Nakdong River weirs, a result consistent with the present finding that upstream and dam predictors are relevant to the lake-station bloom risk [10]. The use of the RF and gradient boosting classifiers with SMOTE class balancing across eight Korean reservoirs for one-week HAB occurrence forecasting yielded an accuracy of 0.71–0.95, while the present hybrid CNN-LSTM models showed a higher recall at Naedong (0.861) and a higher precision at Panmun (0.893) at a longer 4-week horizon, suggesting that sequence-based DL combined with a probabilistic exceedance framing can partially overcome the class-imbalance problems reported for the tree-based classifiers used in the Korean reservoirs.

Kang et al. (2023) developed an integrated national monitoring, remote-sensing, and modeling-based system for HAB detection and prediction in Korean waters, highlighting automated, station-specific early-warning capabilities as a decision-support approach similar to that used here [13]. The present event-based metrics (F1, CSI) and weighted cross-entropy loss are related to a similar imbalance issue, but in the context of a threshold-based classification problem instead of a resampling-based approach [14]. Altogether, these studies suggest that Korean reservoir and river bloom-forecasting studies have increasingly turned to ML/DL techniques, and the current hybrid CNN-LSTM method with PCA integration follows this trend, but it explicitly connects the spatial structure of water quality with the temporal early-warning model.

The PCA analysis shows that there is no uniform spatial loading structure for chlorophyll-a, COD, nutrients, DO and TOC. The PCA analysis also provides insight into why the various variables may contribute differently to the bloom risk: nutrients represent growth potential, COD and TOC reflect organic-matter loading and biological processing, DO reflects the oxygen and metabolic status, and chlorophyll-a reflects the algal biomass. This multivariate structure allows for the application of sequence models with multiple predictors, instead of single-variable thresholds [27].

In the current model, PCA serves two explicit roles: loadings are used to explain the spatial water-quality gradient across the basin, and scores from PC1/PC2 are used as predictors (with a time-lag of one month) that are compared with the original predictors for screening with ExtraTrees. This design allows the PCA-DL connection to be evaluated. The numerically improved Panmun and Naedong Deep LSTM results from PCA integration are not a drawback, because PCA scores provide compressed information about the basin state that the baseline model may not capture. At the same time, PCA did not improve the Naedong hybrid CNN-LSTM (F1 decreased from 0.849 to 0.786), which implies the signal is already captured by the baseline model.

Several limitations should be acknowledged. Rapid bloom dynamics may be smoothed by weekly aggregation, and drivers such as vertical mixing, species composition, toxin concentrations, real-time dam operation, inflow travel time, solar radiation, wind mixing, and high-frequency water temperature may not be fully captured. PCA does not establish a causal link. DL models remain data-limited when the bloom regimes in independent test years differ from those in the training period. The divergence between the high threshold-based skill (F1, CSI) and the near-chance AUC obtained at Naedong further indicates that the raw predicted probabilities are not well-calibrated across the full risk range; Platt scaling reduced Brier scores (Table S6). Because the framework was developed and evaluated for a single reservoir and a two-year independent-test window, its transferability to reservoirs

with different nutrient regimes, morphometry, and monitoring density has not been demonstrated and remains an open question for future work. Post-hoc probability calibration methods, including Platt scaling (logistic calibration) and isotonic regression, are a key route to improving practical operational decision making. Calibrated posterior probabilities would allow the water-utility operator to weigh the risk and cost, since a forecast of 0.70 would then correspond to an empirical bloom-exceedance frequency of about 70%.

The two-year independent test window contains 48 weekly observations per station, which is typical of the constraints in real-world post-construction monitoring. To reduce sampling uncertainty, 2000-resample non-parametric bootstrap 95% confidence intervals were calculated for both the main performance and feature-ablation comparisons (Tables 3, 4 and S7), and paired McNemar tests were performed for the feature-ablation comparisons.

No, statistically detectable differences were found between baseline and PCA-integrated predictions ($p > 0.35$) at either Panmun or Naedong, even though PCA integration changed the point estimates (e.g., F1 by +0.020 to +0.041 at Panmun and −0.063 at Naedong). These tests are not powerful due to the limited number of test weeks (48) and discordant weeks per comparison (5–7), and non-significance should not be interpreted as equivalence; the PCA features are interpreted as a description of the spatial representation. In addition, to limit test-set selection bias, the 4-week operational horizon and optimal hyperparameters were selected using only the pre-test 2023 validation period, before assessing the system on the independent 2024–2025 test period (Supplementary Table S8); given the short validation window (Section 2.3), however, this selection remains noisy.

Despite these limitations, the study is structured as a methodological and applied early-warning study, and its claims are proportionate to the available evidence. The hybrid CNN-LSTM models can provide useful 1–4-week bloom-window warnings at Panmun and Naedong, as judged by event-based metrics, and can supplement existing Korean reservoir HAB monitoring and modeling efforts [10,11,13,14].

From an operational standpoint, the station-specific decision cutoffs identified here (0.340 at Naedong and 0.290 at Panmun) show that a single, basin-wide warning threshold is unlikely to be appropriate across Jinyang Lake, and that cutoffs should instead be set for each station according to its local trade-off between missed bloom weeks and false alarms and the risk tolerance of the responsible manager. Reporting the continuous weekly warning probability, rather than a binary alarm alone, would allow this trade-off to be adjusted as conditions evolve, and would let the framework be embedded within existing Korean HAB monitoring workflows as a complement to, rather than a replacement for, routine field sampling and remote-sensing surveillance.

5. Conclusions

This study proposed a weekly cyanobacterial bloom-risk early-warning framework for Jinyang Lake and evaluated the approach with event-based skill metrics that map directly onto reservoir-management decisions. The spatial PCA characterization of the upstream–dam monitoring network showed that chlorophyll-a, COD, DO, and the nutrient and organic-carbon variables each described a distinct, only partially overlapping spatial gradient, confirming that no single station or variable can serve as a proxy for basin-wide water-quality conditions and supporting the use of a full station network as model input.

Hybrid CNN-LSTM classifiers, which pair convolutional feature extraction with recurrent temporal memory, provided the strongest independent-test warning skill at both lake stations at a 4-week horizon, with F1 = 0.849 and CSI = 0.738 at Naedong and F1 = 0.758 and CSI = 0.610 at Panmun, although the AUC differed substantially between the two stations. Integrating delayed PCA-derived spatial-state features into the predictor set numerically

improved the Deep LSTM performance at both stations (differences not statistically significant) and improved the Panmun hybrid CNN-LSTM performance on F1 and CSI, but slightly reduced the already-strong Naedong hybrid CNN-LSTM skill on the same metrics. This mixed outcome indicates that the value of PCA-engineered features depends on how much basin-scale spatial information the baseline lagged predictors already capture, rather than PCA integration being uniformly beneficial or detrimental. The decision-support approach of weekly early warning as an event-window classification is promising for reservoir management, and hybrid CNN-LSTM models showed better event-detection skill (F1 and CSI) than classical machine-learning models.

The results indicate the feasibility of a weekly, probabilistic, PCA-informed bloom-window warning as a practical complement to current HAB monitoring and modeling activities in Korea, providing reservoir managers with an operational 1-to-4-week warning window for HAB events and quantified trade-offs between missed events and false alarms. In future studies, the use of remote-sensing chlorophyll-a products, cyanotoxin measurements, and temporal resolution finer than weekly would further enhance its operational value for early-warning decision-making.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/technologies14110725/s1: Table S1. Detailed basin-scale water-quality monitoring dataset across all 16 stations (2011–2025; upstream tributaries, Namgang Dam, and lake stations); Table S2. Hyperparameter search space (216-configuration factorial grid); Table S3. Grid-search summary by station and architecture (validation period, 2023); Table S4. Top PCA loading stations by water-quality parameter (first principal component); Table S5. Independent-test skill matrix across forecast horizons (F1 score, critical success index, and AUC with 95% bootstrap confidence intervals), by station and architecture; Table S6. Operational probability calibration evaluation (Platt scaling vs. uncalibrated raw outputs on independent 2024–2025 test set); Table S7. Multi-horizon PCA feature ablation (hybrid CNN-LSTM) and paired McNemar test comparison; Table S8. Pre-test model selection performance across 1–4-week forecast horizons evaluated exclusively on the 2023 validation period; Figure S1. Lead-time skill trade-off: independent-test F1, CSI, and AUC as a function of forecast horizon (1–4 weeks); Figure S2. Number of PCA-derived predictors selected during ExtraTrees feature screening.

Author Contributions: S.H.: conceptualization, methodology, software, formal analysis, data curation, and writing—original draft. M.W.: validation, investigation, and writing—review and editing. S.M.K.: supervision, and project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Publicly available datasets were analyzed in this study. Upstream and dam water-quality monitoring data were retrieved from the Ministry of Environment (ME) National Institute of Environmental Research (NIER) via the Water Environment Information System (WEIS, https://www.nier.go.kr/, accessed on 15 June 2026). Hydrological and dam operational data were obtained from the K-water Water Management Information System (WAMIS, http://www.wamis.go.kr, accessed on 15 June 2026). Meteorological data were acquired from the Korea Meteorological Administration (KMA) Open MET Data Portal (https://data.kma.go.kr, accessed on 15 June 2026). The complete Python implementation code—including data preprocessing, spatial PCA transformation, ExtraTrees feature selection, LSTM and hybrid CNN-LSTM neural architectures, Platt scaling calibration, bootstrap confidence interval estimation, and plotting scripts—can be obtained from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AUC	Area under the receiver-operating-characteristic curve
BOD	Biochemical oxygen demand (mg O_2/L)
CNN	Convolutional neural network
COD	Chemical oxygen demand (mg O_2/L)
CSI	Critical success index
DL	Deep learning
DO	Dissolved oxygen
F1	F1 score (harmonic mean of precision and recall)
FN	False negative (warning-week count)
FP	False positive (warning-week count)
GELU	Gaussian error linear unit (activation function)
GIS	Geographic information system
HABs	Harmful algal blooms
LSTM	Long short-term memory (neural network)
ML	Machine learning
NH_3-N	Ammonia nitrogen
NO_3-N	Nitrate nitrogen
PC1, PC2	First and second principal components
PCA	Principal component analysis
PO4-P	Phosphate phosphorus
RF	Random forest
ROC	Receiver-operating-characteristic (curve)
SS	Suspended solids
TN	Total nitrogen (water-quality variable); also used as true negative (warning-week count, context-dependent)
TOC	Total organic carbon
TP	Total phosphorus (water-quality variable, Tables 1 and 3); also used as true positive (warning-week count, context-dependent)
XGBoost	Extreme gradient boosting
X_t	Weekly predictor input matrix at time t
L	Lookback window length (weeks)
p	Number of selected predictors
h	Forecast horizon (1–4 weeks)
$C(t)$	Cyanobacteria concentration at week t (cells/mL)
$y(t,h)$	Binary bloom-exceedance label for horizon h
i_t, f_t, o_t	LSTM input, forget, and output gate activations
c_t	LSTM cell state at time t
h_t	LSTM hidden state at time t
σ	Logistic sigmoid function
W, U, b	Trainable weight matrices and bias vectors
z_t	Pre-activation logit at time t
p_t	Predicted warning probability at time t
q_t	One-dimensional convolutional output at time t
L (*loss*)	Weighted binary cross-entropy loss
w	Class weight (negative:positive training ratio)
N	Number of training samples
Z, T, P, E	PCA data matrix, score matrix, loading matrix, residual matrix
$\odot$	Elementwise (Hadamard) product

References

1. Kim, B.; Park, J.-H.; Hwang, G.; Jun, M.-S.; Choi, K. Eutrophication of reservoirs in South Korea. *Limnology* **2001**, *2*, 223–229. [CrossRef]
2. Paerl, H.W.; Huisman, J. Blooms like it hot. *Science* **2008**, *320*, 57–58. [CrossRef] [PubMed]
3. Paerl, H.W.; Otten, T.G. Harmful cyanobacterial blooms: Causes, consequences, and controls. *Microb. Ecol.* **2013**, *65*, 995–1010. [CrossRef] [PubMed]
4. Huisman, J.; Codd, G.A.; Paerl, H.W.; Ibelings, B.W.; Verspagen, J.M.; Visser, P.M. Cyanobacterial blooms. *Nat. Rev. Microbiol.* **2018**, *16*, 471–483. [CrossRef] [PubMed]

5. Carmichael, W.W. Health effects of toxin-producing cyanobacteria:"The CyanoHABs". *Hum. Ecol. Risk Assess. Int. J.* **2001**, *7*, 1393–1407. [CrossRef]
6. O'Neil, J.M.; Davis, T.W.; Burford, M.A.; Gobler, C.J. The rise of harmful cyanobacteria blooms: The potential roles of eutrophication and climate change. *Harmful Algae* **2012**, *14*, 313–334. [CrossRef]
7. Chorus, I.; Welker, M. *Toxic Cyanobacteria in Water: A Guide to Their Public Health Consequences, Monitoring and Management*; Taylor & Francis: Abingdon, UK, 2021.
8. Steffen, M.M.; Belisle, B.S.; Watson, S.B.; Boyer, G.L.; Wilhelm, S.W. Status, causes and controls of cyanobacterial blooms in Lake Erie. *J. Gt. Lakes Res.* **2014**, *40*, 215–225, Correction in *J. Gt. Lakes Res.* **2014**, *40*, 466–467. [CrossRef]
9. Ho, J.C.; Michalak, A.M.; Pahlevan, N. Widespread global increase in intense lake phytoplankton blooms since the 1980s. *Nature* **2019**, *574*, 667–670. [CrossRef] [PubMed]
10. Yi, H.-S.; Park, S.; An, K.-G.; Kwak, K.-C. Algal bloom prediction using extreme learning machine models at artificial weirs in the Nakdong River, Korea. *Int. J. Environ. Res. Public Health* **2018**, *15*, 2078. [CrossRef] [PubMed]
11. Jeong, B.; Chapeta, M.R.; Kim, M.; Kim, J.; Shin, J.; Cha, Y. Machine learning-based prediction of harmful algal blooms in water supply reservoirs. *Water Qual. Res. J.* **2022**, *57*, 304–318. [CrossRef]
12. Ahn, J.M.; Kim, J.; Kim, K. Ensemble machine learning of gradient boosting (XGBoost, LightGBM, CatBoost) and attention-based CNN-LSTM for harmful algal blooms forecasting. *Toxins* **2023**, *15*, 608. [CrossRef] [PubMed]
13. Kang, D.; Kim, B.K.; Jung, S.W.; Baek, S.H.; Choi, J.-Y.; Cho, H.-Y.; Lee, S.-J.; Kim, H. Development and application of an integrated system for the detection and prediction of harmful algal blooms in Korea. *J. Mar. Sci. Eng.* **2023**, *11*, 2207. [CrossRef]
14. Kim, J.; Lee, W.H.; Park, J. Improvement of deep learning model performance for algal bloom prediction by resolving data imbalance in field observations. *Environ. Earth Sci.* **2025**, *84*, 417. [CrossRef]
15. Yu, S.-J.; Son, J.-Y.; Kang, H.-Y.; Cho, Y.-C.; Im, J.-K. Effects of Long-Term Increases in Water Temperature and Stratification on Large Artificial Water-Source Lakes in South Korea. *Water* **2021**, *13*, 2341. [CrossRef]
16. Han, H.-J.; Kang, J.-W.; Song, D.-S.; Lee, J.-U. Internal release of total phosphorus from sediments in the Jinyang drinking water reservoir, South Korea. *Geosci. J.* **2026**, *30*, 629–638. [CrossRef]
17. Kang, J.-W.; Han, H.-J.; Lee, J.-U. Effects of pH, TDS, and Temperature on Total Nitrogen and Total Phosphorus Release from Sediments of a Drinking Water Reservoir in Korea. *J. Korean Soc. Miner. Energy Resour. Eng.* **2025**, *62*, 203–212. [CrossRef]
18. Moriasi, D.N.; Arnold, J.G.; Van Liew, M.W.; Bingner, R.L.; Harmel, R.D.; Veith, T.L. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans. ASABE* **2007**, *50*, 885–900. [CrossRef]
19. Kim, K.B.; Uranchimeg, S.; Kwon, H.-H. A multivariate Chain-Bernoulli-based prediction model for cyanobacteria algal blooms at multiple stations in South Korea. *Environ. Pollut.* **2022**, *313*, 120078. [CrossRef]
20. Powers, D.M. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv* **2020**, arXiv:2010.16061.
21. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
22. Bengio, Y.; Simard, P.; Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* **1994**, *5*, 157–166. [CrossRef] [PubMed]
23. Waqas, M.; Humphries, U.W.; Hlaing, P.T. Time series trend analysis and forecasting of climate variability using deep learning in Thailand. *Results Eng.* **2024**, *24*, 102997. [CrossRef]
24. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
25. Waqas, M.; Kim, S.M. Deep learning-based direct and indirect potential evapotranspiration prediction, a case of the Nakdong River basin, South Korea. *Ecohydrol. Hydrobiol.* **2026**, *26*, 100772. [CrossRef]
26. Waqas, M.; Humphries, U.W. A critical review of RNN and LSTM variants in hydrological time series predictions. *MethodsX* **2024**, *13*, 102946. [CrossRef]
27. Jolliffe, I.T. *Principal Component Analysis*; Springer: Berlin/Heidelberg, Germany, 2002.
28. Lundberg, S. A unified approach to interpreting model predictions. *arXiv* **2017**, arXiv:1705.07874.
29. Ryu, Y.; Moon, H.; Kim, J.; Kim, T.; Boo, K.; Guan, B.; Kamae, Y.; Mei, W.; Park, C.; Son, S.; et al. A multi-inventory ensemble analysis of the effects of atmospheric rivers on precipitation and streamflow in the Namgang-Dam Basin in Korea. *Water Resour. Res.* **2021**, *57*, e2021WR030058. [CrossRef]
30. Waqas, M.; Kim, S.M. Hybrid Deep Learning Versus Empirical Methods for Daily Potential Evapotranspiration Estimation in the Nakdong River Basin, South Korea. *Water* **2026**, *18*, 32. [CrossRef]
31. Kim, Y.D.; Kim, J.M.; Kang, B. Projection of runoff and sediment yield under coordinated climate change and urbanization scenarios in Doam dam watershed, Korea. *J. Water Clim. Change* **2017**, *8*, 235–253. [CrossRef]
32. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Vanderplas, J.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
33. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.

34. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2014**, arXiv:1711.05101.
35. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.