sustainability

MDPI

*Article*

# Parameter Sensitivity and Limits to Ecological Prediction in a Coupled GOTM–FABM–WET Application to a Shallow Lake in China

**Tangwu Yang [1,2,3], Jingyan Wang [3], Yiqi Shao [3], Pingyan Ge [3], Dehua Zhao [1], Dianpeng Li [4,5,*] and Shuqing An [1,3,*]**

[1] School of Life Sciences, Nanjing University, Nanjing 210046, China; liib@smail.nju.edu.cn (T.Y.); dhzhao@nju.edu.cn (D.Z.)
[2] Key Laboratory of Water Management in Taihu Basin, Ministry of Water Resources, Nanjing 210098, China
[3] Nanjing University Ecological Research Institute of Changshu, Suzhou 215500, China
[4] Xiangyang Institute of History and Culture, Hubei University of Arts and Science, Xiangyang 441053, China
[5] Hanjiang Institute, Hubei University of Arts and Science, Xiangyang 441053, China
* Correspondence: lldp0628@gmail.com (D.L.); anshq@nju.edu.cn (S.A.)

**Abstract**

We evaluated the diagnostic and predictive performance of the coupled General Ocean Turbulence Model–Framework for Aquatic Biogeochemical Models–Water Ecosystems Tool (GOTM–FABM–WET) in shallow Kun Cheng Lake, China. Monthly near-surface observations supported model calibration during 2021–2023 and out-of-sample ecological validation during 2024–2025. Model skill was assessed through one-at-a-time parameter screening, moment-independent Delta sensitivity analysis with sample-size convergence diagnostics, sequential multi-parameter calibration, and comparisons against statistical and nutrient mass-balance benchmarks. Water temperature calibration was strong (Nash–Sutcliffe efficiency, NSE = 0.933), although temperature observations were available only for 2021–2023. Dissolved oxygen (DO) exhibited modest validation skill (NSE = 0.290). Nitrogen parameter refinement improved correlation direction and simulated variability, yet the simulated seasonal TN peak remained five months earlier than observed. Total nitrogen (TN), total phosphorus (TP), and chlorophyll a (Chl-a) all retained negative validation NSE values. In particular, Chl-a calibration remained inadequate (calibration NSE = −0.405; validation NSE = −0.481), and its dedicated calibration stage retained the incoming seed parameters. Only 0.50–0.78% of evaluated candidates satisfied the cross-variable calibration guardrails. Temporally constant nutrient boundaries, coarse TP reporting resolution, and frequent numerical state repairs highlight key priorities for model improvement. Overall, the current evidence supports retrospective near-surface thermal diagnosis and conditional interpretation of broad seasonal DO patterns, whereas reliable nutrient and algal prediction will require time-varying inflow and nutrient chemistry, process-level field observations, and independent verification of seasonal timing and spatial representativeness.

**Keywords:** aquatic ecosystem model; calibration diagnostics; ecological prediction; GOTM–FABM–WET; model evaluation; parameter sensitivity; shallow lake; sustainable lake management

Academic Editor: Giovanni De Feo

Received: 8 October 2026
Revised: 2 November 2026
Accepted: 3 November 2026
Published: 5 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

Aquatic ecosystem models are widely used to diagnose lake dynamics, test restoration hypotheses, and evaluate environmental management scenarios. These applications directly support United Nations Sustainable Development Goal (SDG) 6, particularly Target 6.3 on improving water quality and reducing pollution and Target 6.6 on protecting and restoring water-related ecosystems, including lakes [1]. In practice, however, the utility of any model depends on whether its predictions are empirically supported for the intended decision context. Greater mechanistic detail does not automatically confer predictive reliability: a coupled model can capture a dominant physical signal while missing the seasonal timing, amplitude, or biogeochemical feedbacks that govern ecological responses. Recent syntheses therefore emphasize aligning model complexity and evaluation with the specific management question, observational coverage, and resolved spatiotemporal scales [2,3]. Such alignment is especially critical when projecting responses to nutrient-load reductions, climate warming, or regime shifts, where poorly constrained boundary inputs, process parameterizations, or calibration records can yield projections that appear precise yet remain weakly grounded [4,5].

Parameter calibration is necessary but cannot substitute for rigorous model diagnosis. Standard goodness-of-fit metrics capture distinct and occasionally competing aspects of model behavior: root mean squared error (RMSE) penalizes large residuals, signed bias measures systematic offset, the coefficient of determination ($R^2$) reflects linear association rather than 1:1 agreement, and Nash–Sutcliffe efficiency (NSE) benchmarks squared prediction error against the observed mean [6,7]. Consequently, a parameterization that minimizes aggregate error can still dampen observed variance or match the annual mean while misrepresenting the timing of the seasonal peak. Expanding the number of free parameters may lower an optimization objective, but when observations are sparse relative to model complexity, it also widens the subspace of compensating parameter combinations (equifinality) [8,9]. Global sensitivity analysis is therefore most informative when conducted prior to calibration, targeted at decision-relevant diagnostics, and paired with explicit parameter bounds, out-of-sample validation, and uncertainty checks [10,11]. For nonlinear ecological models, moment-independent Delta sensitivity analysis is particularly well suited because it quantifies shifts across the entire output distribution rather than variance alone [12,13].

Evaluating aquatic ecosystem models also requires looking beyond pointwise agreement in state variables. Hipsey et al. [14] formulated a Concept–State–Process–System evaluation hierarchy that distinguishes conceptual plausibility, state-variable fit, process-rate fidelity, and emergent system-level behavior. Incorporating system-level diagnostics into calibration can expose structural or dynamic flaws that remain masked when objective functions are dominated by strongly seasonal physical variables or uneven sampling density [15]. From an environmental-observatory perspective, model–data discrepancies serve as diagnostic signals to refine boundary forcing, monitoring design, parameterizations, or model structure rather than errors to be masked through aggressive tuning [16]. Model adequacy is thus inherently conditional: a given configuration may be well suited for retrospective thermal diagnosis yet inadequate for nutrient forecasting or setting management thresholds [17].

The coupled GOTM–FABM–WET modeling framework is well suited to such diagnostic evaluation. Within this modular architecture, GOTM resolves one-dimensional (1D) vertical hydrodynamics, FABM mediates two-way coupling between physical and biogeochemical modules, and WET simulates pelagic and benthic nutrient cycling, detritus, phytoplankton functional groups, submerged macrophytes, and food-web interactions within the PCLake model lineage [18–20]. Ensemble PCLake experiments have likewise shown that projected lake responses to climate and nutrient scenarios

depend strongly on parameter and forcing uncertainties [21]. Prior coupled applications illustrate the framework's capabilities when supported by comprehensive field records. In Lake Hinge, Denmark, Andersen et al. [22] leveraged a 14-year dataset covering inorganic nutrients, phytoplankton, zooplankton, and macrophytes to sequentially calibrate more than 140 parameters before analyzing ecosystem state transitions. In Jihongtan Reservoir, China, Chou et al. [23] calibrated 124 parameters across physical, dissolved oxygen, nitrogen, phosphorus, and plankton stages using time-varying inflow hydrology and nutrient loads over an eight-year calibration window and a two-year validation window. More recently, a GOTM–WET application to Meiliang Bay in Lake Taihu combined in situ water temperature, TN, and TP records with MODIS-derived Chl-a, applying Delta sensitivity screening with a dummy parameter across 220,761 simulations [24]. While these studies demonstrate the value of staged calibration, adopting their parameter sets or objective functions cannot compensate for missing boundary forcing or sparse process observations in a data-limited lake.

Lake morphometry and mixing regime further shape parameter sensitivity. Because parameter rankings shift with depth, surface area, underwater light climate, and thermal stratification, sensitivity hierarchies derived for a deep stratified reservoir cannot be assumed to transfer directly to a shallow polymictic lake [25]. Although 1D vertical models can resolve transient stratification and mixing events, monthly sampling at a single near-surface depth provides little direct constraint on vertical eddy diffusivity, benthic–pelagic fluxes, or trophic interactions. Consequently, studies equipped with high-frequency forcing and depth-resolved profiles typically rely on large ensemble sizes and system-level metrics to disentangle physical transport from biological turnover [26]. The fundamental question is whether the available boundary inputs and field observations contain sufficient information to identify the processes governing the target prediction—a question that cannot be answered by reducing optimization error alone.

Kun Cheng Lake in Jiangsu Province, China, is a shallow urban lake (maximum depth ~8 m; mean depth ~4 m) with no documented persistent year-round thermal stratification. In our initial coupled configuration, the model captured near-surface water temperature accurately and dissolved oxygen (DO) moderately well, but exhibited poor seasonal agreement for total nitrogen (TN), total phosphorus (TP), and chlorophyll a (Chl-a). This divergence motivated a structured diagnostic investigation to determine what scope of model application is genuinely supported by the available data. For sustainable lake management, the central question is whether staged parameter refinement under sparse forcing and monthly surface monitoring can justify predictive ecological scenario analysis. Specifically, we addressed four questions: (1) Which internal nitrogen, phosphorus, and phytoplankton parameters govern state error, seasonal correlation, and amplitude? (2) Can sensitivity-guided sequential calibration improve ecological phase and out-of-sample validation performance without degrading water temperature or DO? (3) What constraints in the parameter search, boundary forcing, observational resolution, and numerical integration remain unresolved? (4) What domain of model applicability is warranted by the resulting evidence? By reporting both calibration gains and persistent predictive failures, we make the evidentiary limits of the model explicit and define the monitoring and analytical steps required before complex ecosystem models are used to guide lake restoration or water-quality policy.

## 2. Materials and Methods

### *2.1. Study System and Intended-Use Definition*

Kun Cheng Lake is located immediately south of urban Changshu in Jiangsu Province, eastern China, and is the municipality's largest lake, with an officially reported water surface area of approximately 18.4 km$^2$ [27]. Previous local modeling studies examined horizontal circulation and pollutant transport in this shallow urban basin [28,29]. Situated within the lower Yangtze alluvial lake plain, Kun Cheng Lake shares the region's characteristic shallow morphometry, nutrient enrichment, and tight benthic–pelagic coupling [30]. The lake is embedded within the dense river network of the Taihu Plain. Earlier hydrodynamic modeling of Kun Cheng Lake indicated generally weak background currents with a net northwest-to-southeast flow tendency [29], while broader Taihu Basin studies demonstrate that monsoon storm runoff and interconnected river channels can drive pronounced temporal fluctuations in inflow and horizontal water-quality gradients [31]. Episodic rainfall–runoff pulses and river–lake exchange are therefore plausible drivers of nutrient variability at the pelagic monitoring station, although the available study archive lacks the event-scale discharge measurements needed to quantify their contribution.

The representative model column was located at 31.58° N, 120.74° E. Basin survey records describe a maximum lake depth of approximately 8 m and a mean depth of approximately 4 m, which we represented using an 8 m vertical water column discretized into 16 layers. The supplied hypsographic file (KCH02A/hypsograph.dat) prescribes a constant horizontal area of 15,632,844 m$^2$ from 0 to 8 m depth, yielding a modeled volume of 125,062,752 m$^3$. This idealized box geometry also defines the hydrological–nutrient mass-balance benchmark in Supplementary Methods S6. The 15.63 km$^2$ model plan area and the approximately 18.4 km$^2$ municipal surface-area estimate represent the idealized 1D column footprint and the full basin area, respectively. Horizontal representativeness could not be evaluated against multi-station field surveys or a spatially resolved hydrodynamic model. Furthermore, while the lake does not sustain persistent annual stratification, monthly near-surface measurements cannot distinguish continuous polymictic mixing from transient diurnal or synoptic stratification events; we therefore classify the system simply as shallow and non-permanently stratified. Prior to calibration, we ranked prospective model applications in order of increasing evidentiary demand: (i) retrospective near-surface thermal diagnosis, (ii) seasonal DO interpretation, and (iii) nutrient and Chl-a scenario projection—the latter contingent on successful out-of-sample validation. Accordingly, we evaluate these foundational capabilities here without conducting climate-warming, nutrient-reduction, regime-shift, or management-threshold projections.

### *2.2. Observations, Forcing, and Data-Quality Audit*

Monthly field monitoring encompassed water temperature, DO, TN, TP, and Chl-a. Water temperature and Chl-a were sampled at 0.2 m depth, whereas DO, TN, and TP were sampled at 0.5 m depth. Temperature and DO were measured in situ following China's national standard thermometer and electrochemical-probe protocols (GB 13195-91 [32] and HJ 506-2009 [33]), respectively. Discrete water samples were collected monthly for laboratory determination: TN via alkaline potassium persulfate digestion and UV spectrophotometry (HJ 636-2012 [34]), TP via ammonium molybdate spectrophotometry (GB 11893-89 [35]), and Chl-a via acetone extraction and spectrophotometry (HJ 897-2017 [36]). Field sampling and sample handling adhered to national surface-water monitoring specifications (HJ/T 91-2002 [37] and HJ 91.2-2022 [38]). Specific instrument models, sample holding times, procedural blanks, laboratory replicate precision, certified

reference recoveries, and batch detection limits were not separately tabulated in the municipal monitoring dataset.

The calibration window spanned 1 January 2021 to 31 December 2023, providing 36 monthly observations per variable (Table 1). For DO, TN, TP, and Chl-a, an additional 24 monthly records from 2024 to 2025 were held out strictly for out-of-sample validation and excluded from all parameter-selection steps. The supplied TN and TP observation files represent the primary field measurements used across all model evaluations; their exact sampling dates, depths, and calibration–validation split are documented in Supplementary Methods S6. Because no water temperature measurements were available after December 2023, thermal evaluation is restricted to the 2021–2023 calibration window. Prior to computing performance metrics, we audited all observation files for duplicate entries (by variable, date, depth, and period), sample size, temporal span, unique value counts, minimum reporting increments, and missing values. Auditing the boundary forcing files revealed 2558 daily records from 1 January 2019 through 31 December 2025, but daily inflow discharge, supplied nitrogen concentration, and supplied phosphorus concentration were each held at a single constant value throughout the seven-year simulation (Table S1). We retained these boundary files as supplied rather than substituting in-lake TN or TP observations for inflow concentrations, which would have introduced circularity into the calibration. Auditing the TP observation record further revealed coarse reporting granularity: all 60 calibration and validation measurements fell into just five discrete concentration levels (0.02–0.06 mg P $L^{-1}$ in 0.01 mg P $L^{-1}$ increments). Across the full 2021–2025 record, observed TN averaged 0.512 mg $L^{-1}$ (median 0.515; range 0.13–0.94 mg $L^{-1}$) and observed TP averaged 0.0422 mg $L^{-1}$ (median 0.040; range 0.02–0.06 mg $L^{-1}$), consistent with the 2024 lake-center mean TP of 0.040 mg $L^{-1}$ reported by the Changshu Municipal People's Government [27]. Because monthly sampling was not paired with antecedent precipitation, tributary hydrographs, or storm-event nutrient loads, individual in-lake concentration excursions cannot be attributed to specific hydrological events.

**Table 1.** Summary of field observations used for calibration (2021–2023) and 2024–2025 out-of-sample evaluation. Sampling depth is expressed as negative downward following the model coordinate convention.

| Variable | Unit | Cal. *n* | Val. *n* | First Date | Last Date | Depth (m) | Unique Values |
|---|---|---|---|---|---|---|---|
| WT | °C | 36 | 0 | 13 January 2021 | 17 December 2023 | −0.2 | 36 |
| DO | mg $L^{-1}$ | 36 | 24 | 1 January 2021 | 01 December 2025 | −0.5 | 42 |
| TN | mg $L^{-1}$ | 36 | 24 | 1 January 2021 | 01 December 2025 | −0.5 | 33 |
| TP | mg $L^{-1}$ | 36 | 24 | 1 January 2021 | 01 December 2025 | −0.5 | 5 |
| Chl-a | μg $L^{-1}$ | 36 | 24 | 1 January 2021 | 01 December 2025 | −0.2 | 34 |

### *2.3. Coupled Model Configuration*

Our modeling framework coupled the General Ocean Turbulence Model (GOTM v4.1.0) to the Water Ecosystems Tool (WET v2.3) via the Framework for Aquatic Biogeochemical Models (FABM v3.0.1). Within this system, GOTM resolves 1D vertical hydrodynamics and turbulent mixing, FABM provides the standardized coupling interface between physical and biogeochemical state variables, and WET simulates pelagic and benthic nutrient pools, detritus, three phytoplankton functional groups, dissolved oxygen, submerged macrophytes, and higher trophic levels [18–20]. The 8 m

water column was discretized into 16 uniform vertical layers and integrated at a 600 s time step with daily diagnostic output from 1 January 2019 to 1 January 2026, treating 2019–2020 as a two-year spin-up window (without formal spin-up equilibrium testing). Simulated state variables were linearly interpolated in time and depth to match the exact coordinates of each field observation. Model evaluation focused on five state outputs: water temperature, DO, TN, TP, and Chl-a. While this coupled architecture enables joint evaluation of physical and biogeochemical dynamics, resolving internal ecological processes in the code does not guarantee that those processes are empirically identifiable from monthly surface observations.

The architecture of the coupled system builds directly on the established PCLake model lineage. Whereas classical PCLake represents a shallow lake as a single well-mixed water column coupled to a benthic sediment layer and a food web comprising phytoplankton, zooplankton, macrophytes, and fish [21], FABM-PCLake reformulates these processes into nine modular components within FABM: abiotic water, abiotic sediment, pelagic phytoplankton, benthic phytoplankton, macrophytes, zooplankton, fish, zoobenthos, and auxiliary diagnostics [23]. Coupling FABM-PCLake to GOTM introduces vertical turbulent transport of momentum, heat, and dissolved and particulate constituents across depth layers [22,23,25], as demonstrated in Lake Hinge and Jihongtan Reservoir [22,23]. WET further generalizes FABM-PCLake into a fully modular food-web and biogeochemical configuration [20]; accordingly, we refer to our implementation as GOTM–FABM–WET throughout this study. Full configuration settings for Kun Cheng Lake are detailed in Table S2, and Supplementary Methods S5 documents the software provenance and public code repositories.

Table S3 itemizes the simulation counts across the sensitivity-screening and sequential-calibration stages. Table S2 consolidates the software provenance and complete physical, numerical, and ecological configuration—including vertical discretization, time-stepping, advection and ODE integration schemes, boundary condition mappings, biophysical feedback switches, numerical safeguards, and initial state concentrations—while Table S4 lists the declared bounds, initial seeds, and accepted values for all 29 calibrated parameters. Supplementary Methods S5 links this configuration record to the archival software releases and to the screening, calibration, and benchmark protocols detailed in Sections 2.4–2.6 and Supplementary Methods S1–S8. Together, these disclosures provide a transparent, self-contained basis for reproducing our model configuration and diagnostic workflow.

### *2.4. Candidate-Parameter Inventory and N Screening*

We compiled a candidate inventory of 62 biogeochemical process parameters drawing on the local WET configuration, study calibration records, and prior PCLake sensitivity and calibration studies [22–25,39]. For each parameter, we documented the associated model module, ecological process, physical units, sampling scale (linear or logarithmic), lower and upper bounds, and literature provenance. Rather than calibrating all 62 parameters simultaneously, we used this inventory as a structured pool for staged screening. External boundary forcings, fixed stoichiometric constants, and initial state concentrations were excluded from calibration because the available field data provided no independent basis for estimating them.

Nitrogen sensitivity screening began with 37 internal N-cycle parameters governing organic-matter mineralization, nitrification, denitrification, benthic–pelagic pore-water diffusion, algal N uptake, cyanobacterial N quotas, mortality recycling, and associated Arrhenius temperature multipliers. Each parameter was perturbed one at a time (OAT) to its declared lower and upper bounds while holding all other parameters at their incoming baseline values, yielding 74 two-bound OAT simulations. One organic-detritus settling

parameter was excluded from subsequent global sensitivity analysis because at least one boundary perturbation failed to return all 36 monthly calibration pairs. From the remaining valid parameters, the 18 controls exhibiting the largest OAT composite score ranges were advanced to a 256-point maximin Latin-hypercube global sensitivity design.

Moment-independent Delta indices quantify the expected shift in the full probability distribution of a model output when an input parameter is fixed across its range [12,13]. We computed Delta indices across four calibration-period TN diagnostics: normalized root mean squared error (RMSE), correlation loss, amplitude loss, and a weighted combined dynamic loss. Whereas the historical screening analysis used a single 256-point Latin-hypercube design across the 18 N parameters with 400 bootstrap resamples, we added a formal sample-size convergence and across-design repeatability assessment using two independent 75-point maximin Latin-hypercube realizations (A75 and B75) executed within a version-locked GOTM–FABM–WET reconstruction. Nested paired-prefix pooled samples of $n$ = 50, 100, and 150 completed runs were complemented by 100 fixed-seed random permutations evaluated at cumulative sample sizes of $n$ = 50, 75, 100, 125, and 150 to isolate sample-composition variability. Using a SALib 1.3.8-compatible Borgonovo Delta estimator with a 100-point quadrature grid, 400 paired-row bootstrap resamples, and a dummy-variable screening threshold, we evaluated rank and magnitude convergence via maximum absolute Delta shift ($E$), Spearman's rank correlation (*rho*), maximum rank displacement ($H$), Top-5 retention, and screened-set Jaccard overlap ($J$), following convergence-guided lake-modeling practice [26]. Full experimental protocols, random seeds, and parameter-level stability metrics are presented in Supplementary Methods S7 and Tables S5–S8, while the historical 256-point estimates are retained in Table S9.

### *2.5. Sequential Multi-Parameter Calibration*

Table S4 details the process roles, declared bounds, incoming baseline values, and final accepted values for the 29 parameters varied across the three sequential-calibration stages. Rather than estimating all 29 parameters simultaneously, we followed the staged, bottom-up calibration sequence established in prior coupled PCLake studies [22,23], calibrating a targeted subset of parameters at each stage while enforcing strict cross-variable guardrails to prevent degradation of previously fitted state variables (this sequential strategy was not benchmarked against simultaneous multi-objective optimization). In the first stage (TN), we evaluated 336 candidate parameter sets across the 12 nitrogen parameters that exceeded the combined dynamic-loss dummy threshold. In the second stage (TP), we evaluated 128 candidates across seven parameters governing sediment and pelagic phosphate sorption, iron-binding capacity, redox sensitivity, maximum pore-water phosphate concentration, and molecular diffusion; interpretation of this stage is constrained by the coarse 0.01 mg $L^{-1}$ reporting increment of the TP observations (which reflects record granularity rather than analytical uncertainty or an optimization tolerance). In the third stage (Chl-a), we evaluated 200 candidates across ten phytoplankton growth, respiration, mortality, settling, and chlorophyll-to-biomass quota parameters; because no candidate improved the Chl-a objective while satisfying all cross-variable guardrails, the ten Chl-a parameters retained their incoming seed values. Candidate populations were generated via differential evolution [40] using population sizes and generation counts of 48 × 6 for TN, 32 × 3 for TP, and 40 × 4 for Chl-a. All candidate selection relied exclusively on the 2021–2023 calibration observations, keeping the 2024–2025 validation records strictly out of sample.

Across all three stages, candidate acceptance required that water temperature and DO RMSE not exceed 101% of their pre-refinement baseline values and that cumulative FABM state-repair counts not exceed 110% of baseline, alongside stage-specific guardrails

on non-target ecological variables (Figure 1). The 1% allowance on temperature and DO RMSE served as an operational design tolerance to preserve physical and oxygen skill rather than a statistically derived threshold. In the TN stage, feasible candidates were also required to exhibit a positive signed correlation ($r > 0$), nonzero simulated variance, and a TN RMSE ≤102% of baseline, ranked by a composite objective weighting normalized RMSE (0.45), correlation loss (0.40), and absolute log standard-deviation ratio error (0.15). The TP stage weighted normalized RMSE (0.80), correlation loss (0.10), and amplitude loss (0.10) while permitting TP RMSE up to 110% of baseline to accommodate low ambient P concentrations. The Chl-a stage weighted normalized RMSE (0.50), correlation loss (0.40), and amplitude error (0.10) while additionally requiring preservation of the accepted TN-stage solution (Supplementary Methods S1). In total, the sequential calibration evaluated 664 native-model candidates (336 TN, 128 TP, and 200 Chl-a), reported separately from the 74 OAT screening runs and the 256-point Delta design (Figure 1; Table S3); the 400 bootstrap replicates represent statistical resamples rather than additional model integrations.

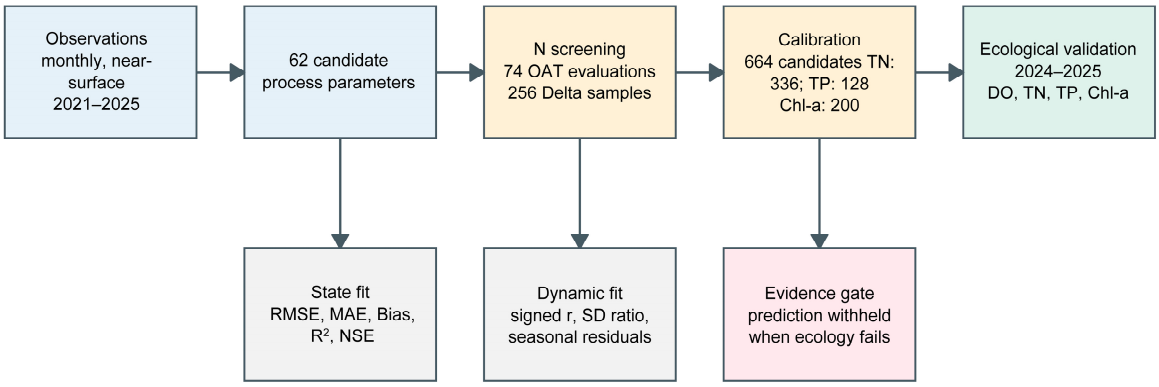



**Figure 1.** Sequential calibration and diagnostic evaluation workflow. Nitrogen sensitivity screening comprised 74 two-bound one-at-a-time (OAT) runs and a 256-point Latin-hypercube Delta design, followed by 664 sequential-calibration evaluations across TN (336), TP (128), and Chl-a (200). These stage counts represent distinct analytical components rather than a cumulative sum across preliminary exploratory trials. Technical completion, pointwise state fit, seasonal dynamic fidelity, and suitability for prospective management prediction were evaluated as distinct diagnostic gates. Abbreviations: MAE, mean absolute error; RMSE, root mean squared error; NSE, Nash–Sutcliffe efficiency; OAT, one-at-a-time; SD, standard deviation.

### *2.6. Performance and Dynamic-Fidelity Metrics*

Pointwise state-variable agreement was quantified using sample size ($n$), RMSE, MAE, signed bias (simulated minus observed), $R^2$, and NSE [6,7]. We report the signed Pearson correlation coefficient ($r$) alongside $R^2$ to preserve the direction of linear association, which is lost upon squaring. Seasonal dynamic fidelity was assessed using the simulated-to-observed standard-deviation (SD) ratio, the correlation between observed and simulated 12-month climatologies (climatological $r$), observed and simulated peak months, the circularly wrapped peak-month phase lag, and seasonal mean residual biases. Together, these diagnostics separate aggregate concentration error from errors in seasonal timing and amplitude. Given the modest sample sizes and serial autocorrelation of monthly lake observations, we refrain from reporting $p$-values or null-hypothesis significance tests.

To contextualize coupled-model skill during the 2024–2025 validation period, we defined three statistical benchmarks a priori: (i) the constant calibration-period mean, (ii) the 12 calendar-month calibration means (monthly climatology), and (iii) a first-order

annual harmonic fitted via ordinary least squares with an intercept and one sine–cosine pair [41]. Each statistical benchmark was parameterized exclusively on the 36 monthly calibration records (2021–2023) and evaluated out of sample on the 24 validation records (2024–2025) for DO, TN, TP, and Chl-a (water temperature lacked post-2023 observations). Observation counts, sampling depths, calendar-month means, and sample variances were held identical across all benchmark evaluations. Benchmark RMSE, MAE, signed bias, and NSE are evaluated on the exact same 2024–2025 validation records and compared directly against the validation scores of the accepted coupled model (Supplementary Methods S4; Table S10).

A low-complexity hydrological–nutrient mass-balance benchmark was also formulated to represent the lake as a constant-volume, completely mixed reactor subject to external nutrient loading, advective flushing, and first-order net nutrient removal [42,43]. Governed by $dV/dt = Q_{in} - Q_{out} = 0$ and $V\,dC/dt = Q(C_{in} - C) - kVC$ for TN and TP separately, the model used the exact KCH02A column volume ($V$ = 125,062,752 m$^3$), balanced inflow and outflow ($Q$ = 86,400 m$^3$ d$^{-1}$), and constant inflow concentrations (2.00 mg N L$^{-1}$ and 0.12 mg P L$^{-1}$). Here, the lumped first-order coefficient $k$ integrates net water-column retention and permanent loss pathways such as sediment burial and denitrification [44]. For each nutrient, a single nonnegative $k$ was fitted by ordinary least squares across the 36 monthly calibration records (2021–2023), initialized at the first calibration observation, and propagated analytically through the 2024–2025 validation window. Both the mass-balance and coupled models were evaluated against the exact same field measurements and period split from the supplied OBS files (Supplementary Methods S6; Table S11).

Stage-specific candidate counts and pass fractions were compiled across the three sequential-calibration stages. Using the complete 336-candidate performance table from the TN stage, we conducted a deterministic post-processing sensitivity analysis by relaxing the shared water-temperature and DO RMSE allowances from 101% to 105% and 110% of the pre-refinement baseline while holding all other guardrails constant (Supplementary Methods S3; Table S12). This analysis tests whether the strict 1% physical/DO tolerance was the primary bottleneck rejecting TN candidates within the evaluated population, without rerunning differential evolution across the subsequent TP and Chl-a stages. Reported pass rates describe the empirical fraction of evaluated candidates satisfying the guardrails rather than the volumetric fraction of feasible parameter space. Finally, we examined the domain-specific FABM state-repair counters and verified the normalized positions of all 29 accepted parameters within their declared bounds. Inspection of the public GOTM–FABM source code confirmed the counting convention: each out-of-bounds state check within the water-column interior or surface array increments its respective counter once, whereas benthic state checks are counted at each active bottom layer, with a single domain check potentially clipping multiple state variables simultaneously. Across all 336 evaluated TN candidates, all four counter fields were complete and satisfied the exact identity total = interior + surface + bottom with zero discrepancy (Supplementary Methods S5).

Throughout our evaluation, a negative NSE value indicates that the model trajectory yielded a larger mean squared residual than a constant predictor equal to the observed period mean.

### *2.7. Evidence Classification and Reproducibility*

We synthesized model performance using an evidence hierarchy adapted from contemporary aquatic model evaluation frameworks [14,15], separating conceptual plausibility, pointwise state fit, seasonal and process dynamics, out-of-sample validation, and fitness for intended management use. High parameter sensitivity in Delta screening or responsiveness during calibration reflects conditional model-output dependence across the sampled bounds rather than field-verified process dominance or unique parameter identifiability. Similarly, constant boundary inputs and the absence of in situ process-rate measurements are treated explicitly as observational constraints. While our controlled seasonal nitrogen-loading experiments isolate boundary-timing sensitivity within a version-locked model reconstruction, attributing the historical residuals requires weighing boundary forcing, structural dimensionality, and numerical state interventions together. Tables S1–S4 and Supplementary Methods S1–S8 document the software provenance, model configuration, and evaluation protocols needed to interpret and reproduce the workflow.

For future monthly predictive use, we establish an explicit, study-specific screening rule requiring four simultaneous criteria: (1) independent out-of-sample temporal validation, (2) validation NSE > 0.3, (3) an absolute circular peak-month phase lag of at most one month, and (4) a validation RMSE lower than that of a calibration-fitted calendar-month climatology evaluated on the same out-of-sample records. While these operational criteria govern monthly concentration forecasting, event-scale and management-scenario projections additionally require direct observations of the relevant biogeochemical fluxes and response timescales.

### *2.8. Controlled Seasonal Nitrogen-Loading Experiment*

To test whether absent seasonal boundary forcing alone could explain the simulated TN phase error, we conducted five controlled native GOTM–FABM–WET simulations within the version-locked model reconstruction used for the Delta convergence assessment (Supplementary Methods S7). We compared the constant-nitrogen baseline against four idealized sinusoidal inflow N regimes peaking in winter, spring, summer, and autumn with a ±50% concentration amplitude, while holding the annual mean N load, water discharge, phosphorus loading, calibrated ecological parameters, initial conditions, meteorological forcing, and numerical settings strictly identical. Simulations spanned 2019–2025, and simulated TN at 0.5 m depth was linearly interpolated to the 60 monthly observation dates during 2021–2025 to evaluate climatological peak month, signed circular phase offset, seasonal range, and annual harmonic semi-amplitude (see Supplementary Methods S8 for the load-conserving formulation and verification).

## 3. Results

### *3.1. Observation and Forcing Inventory*

The audited observation dataset comprised 36 monthly calibration measurements (2021–2023) for each of the five state variables and 24 held-out evaluation records (2024–2025) for DO, TN, TP, and Chl-a (Table 1). All field measurements were collected monthly at a single near-surface depth (0.2 m or 0.5 m). Across the full record, water temperature contained 36 unique values, DO contained 42, Chl-a contained 34, and TN contained 33, whereas TP contained only five discrete values (0.02–0.06 mg P $L^{-1}$). Auditing the boundary forcing files confirmed 2558 daily entries from 2019 through 2025 with zero temporal variation in inflow discharge, supplied TN concentration, or supplied TP concentration (Table S1).

### *3.2. State-Variable Performance and Benchmark Comparisons*

Over the 2021–2023 calibration window, simulated near-surface water temperature achieved an RMSE of 2.224 °C, a mean absolute error (MAE) of 1.797 °C, a signed bias of +0.935 °C, an $R^2$ of 0.969, and an NSE of 0.933 (Table 2). The coupled model accurately captured the July thermal maximum and achieved a monthly climatological correlation of $r$ = 0.989. Residuals were seasonally structured, exhibiting a warm bias in spring (+2.54 °C) and summer (+2.45 °C) and a slight cool bias in autumn (−0.28 °C) and winter (−0.97 °C); thus, strong annual phase agreement coexisted with a warm-season offset. Dissolved oxygen demonstrated moderate, positive predictive skill: calibration RMSE was 1.544 mg $L^{-1}$ ($R^2$ = 0.591; NSE = 0.368), and out-of-sample validation RMSE was 1.316 mg $L^{-1}$ ($R^2$ = 0.458; NSE = 0.290) (Table 2). Simulated and observed DO maxima coincided in February during calibration and differed by one month during validation, with a calibration climatological correlation of $r$ = 0.854. The primary DO discrepancy was a spring overestimation (+2.10 mg $L^{-1}$ in calibration; +1.52 mg $L^{-1}$ in validation) paired with a summer validation underestimation (−0.83 mg $L^{-1}$). By contrast, nutrient and phytoplankton predictions performed poorly. Final calibration $R^2$ and NSE values were 0.021 and −0.156 for TN, 0.007 and −0.983 for TP, and 0.061 and −0.405 for Chl-a (Table 2). Out-of-sample evaluation during 2024–2025 confirmed this persistent ecological mismatch, yielding validation $R^2$ and NSE values of 0.016 and −0.156 for TN, 0.112 and −0.025 for TP, and 0.164 and −0.481 for Chl-a (Figure 2; Table 2).

**Table 2.** Pointwise performance metrics of the accepted model refinement. Signed bias is defined as simulated minus observed; water temperature lacks post-2023 validation observations.

| Variable | Period | *n* | RMSE | MAE | Bias | $R^2$ | NSE |
|---|---|---|---|---|---|---|---|
| WT | Calibration | 36 | 2.224 | 1.797 | 0.935 | 0.969 | 0.933 |
| DO | Calibration | 36 | 1.544 | 1.305 | 0.866 | 0.591 | 0.368 |
| DO | Validation | 24 | 1.316 | 1.069 | −0.0082 | 0.458 | 0.29 |
| TN | Calibration | 36 | 0.194 | 0.153 | 0.04 | 0.021 | −0.156 |
| TN | Validation | 24 | 0.156 | 0.137 | 0.021 | 0.016 | −0.156 |
| TP | Calibration | 36 | 0.013 | 0.01 | −0.0061 | 0.0071 | −0.983 |
| TP | Validation | 24 | 0.0094 | 0.0075 | 0.0001 | 0.112 | −0.025 |
| Chl-a | Calibration | 36 | 8.933 | 6.385 | −3.293 | 0.061 | −0.405 |
| Chl-a | Validation | 24 | 7.146 | 5.862 | −3.148 | 0.164 | −0.481 |

During calibration, signed Pearson correlations ($r$) for the accepted refinement were +0.147 for TN, −0.084 for TP, and +0.248 for Chl-a (Figure 3b; Figure S1), accompanied by simulated-to-observed SD ratios of 0.506, 0.638, and 0.773, respectively (Figure 3b). Thus, simulated TN captured roughly half of the observed temporal dispersion, TP retained a weak inverse association, and Chl-a exhibited weak positive correlation despite a closer variance ratio.

The benchmark comparison gave distinct, variable-specific outcomes across the 2024–2025 validation period (Table S10). For DO, the coupled model's validation RMSE (1.316 mg $L^{-1}$) outperformed the calibration-period mean (1.693 mg $L^{-1}$) but slightly trailed monthly climatology (1.197 mg $L^{-1}$) and the annual harmonic (1.289 mg $L^{-1}$). For TN, all three statistical benchmarks achieved lower validation RMSE (0.1448, 0.0577, and 0.1395 mg $L^{-1}$) than the coupled model (0.1556 mg $L^{-1}$); notably, calendar-month climatology attained a validation NSE of 0.841 compared with −0.156 for the coupled model. Conversely, the coupled model achieved a lower validation RMSE for TP (0.00939 mg $L^{-1}$) than all three statistical benchmarks (0.01088–0.01249 mg $L^{-1}$), although NSE remained negative across all four models. For Chl-a, both the calibration mean (5.881 μg $L^{-1}$) and annual harmonic (5.738 μg $L^{-1}$) yielded lower validation RMSE than the coupled model

(7.146 μg $L^{-1}$), whereas monthly climatology had a higher RMSE (7.581 μg $L^{-1}$). Overall, the coupled mechanistic model did not hold a consistent predictive advantage over simple statistical predictors.

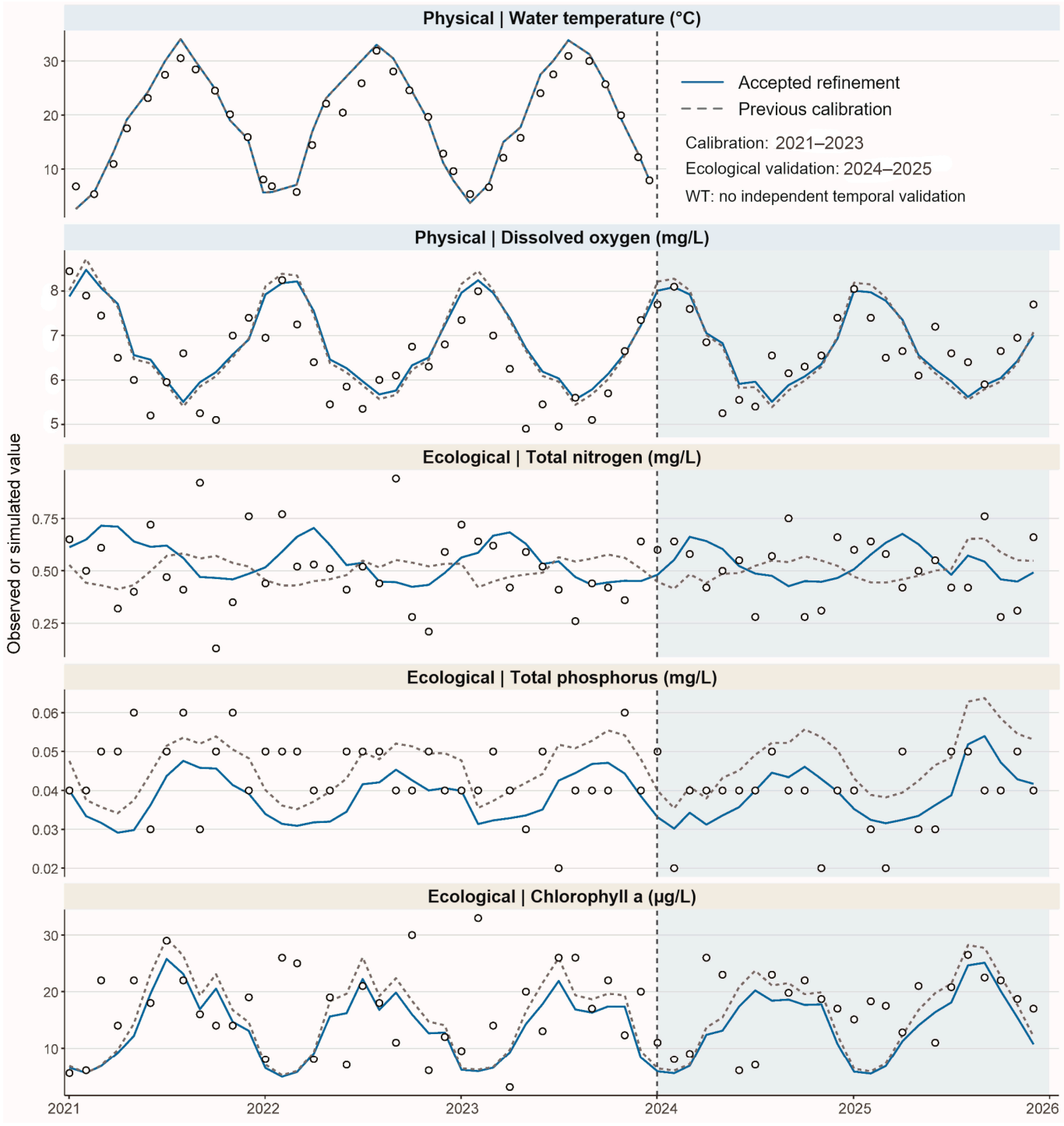



**Figure 2.** Observed and simulated time series of water temperature (WT, °C), dissolved oxygen (DO, mg $L^{-1}$), total nitrogen (TN, mg $L^{-1}$), total phosphorus (TP, mg $L^{-1}$), and chlorophyll a (Chl-a, μg $L^{-1}$) before and after sequential parameter refinement. Open circles denote monthly field observations; grey dashed and solid blue curves represent the previous calibration and accepted refinement, respectively. The vertical dashed line and shaded background demarcate the calibration window (2021–2023) and out-of-sample ecological evaluation window (2024–2025). Post-2023 observations were unavailable for water temperature.

The mass-balance comparison also showed a contrasting, nutrient-specific return on added model complexity (Table S11). For TN, the parsimonious mass-balance model and the coupled model yielded validation RMSEs of 0.15006 and 0.15564 mg $L^{-1}$ (NSE = −0.075 and −0.156), respectively, corresponding to a 7.58% higher mean squared error in the coupled model. For TP, validation RMSEs were 0.01112 and 0.00939 mg $L^{-1}$ (NSE = −0.438 and −0.025), representing a 28.68% reduction in mean squared error and a reduction in mean signed bias from +0.00615 to +0.00010 mg $L^{-1}$ in the coupled model. Evaluated on identical out-of-sample field records, resolving coupled biogeochemical processes reduced aggregate concentration error for TP whereas the simpler loading–flushing–removal balance proved more accurate for TN.

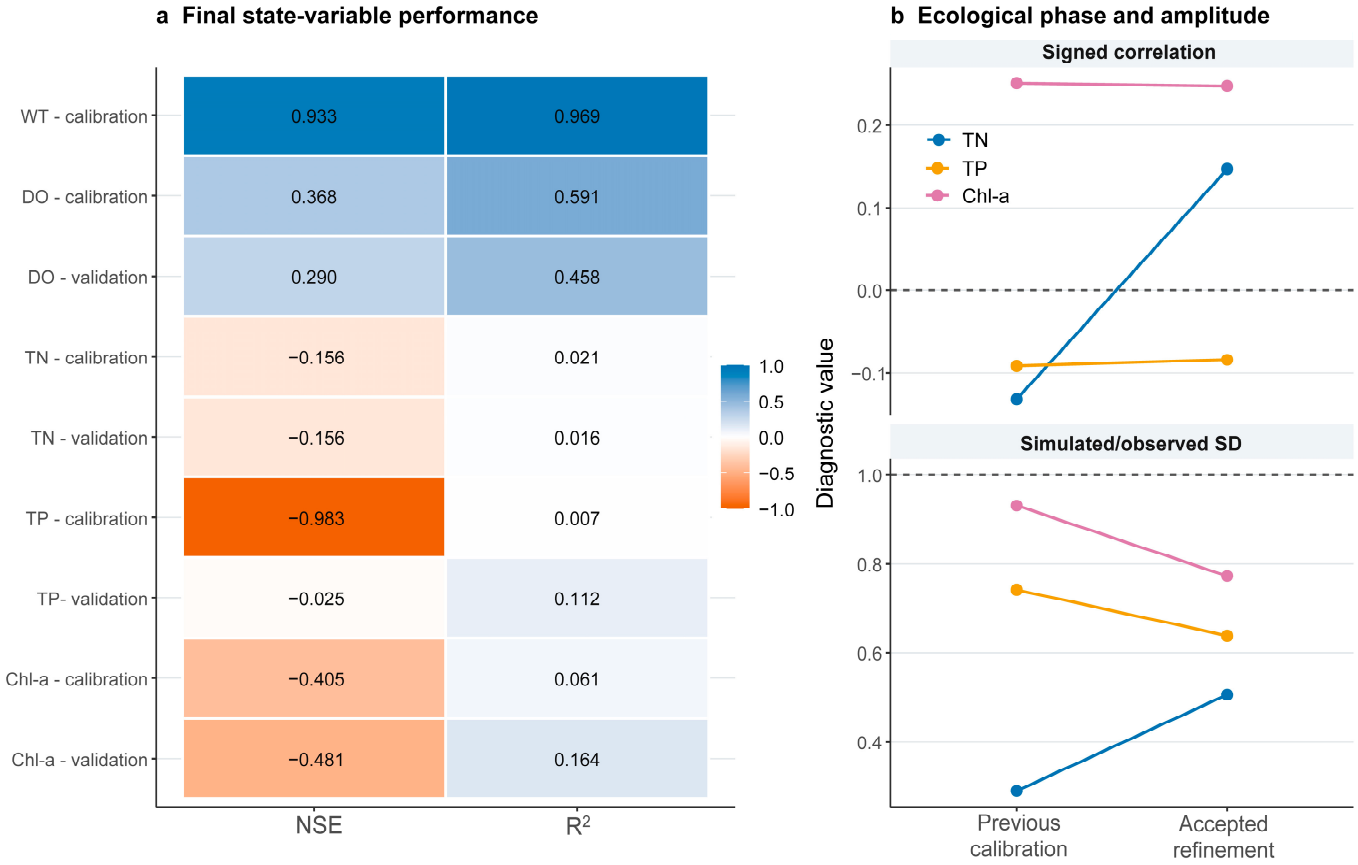



**Figure 3.** Summary of pointwise state-variable skill and ecological dynamic fidelity. (**a**) Nash–Sutcliffe efficiency (NSE) and coefficient of determination ($R^2$) across calibration (2021–2023) and validation (2024–2025). (**b**) Signed Pearson correlation ($r$) and simulated-to-observed standard-deviation (SD) ratio for TN, TP, and Chl-a before and after the accepted parameter refinement. Horizontal dashed lines denote $r = 0$ and a 1:1 SD ratio.

### *3.3. Delta Sensitivity and Ranking Stability*

Following two-bound OAT screening across all 37 candidate nitrogen parameters, the 18 most influential valid controls were evaluated in a 256-point maximin Latin-hypercube Delta sensitivity design. Twelve nitrogen parameters exceeded the archived dummy-variable threshold for the combined dynamic-loss diagnostic (Figure 4; Table S9); this 12-parameter subset defined the active search space for TN calibration and was subsequently evaluated for sample-size convergence. In the archived 256-point design, normalized TN RMSE was most sensitive to the sediment detritus mineralization rate (*kDMinDetS*, Delta = 0.0344) and sediment nitrification rate (*kNitrS*, 0.0335), followed by the refractory fraction of sediment organic matter (*fRefrDetS*, 0.0286), molecular

ammonium diffusion (*kNDifNH4S*, 0.0280), and the bioturbation multiplier for pore-water diffusion (*cTurbDifNutS*, 0.0272) (Figure 4; Table 3). Molecular nitrate diffusion (*kNDifNO3S*), pelagic and benthic mineralization temperature coefficients (*cThetaMinW*, *cThetaMinS*), effective diffusion depth (*fDepthDifS*), pore-water diffusion temperature dependence (*cThetaDifS*), pelagic detritus mineralization (*kDMinDetW*), and maximum cyanobacterial N quota (*cNDBlueMaxW*) completed the 12-parameter set.

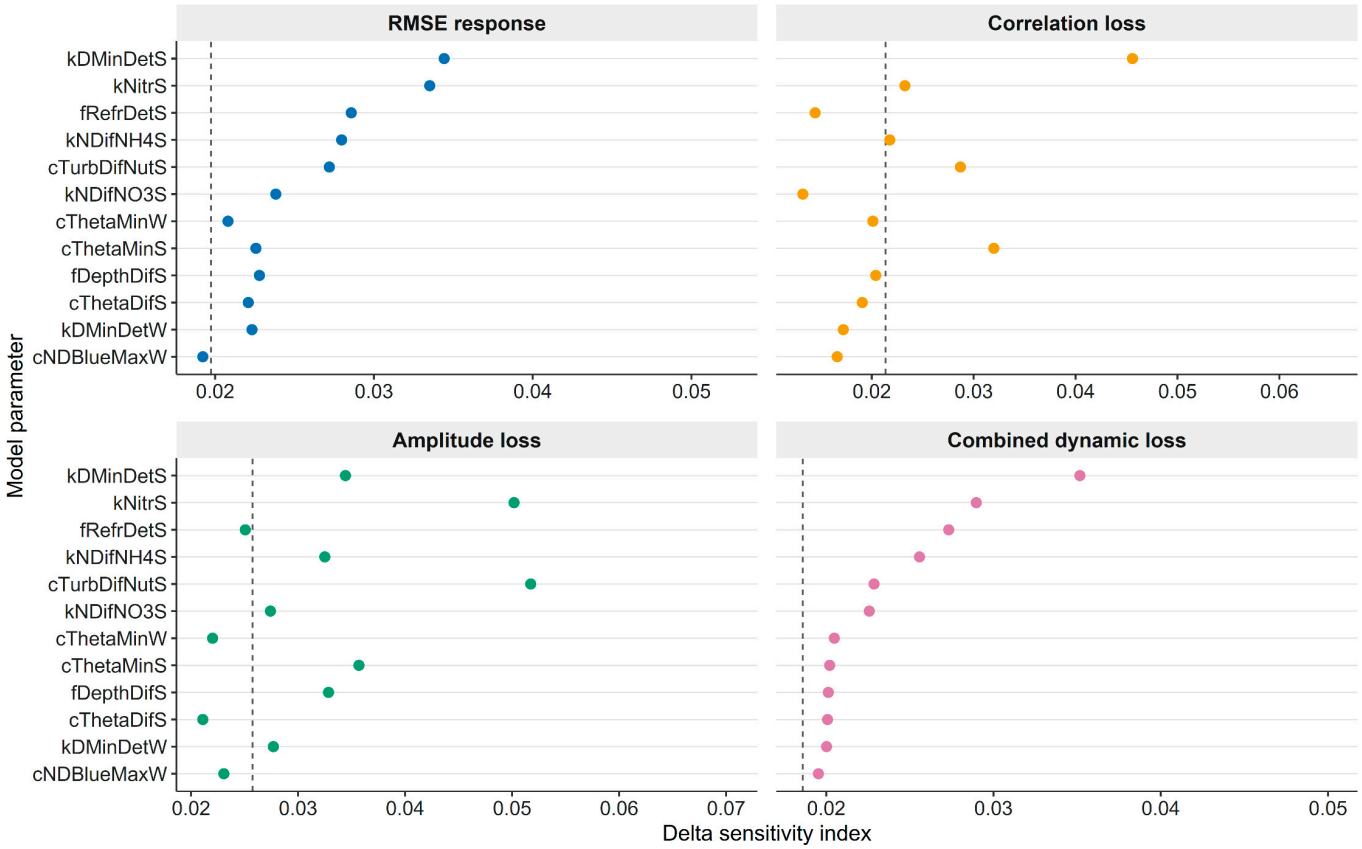



**Figure 4.** Moment-independent Delta sensitivity point estimates for the 12 nitrogen parameters exceeding the combined dynamic-loss screening threshold in the 256-point Latin-hypercube design. Vertical dashed lines indicate the metric-specific dummy thresholds across normalized TN RMSE, correlation loss, amplitude loss, and combined dynamic loss. Delta is dimensionless and quantifies distributional sensitivity across the prescribed parameter bounds rather than variance partitioning or effect sign. Nominal 95% bootstrap bounds from the 256-point screening design are documented in Table S9 and Supplementary Methods S2 and are omitted here.

Parameter rankings shifted across individual diagnostic metrics. Within the 12 retained parameters, *cTurbDifNutS* (0.0517) and *kNitrS* (0.0502) exerted the strongest influence on amplitude loss, whereas *kDMinDetS* dominated correlation loss (0.0456) (Figure 4; Table 3). Several parameters exceeded the dummy threshold for a single diagnostic component without passing the combined dynamic-loss screen, illustrating that parameter sensitivity depends on the chosen evaluation metric. Overall, benthic mineralization, nitrification, and sediment–water diffusive transport emerged as the primary internal controls on simulated TN dynamics within the sampled space.

In the version-locked sample-size convergence assessment for combined dynamic loss ($D$), increasing sample size from P100 to P150 yielded a Spearman rank correlation of $rho$ = 0.5501, a maximum Delta shift of $E$ = 0.07768, a maximum rank displacement of $H$ = 11, a Top-5 retention of 0.80, and a screened-set Jaccard overlap of $J$ = 0.40, while comparing the two independent 75-point realizations (A75 versus B75) yielded $rho$ =

0.3251, $E$ = 0.09497, $H$ = 13, Top-5 retention = 0.40, and $J$ = 1.00 for the two-parameter screened core (Tables S6 and S7). Across 100 random cumulative subsamples, advancing from $n$ = 125 to $n$ = 150 achieved a median *rho* of 0.8720, median $E$ of 0.02894, median $H$ of 7, and median Top-5 retention of 1.00, meeting the strict three-metric joint convergence criterion in 4% of replicates. Crucially, *kDMinDetS* and *kDMinDetW* ranked first and second in 100% of the $n$ = 125 subsamples, while *cTurbDifNutS*, *cThetaDifS*, and *cThetaMinS* appeared in the top five in 93%, 77%, and 77% of subsamples, respectively (Table S8). All 29 accepted parameters across the sequential-calibration stages fell strictly within the interior of their declared bounds (Table S4; Figure S2).

**Table 3.** Moment-independent Delta sensitivity (256-point Latin-hypercube design) and accepted values for the 12 N parameters retained by the combined dynamic-loss threshold.

| Parameter | Process Description | Delta RMSE | Delta Corr. | Delta Amp. | Delta Dynamic | Accepted Value |
|---|---|---|---|---|---|---|
| *kDMinDetS* | Sediment detritus mineralization rate for organic N | 0.0344 | 0.0456 | 0.0344 | 0.0352 | 0.0134 |
| *kNitrS* | Sediment nitrification rate | 0.0335 | 0.0233 | 0.0502 | 0.0290 | 0.4333 |
| *fRefrDetS* | Sediment detritus refractory fraction of sediment particulate organic matter | 0.0286 | 0.0144 | 0.0251 | 0.0273 | 0.2143 |
| *kNDifNH4S* | Sediment–water exchange molecular ammonium diffusion coefficient | 0.0280 | 0.0218 | 0.0325 | 0.0256 | $9.4616251 \times 10^{-5}$ |
| *cTurbDifNutS* | Sediment–water exchange bioturbation multiplier for nutrient diffusion | 0.0272 | 0.0287 | 0.0517 | 0.0229 | 8.0581 |
| *kNDifNO3S* | Sediment–water exchange molecular nitrate diffusion coefficient | 0.0238 | 0.0132 | 0.0274 | 0.0226 | $4.5847177 \times 10^{-5}$ |
| *cThetaMinW* | Water detritus temperature coefficient for particulate detritus mineralization | 0.0208 | 0.0201 | 0.0220 | 0.0205 | 1.0175 |
| *cThetaMinS* | Sediment detritus temperature coefficient for sediment detritus mineralization | 0.0226 | 0.032 | 0.0357 | 0.0202 | 1.0053 |
| *fDepthDifS* | Sediment–water exchange diffusion distance as fraction of active sediment depth | 0.0228 | 0.0204 | 0.0328 | 0.0201 | 0.7697 |
| *cThetaDifS* | Sediment–water exchange temperature coefficient for pore-water diffusion | 0.0221 | 0.0191 | 0.0211 | 0.0201 | 1.0513 |
| *kDMinDetW* | Water detritus particulate detritus mineralization rate carrying organic N | 0.0223 | 0.0172 | 0.0277 | 0.0200 | 0.0110 |
| *cNDBlueMaxW* | Maximum cyanobacterial N quota | 0.0192 | 0.0166 | 0.0231 | 0.0195 | 0.0809 |

Note: Accepted parameter values are reported to sufficient decimal precision to confirm their interior placement relative to the declared bounds. Physical units and boundary ranges are detailed in Table S4.

### *3.4. TN Phase and Amplitude After Refinement*

Sequential TN refinement yielded modest reductions in pointwise concentration error alongside pronounced gains in two dynamic diagnostics. During calibration, TN RMSE edged down from 0.1942 to 0.1938 mg $L^{-1}$ and NSE improved from −0.1620 to −0.1562, whereas signed correlation reversed from $r$ = −0.1319 to +0.1466 and the simulated-to-observed SD ratio nearly doubled from 0.2896 to 0.5062. These dynamic improvements carried over to the out-of-sample validation period: TN RMSE declined from 0.1666 to 0.1556 mg $L^{-1}$, NSE improved from −0.3250 to −0.1560, signed correlation shifted from $r$ = −0.1697 to +0.1258, and the SD ratio rose from 0.4250 to 0.5137. Tuning internal nitrogen parameters thus corrected an inverse temporal association and

recovered roughly half of the observed concentration SD. At the multi-year mean scale, observed and simulated TN were closely aligned: 0.5122 versus 0.5519 mg $L^{-1}$ (+7.7%) in calibration and 0.5117 versus 0.5328 mg $L^{-1}$ (+4.1%) in validation.

Despite the improved correlation sign and variance ratio, simulated and observed TN climatologies remained substantially out of phase. Whereas observed TN peaked in September during both calibration and validation, simulated TN peaked in April, resulting in a persistent circular phase advance of −5 months in both windows (Figure 5a; Table 4). Consequently, monthly climatological correlations remained weak ($r$ = 0.136 in calibration; $r$ = 0.166 in validation), and the simulated climatological amplitude reached only about half of the observed seasonal range (0.255 versus 0.490 mg $L^{-1}$ in calibration; 0.211 versus 0.475 mg $L^{-1}$ in validation). Seasonal mean residuals were positive in spring (+0.168 and +0.141 mg $L^{-1}$) and negative in winter (−0.085 and −0.121 mg $L^{-1}$), while autumn exhibited the largest seasonal RMSE during calibration (Figure 5b). Internal parameter refinement thus improved overall dispersion and correlation direction without resolving the underlying seasonal phase mismatch.




**Figure 5.** Seasonal phase, amplitude, and residual bias diagnostics for TN, TP, and Chl-a. (**a**) Observed and simulated calendar-month climatologies during calibration (2021–2023) and out-of-sample evaluation (2024–2025). (**b**) Seasonal mean residual bias (simulated minus observed). In both

periods, simulated TN peaked in April whereas observed TN peaked in September, yielding a circular peak-month phase offset of −5 months. Red annotations highlight the simulated April and observed September TN peaks in both panels, illustrating the persistent five-month phase advance (signed peak-month lag = −5 months).

**Table 4.** Calendar-month climatological phase and amplitude diagnostics for the accepted refinement. Peak-month phase lag (simulated minus observed) is wrapped circularly on [−6, +5] months, where negative values denote an earlier simulated peak.

| Variable | Period | Obs. Peak | Sim. Peak | Lag (Months) | Obs. Amplitude | Sim. Amplitude | Clim. *r* |
|---|---|---|---|---|---|---|---|
| WT | Calibration | 7 | 7 | 0 | 25.252 | 29.664 | 0.989 |
| DO | Calibration | 2 | 2 | 0 | 5.267 | 5.445 | 0.854 |
| DO | Validation | 1 | 2 | 1 | 4.4 | 4.941 | 0.764 |
| TN | Calibration | 9 | 4 | −5 | 0.49 | 0.255 | 0.136 |
| TN | Validation | 9 | 4 | −5 | 0.475 | 0.211 | 0.166 |
| TP | Calibration | 11 | 9 | −2 | 0.02 | 0.015 | −0.068 |
| TP | Validation | 8 | 9 | 1 | 0.025 | 0.017 | 0.454 |
| Chl-a | Calibration | 7 | 7 | 0 | 17.667 | 17.885 | 0.363 |
| Chl-a | Validation | 8 | 9 | 1 | 16.25 | 16.407 | 0.513 |

### *3.5. TP Performance Across the Reported Concentration Levels*

In the accepted refinement, calibration TP RMSE was 0.01262 mg $L^{-1}$—roughly equivalent to a single 0.01 mg $L^{-1}$ reporting step—yet $R^2$ was 0.007 and NSE was −0.983. During out-of-sample validation, TP RMSE declined to 0.00939 mg $L^{-1}$, $R^2$ rose to 0.112, and NSE approached zero (−0.025). Simulated and observed validation peaks occurred in September and August (a +1 month lag), compared with a −2 month offset during calibration (Figure 5a; Table 4). Although the monthly climatological correlation improved from $r$ = −0.068 in calibration to $r$ = 0.454 in validation, this gain must be interpreted cautiously given that all 60 observations were quantized into five discrete reporting levels. Multi-year mean TP concentrations were 0.0444 (observed) versus 0.0383 mg $L^{-1}$ (simulated; −13.8%) in calibration and 0.0388 versus 0.0389 mg $L^{-1}$ (+0.3%) in validation.

The TP calibration stage varied seven parameters controlling sediment and water-column phosphate sorption, iron-related adsorption capacity, redox sensitivity, maximum pore-water phosphate concentration, and molecular phosphate diffusion. The accepted TP candidate satisfied the relaxed 110% TP RMSE guardrail and reduced out-of-sample validation RMSE relative to the prior calibration, while calibration RMSE and NSE worsened.

### *3.6. Chlorophyll a Performance and Parameter Responsiveness*

Across the full sequential refinement, Chl-a calibration RMSE edged down from 9.036 to 8.933 μg $L^{-1}$ (NSE improving from −0.438 to −0.405), and validation RMSE declined from 7.255 to 7.146 μg $L^{-1}$ (NSE improving from −0.527 to −0.481; $R^2$ increasing from 0.145 to 0.164). These marginal gains were small relative to the remaining prediction error, and mean underestimation deepened from −1.48 to −3.29 μg $L^{-1}$ in calibration and from −1.17 to −3.15 μg $L^{-1}$ in validation. Refinement reduced the calibration SD ratio from 0.931 to 0.773 while bringing the validation SD ratio closer to unity (from 1.175 to 1.004), improving out-of-sample amplitude agreement at the expense of larger negative bias. In two-bound OAT screening, the Chl-a objective was overwhelmingly dominated by cyanobacterial parameters (Figure S3): perturbing the maximum cyanobacterial chlorophyll-to-dry-weight ratio (*cChDBlueMax*) shifted the composite score by 0.722, followed by cyanobacterial mortality (*kMortBlueW*, 0.383), maintenance respiration

(*kDRespBlue*, 0.273), and maximum specific growth rate (*cMuMaxBlue*, 0.268), whereas most diatom and green-algal parameters altered the composite score by less than 0.001 (Figure S3). Despite this clear parameter responsiveness, the dedicated 200-candidate Chl-a optimization stage retained its incoming parameter seed because no candidate improved the Chl-a objective without violating the cross-variable guardrails. Multi-year mean Chl-a was 16.897 (observed) versus 13.604 μg $L^{-1}$ (simulated; −19.5%) in calibration and 17.238 versus 14.090 μg $L^{-1}$ (−18.3%) in validation.

Simulated Chl-a underestimated observations in spring (−6.50 μg $L^{-1}$ in calibration; −7.49 μg $L^{-1}$ in validation) and winter (−7.74 and −7.07 μg $L^{-1}$), whereas summer mean bias was near zero in calibration but positive in validation (+3.47 μg $L^{-1}$) (Figure 5b). Monthly climatological correlations remained modest ($r$ = 0.363 in calibration; $r$ = 0.513 in validation). The model matched the July calibration peak and lagged the August validation peak by one month, while pronounced multi-season biases and negative NSE values remained (Figure 5a; Table 4). The slight improvement between the previous and accepted Chl-a trajectories arose from upstream nutrient-stage adjustments; the direct Chl-a stage retained its incoming seed.

### *3.7. Candidate Feasibility and Guardrail Sensitivity*

Across the three sequential-calibration stages, 664 parameter candidates were evaluated, tabulated separately from the OAT and Delta sensitivity runs (Table S3). Only 2 of the 336 TN candidates (0.60%) satisfied all TN-stage guardrails simultaneously (Figure 6; Table S13), while 1 of 128 TP candidates (0.78%) and 1 of 200 Chl-a candidates (0.50%) satisfied their respective stage guardrails (Table S13). These empirical pass rates reflect the joint intersection of technical completion, full sample coverage, physical and DO RMSE tolerances, cross-variable ecological RMSE limits, positive TN correlation, nonzero variance, and state-repair caps within the sampled differential-evolution populations. Figure 6a illustrates the candidate-level trade-off surface across all 336 evaluated TN parameter sets, showing that the accepted TN solution achieved positive correlation while maintaining near-baseline RMSE. Across all five state variables, sequential refinement produced modest and occasionally opposing shifts in RMSE, $R^2$, and NSE between calibration and validation (Figure 6b).

In our retrospective guardrail audit, relaxing the shared water-temperature and DO RMSE allowances from 101% to 105% and 110% of baseline left the exact same 2 of 336 TN candidates feasible (0.60%) and preserved the same optimal candidate (Table S12). Although relaxing the temperature and DO tolerances alone increased the number of passing candidates from 245 to 282 and 303, every newly admitted candidate violated at least one ecological or state-repair guardrail. Consequently, the strict 101% physical/DO tolerance was not the sole bottleneck limiting TN feasibility within the archived candidate population.

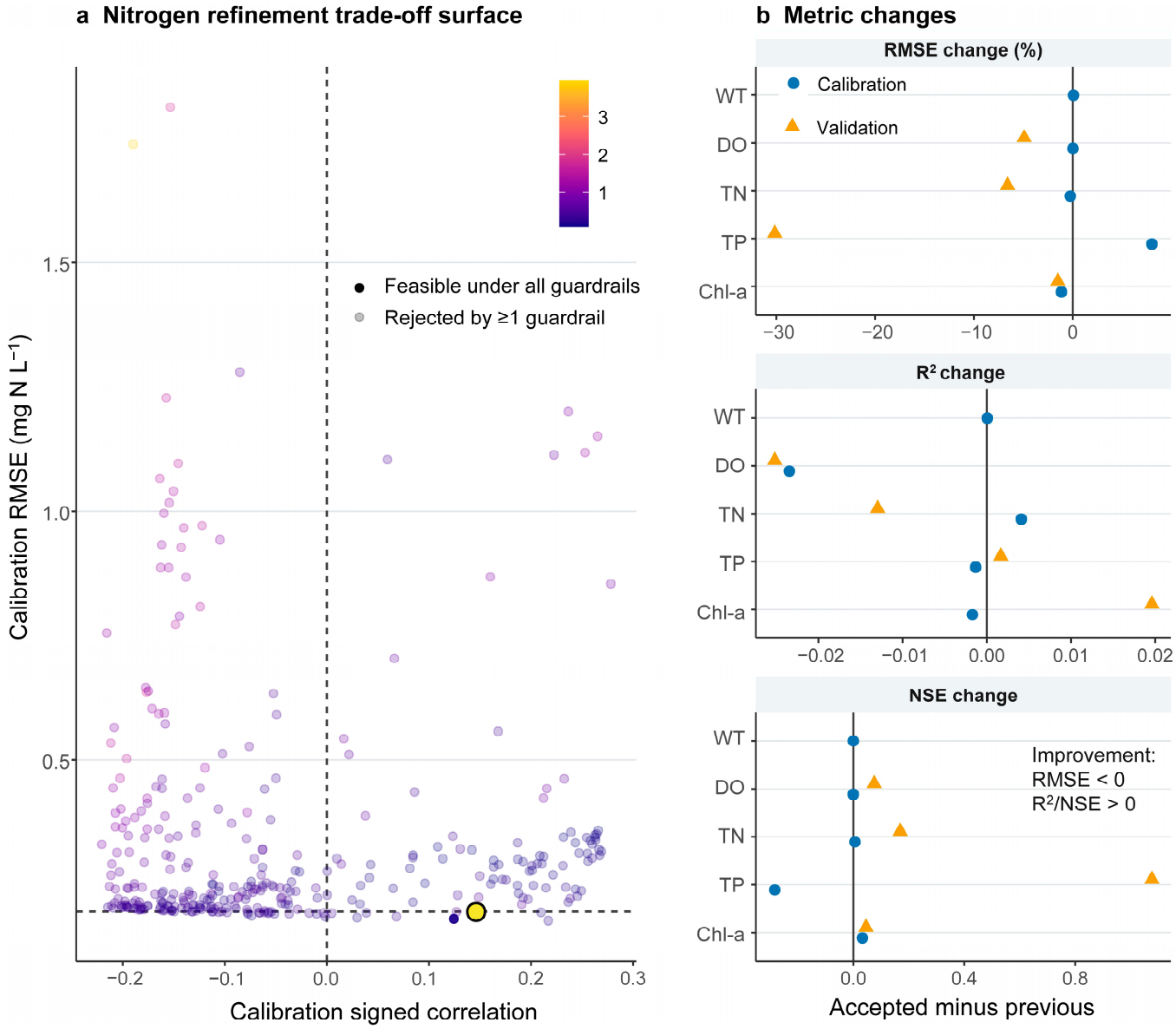



**Figure 6.** Nitrogen calibration trade-off surface and resulting shifts in state-variable performance metrics. (**a**) Calibration signed correlation (*r*) versus TN RMSE across all 336 evaluated TN candidates; symbol color indicates the simulated-to-observed SD ratio, and the enlarged yellow circle highlights the accepted candidate. Dashed lines denote *r* = 0 and the pre-refinement baseline TN RMSE. (**b**) Changes in performance metrics (accepted refinement minus previous calibration), where RMSE shifts are expressed as percentages and $R^2$ and NSE shifts as absolute differences. Panel (**b**) compares the two accepted parameterizations rather than candidate-level trade-offs for TP or Chl-a.

### *3.8. Reported Numerical Interventions*

The final accepted run contained 2,686,380 reported FABM state-repair events, comprising 1,760,786 water-column interior (65.54%), zero surface, and 925,594 benthic (34.46%) occurrences (Table S2). This cumulative count was 6.8% higher than the pre-refinement baseline and remained within the 110% count guardrail. Evaluating all 336 TN candidate runs confirmed that all four counter fields were complete and satisfied the exact identity total = interior + surface + bottom with zero discrepancy, providing a verified record of numerical clipping frequency across model domains.

### *3.9. Response to Controlled Seasonal Nitrogen Loading*

Across the five native simulations executed in the version-locked model reconstruction (Supplementary Methods S7), imposing ±50% sinusoidal inflow nitrogen cycles with winter, spring, summer, or autumn maxima while conserving annual N loading altered the annual harmonic TN semi-amplitude by −1.61% to +4.55%, yet the simulated climatological TN peak remained locked in May across all scenarios—four months ahead of the observed September peak (Table S14). Because the uncalibrated public-WET reconstruction operates at a higher baseline concentration scale (period-mean simulated TN of 7.337–7.352 mg $L^{-1}$) than the calibrated setup (0.512 mg $L^{-1}$ in the matched observations), this perturbation experiment specifically isolates relative responses in seasonal phase and harmonic amplitude relative to the constant-loading baseline.

## 4. Discussion

### *4.1. Information Content of the Observations and Forcing*

The monthly monitoring record provided sufficient temporal structure to evaluate the annual near-surface thermal cycle and broad seasonal DO patterns, but offered little direct constraint on vertical stratification gradients, benthic–pelagic nutrient fluxes, or time-varying external nutrient loads. These informational asymmetries shaped the calibration problem before any optimization algorithm was executed. Near-surface water temperature is governed primarily by synoptic meteorological forcing at the air–water interface—which was resolved daily—whereas in-lake nutrient concentrations depend jointly on external tributary loading, internal biogeochemical cycling, sediment–water exchange, and biological uptake. Prescribing a constant product of inflow discharge and nutrient concentration removes a primary driver of seasonal nutrient variability at the model boundary. While adjusting internal mineralization, nitrification, or diffusion rates can alter internal residence times and phase lags, forcing internal parameters to compensate for unmeasured storm runoff, water diversions, wastewater inputs, or tributary pulses risks severe parameter distortion.

Previous coupled GOTM–FABM–PCLake and GOTM–WET studies relied on substantially richer boundary and process datasets than were available for Kun Cheng Lake. In Jihongtan Reservoir, Chou et al. [23] drove their model with monthly, time-varying inflow, outflow, TN, and TP series over a 14-year window (2000–2013, spanning spin-up, eight calibration years, and two validation years). In Lake Hinge, Andersen et al. [22] calibrated against two seven-year windows encompassing speciated inorganic nutrients, phytoplankton, zooplankton, and macrophyte coverage. In Meiliang Bay, Yue et al. [24] paired in-lake temperature, TN, and TP measurements with water-balance exchange fluxes, speciated nutrient inputs, and satellite-derived Chl-a. While these comparisons do not imply that one system is intrinsically simpler to model, they underscore why calibrating a similar number of parameters does not yield equivalent evidentiary support when boundary and process data are sparse. Our diagnostic workflow therefore separates technical execution, pointwise state fit, dynamic fidelity, and fitness for prospective management prediction (Figure 1), recognizing that model adequacy is always relative to the target question and supporting evidence [14,17].

Because water temperature measurements were restricted to the 2021–2023 calibration window, our results support retrospective near-surface thermal diagnosis while highlighting out-of-sample temporal and vertical profile verification as immediate priorities (Figure 2). Simulated DO captured the broad annual cycle with season-specific residual biases, whereas for TN, TP, and Chl-a, near-zero $R^2$ values show that approximate agreement in multi-year mean concentrations coexisted with weak temporal covariation (Figure S1).

Reporting $R^2$ alone masks the sign of linear association (Figure 3). Pointwise error, seasonal phase, and amplitude metrics yielded contrasting assessments of model skill, reinforcing the need for multi-metric evaluation of environmental models [7] and system-inspired diagnostics that expose dynamic flaws hidden by aggregate error norms [15]. Importantly, the stark contrast between physical and biogeochemical performance in Kun Cheng Lake does not indict the underlying PCLake/WET model structure, which has successfully reproduced nutrient and food-web dynamics in better-constrained lakes [19,22,23]. Rather, it defines a site- and data-specific diagnostic boundary: under constant nutrient forcing and monthly near-surface monitoring, the coupled model captured the meteorologically forced thermal cycle and broad DO seasonality but could not identify the seasonal nutrient and algal dynamics required for prognostic scenario analysis.

### *4.2. Sensitivity Screening and Calibration Trade-Offs*

The Borgonovo Delta index is a dimensionless measure of distributional shift across a parameter's sampled range; it does not quantify a fraction of output variance, the sign of an ecological effect, or an in situ biogeochemical rate. The metric-dependent parameter rankings observed in Section 3.3 align with prior findings that lake morphometry and diagnostic choice reshape sensitivity hierarchies [25]. Moreover, high sensitivity of simulated TN to benthic mineralization, nitrification, and pore-water diffusion parameters over their declared bounds does not imply that sediment mineralization or internal nitrogen loading is exceptionally high in Kun Cheng Lake. Because no field measurements of benthic nutrient fluxes, pore-water profiles, denitrification, or sediment organic-matter turnover were available, internal parameters can appear highly influential simply because they compensate for constant boundary forcing or 1D spatial aggregation. Thus, the screened parameters represent conditional model controls whose unique field values remain unidentifiable from monthly surface concentrations alone.

The added assessment followed the convergence-guided sensitivity framework of Mesman et al. [26], applying uniform estimator settings across two independent 75-point maximin Latin-hypercube realizations and nested pooled samples of $n$ = 50, 100, and 150 runs. The results demonstrate a highly stable leading core of dominant parameters alongside a sample-responsive secondary tail, with the full 18-parameter ordering continuing to tighten at $n$ = 150 under our operational convergence criteria. While the interior placement of all 29 accepted parameters confirms that the optimizer did not stall against declared boundary limits, formal parameter identifiability and equifinality remain open questions requiring dedicated posterior uncertainty analysis [8,9].

The low empirical pass rates across the three calibration stages (0.50–0.78%) reflect the specific sequence of parameter subsets, objective weights, search bounds, and cross-variable guardrails imposed during differential evolution. While sequential bottom-up calibration is designed to protect previously fitted physical and chemical states [22,23], locking in upstream parameter values and enforcing tight guardrails in downstream stages can foreclose joint parameter adjustments that a simultaneous multi-objective search might discover. Because we did not benchmark our sequential protocol against simultaneous multi-objective optimization or alternative stage orderings, we cannot isolate how much of the low pass rate stems from sequential locking versus intrinsic structural conflict. Our deterministic reclassification of the 336 archived TN candidates (Table S12) provides an empirical baseline for designing relaxed or simultaneous searches across the TN, TP, and Chl-a stages. Ultimately, however, global sensitivity analysis and algorithmic tuning can only organize the parameter search; they cannot substitute for missing boundary hydrographs or process-level observations [10].

*4.3. Nutrient Timing and Phosphorus Observation Resolution*

The TN period means showed close agreement between simulated and observed concentration scales, whereas the five-month seasonal phase advance persisted across both calibration and out-of-sample validation despite improved correlation direction and variance ratio. This persistent phase discrepancy demonstrates that tuning internal nitrogen parameters within the evaluated search space was insufficient to recover the observed seasonal TN cycle. While our forcing and configuration audits document constant boundary inputs and a 1D water column as factual features of the setup, the specific mechanisms linking those features to the five-month TN phase offset remain testable hypotheses: unmeasured seasonal inflow and load pulses could shift external nutrient delivery; horizontal advection could redistribute tributary plumes and resuspended sediment; monthly sampling could alias synoptic flushing events; and alternative benthic or food-web parameterizations could alter internal nitrogen turnover timescales. Disentangling their relative contributions requires controlled factorial experiments that perturb one candidate mechanism at a time against a common validation dataset. By comparison, Chou et al. [23] achieved TN correlations of $r = 0.83$ in calibration and $r = 0.69$ in validation by driving their model with monthly time-varying inflow hydrology and nutrient chemistry over a longer calibration record, while Andersen et al. [22] constrained internal cycling with multi-trophic observations. Together, these studies highlight the specific forcing and process datasets needed to advance the present Kun Cheng Lake configuration.

Our controlled seasonal nitrogen-loading experiments (Section 3.9) show that within the version-locked model reconstruction, the timing of the simulated May TN peak is remarkably insensitive to idealized ±50% sinusoidal boundary timing when annual mean loading and all other model settings are held fixed. Observing a measurable response in harmonic amplitude (−1.61% to +4.55%) alongside an invariant May peak refines our earlier hypothesis regarding constant boundary forcing: simple smooth seasonal concentration cycles alone do not correct the phase offset in the reconstructed setup, pointing instead toward episodic hydrological flushing, coupled discharge–concentration covariance, or internal benthic–pelagic turnover parameterization as priorities for future testing. Note that the four-month phase offset pertains specifically to the version-locked public-WET reconstruction, whereas the five-month phase advance remains the documented diagnostic outcome of the historical accepted calibration.

For TP, the apparent paradox between a small absolute RMSE (~0.01 mg $L^{-1}$) and a strongly negative calibration NSE (−0.983) stems directly from the narrow ambient concentration range and coarse reporting resolution. When observed variance is small, residuals on the order of a single 0.01 mg $L^{-1}$ reporting increment can easily exceed the variance of the observations around their period mean. Because quantization into five discrete concentration levels compresses temporal covariance, $R^2$ must be interpreted jointly with absolute error, seasonal phase diagnostics, and out-of-sample validation. Overall, the available TP record weakly constrains the multi-year mean concentration scale—as reflected in the close validation period mean reported in Section 3.5—while negative NSE and low $R^2$ confirm that monthly predictive skill remains inadequate.

Reporting TP at five levels separated by 0.01 mg $L^{-1}$ compresses the concentration gradients needed to discriminate among competing phosphorus mechanisms, allowing distinct combinations of pelagic/benthic sorption, redox-mediated release, and pore-water diffusion to yield nearly indistinguishable fit scores. Whereas the 0.01 mg $L^{-1}$ step reflects reporting granularity, true observation uncertainty also encompasses analytical precision, digestion recovery, and field sampling variability. As demonstrated by Stow and Reckhow [45], observation error in lake phosphorus datasets can systematically bias parameter estimates in ways that depend on both model structure and sampling regime,

allowing acceptable concentration predictions to coexist with substantial parameter bias. In our configuration, the accepted TP parameters therefore represent an effective, composite parameterization conditioned on the quantized TP series, constant boundary inputs, objective weights, and cross-variable guardrails. While the close validation mean confirms reasonable concentration-scale agreement, mechanistically identifying individual sorption, redox, and diffusion parameters will require pairing high-precision dissolved and particulate P measurements with pore-water profiles and direct benthic flux incubations.

### *4.4. Chlorophyll a Processes and Monitoring Priorities*

Achieving credible Chl-a calibration remains a prerequisite for prognostic ecological modeling. In-lake Chl-a integrates phytoplankton biomass, taxonomic succession, intracellular pigment acclimation, underwater light attenuation, dissolved nutrient availability, zooplankton grazing, vertical settling, and horizontal advection. Bulk Chl-a measurements constrain only the net outcome of these interacting processes; disentangling the underlying mechanisms requires concurrent observations of the contributing rates and pools. In particular, the strong OAT sensitivity of simulated Chl-a to cyanobacterial growth, loss, and pigment-quota parameters highlights key model controls that should be evaluated against taxon-specific biomass observations. Both Anagnostou et al. [46] and Soares and Calijuri [3] similarly emphasize aligning field monitoring with resolved model processes and target management applications, while regional shallow-lake studies demonstrate how algal biomass responses are jointly mediated by flushing regime, ambient nutrient background, and lake-specific trophic structure [47].

A practical Chl-a evaluation sequence begins with paired field measurements of extracted Chl-a, taxonomic phytoplankton biovolume and biomass, zooplankton abundance, underwater photosynthetically active radiation (PAR) attenuation, suspended solids/turbidity, dissolved inorganic nutrient fractions, and event-scale external loads. Field campaigns should specifically target the spring and winter underestimation windows and the summer validation overestimation window, cross-calibrating any high-frequency fluorometric or satellite retrievals against laboratory extraction records. Such process-level datasets would enable independent verification of chlorophyll-to-carbon stoichiometry, group-specific growth and grazing losses, and light limitation, followed by multi-metric calibration targeting bias, seasonal amplitude, and bloom phenology across held-out validation windows. Multi-station horizontal surveys can then establish whether lateral transport heterogeneity warrants a 2D or 3D coupled framework.

### *4.5. Numerical Intervention and Spatial Representation*

The evaluated configuration shows three documented diagnostic features: TN parameter adjustment improved correlation direction and variability while the seasonal timing discrepancy persisted; supplied inflow discharge and nutrient concentrations were constant; and the observational record consisted of monthly near-surface state measurements (Figure 7; Table 5). Together, these facts establish the achieved performance and available information, whereas attributing the residuals to boundary forcing, parameter uncertainty, spatial representation, or numerical intervention remains a separate research task. The 29 accepted parameter values describe the selected solution within the declared search ranges. Within the hierarchy of Hipsey et al. [14], the strongest evidence concerns near-surface temperature state agreement during calibration and modest temporal validation for DO. Process and system evaluation would add mixing

depth, oxygen minima, bloom phenology, and nutrient budgets, following the rationale of Kurucz et al. [15].

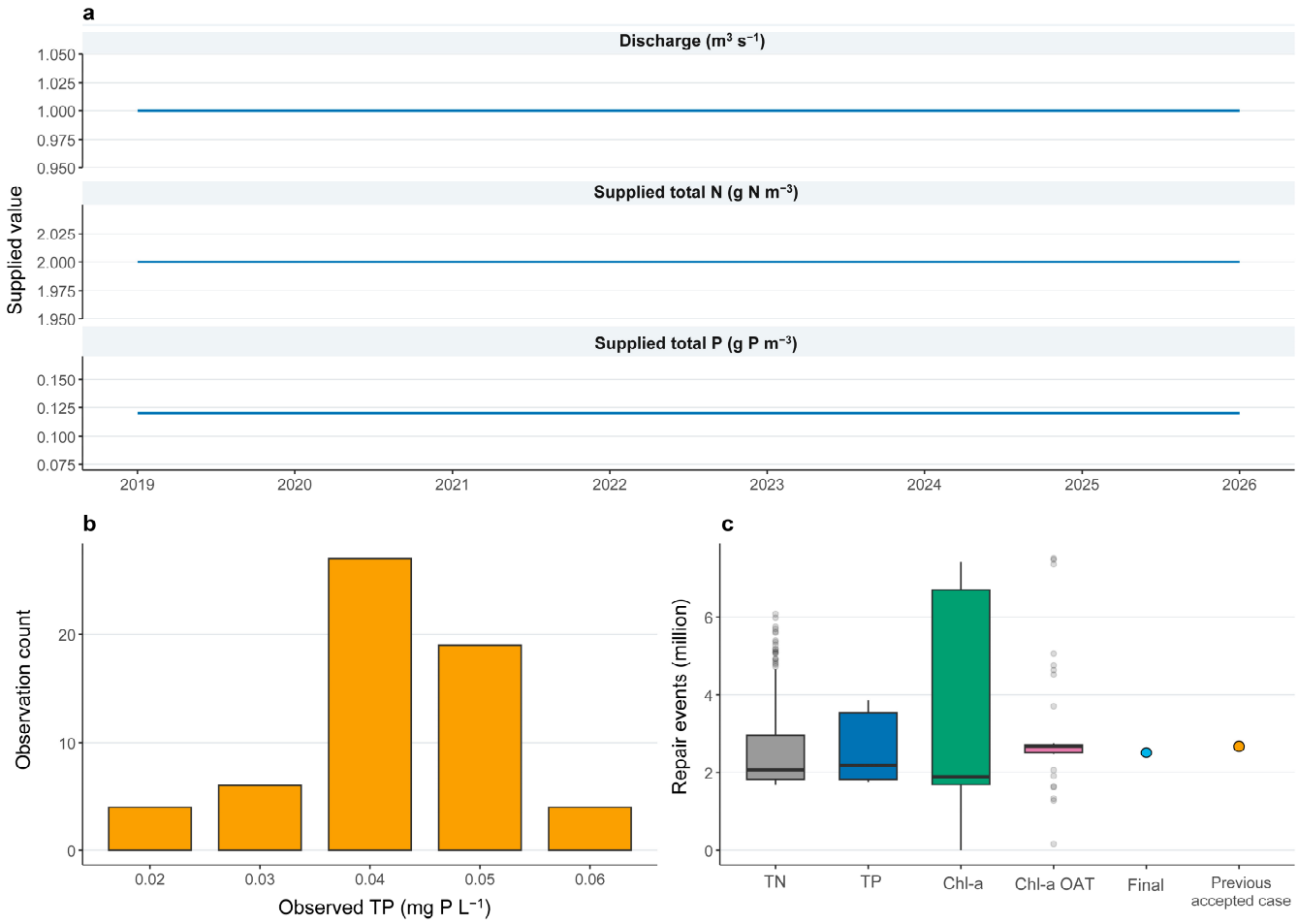



**Figure 7.** Boundary-forcing, observation-granularity, and numerical state-repair diagnostics. (**a**) Supplied daily inflow discharge, TN concentration, and TP concentration each remained invariant across all 2558 daily records (2019–2025). (**b**) Distribution of the 60 monthly TP observations across five discrete reporting levels (0.01 mg P $L^{-1}$ increments). (**c**) Cumulative FABM numerical state-repair counts across calibration stages; the final accepted run recorded 2,686,380 interventions, a 6.8% increase over the pre-refinement baseline. Counters record domain-level state-check triggers across interior, surface, and benthic domains (Table S2; Supplementary Methods S5).

**Table 5.** Variable-specific performance, discrepancy interpretation, and priorities for future evaluation. Performance metrics are drawn from Tables 2, 4 and S10; candidate error sources represent untested hypotheses unless explicitly noted as documented features.

| Variable | Documented Performance | Discrepancy Interpretation | Priority Observations and Evaluation |
|---|---|---|---|
| Water temperature | Calibration $R^2$ = 0.969; NSE = 0.933. Out-of-sample thermal validation remains an immediate priority. | Warm-season positive bias is documented; vertical profile and basin-wide representativeness require field verification. | Collect post-2023 vertical temperature profiles and local meteorological forcing across mixing events; evaluate depth-resolved and out-of-sample skill. |

| | | | |
|---|---|---|---|
| Dissolved oxygen | Calibration NSE = 0.368; validation NSE = 0.290. Broad seasonal cycle is captured with season-specific residual bias. | Air–water reaeration, primary production, respiration, and vertical mixing jointly contribute to the aggregate residual error. | Deploy high-frequency vertical DO and temperature loggers alongside metabolic rate measurements; evaluate hypoxic minima, seasonal bias, and synoptic events. |
| Total nitrogen | Validation NSE = −0.156; simulated seasonal peak occurs five months earlier than observed in both periods. | Constant boundary inputs and 1D column geometry are documented constraints; episodic load timing, lateral exchange, sampling frequency, and internal turnover rates remain causal hypotheses. | Monitor continuous inflow/outflow discharge and speciated N loads across storm events; conduct multi-station and benthic flux surveys; test boundary forcing and horizontal transport on common validation records. |
| Total phosphorus | Calibration NSE = −0.983; validation NSE = −0.025. All 60 observations fall into five discrete reporting levels (0.01 mg $L^{-1}$ steps). | Coarse reporting granularity limits parameter identifiability; analytical uncertainty and compensating sorption, redox-release, and diffusion effects complicate mechanistic interpretation. | Pair unrounded laboratory quality-assurance records with dissolved/particulate P fractions, pore-water profiles, and benthic flux incubations to constrain individual P controls. |
| Chlorophyll a | Calibration NSE = −0.405; validation NSE = −0.481. The dedicated Chl-a optimization stage retained its incoming parameter seed. | Persistent spring/winter underestimation and summer validation overestimation are documented; pigment stoichiometry, light climate, nutrient limitation, grazing, settling, and lateral transport remain hypotheses. | Pair extracted Chl-a with taxonomic phytoplankton biomass, zooplankton, underwater light profiles, and event loads; test growth, loss, and bloom-timing controls on held-out observations. |

Note. The proposed monthly prediction screen requires independent temporal validation, validation NSE > 0.3, an absolute circular peak-month lag ≤1 month, and a lower validation RMSE than calibration-fitted monthly climatology. Under this screen, water temperature supports retrospective calibration-period diagnosis, DO supports conditional seasonal interpretation (validation NSE = 0.290), and TN, TP, and Chl-a remain unsupported for monthly prediction. These quantitative criteria provide an operational gate for prospective monthly applications.

A 1D vertical column resolves vertical turbulent mixing and local pelagic–benthic transformations; near-surface DO skill therefore reflects the net balance of air–water reaeration, temperature-dependent solubility, primary production, respiration, and vertical mixing. By contrast, horizontal river–lake exchange, tributary plumes, spatially heterogeneous wind-wave resuspension, and littoral macrophyte zonation represent untested spatial hypotheses that could contribute to the observed nutrient and algal residuals. Basin morphometry is known to reshape parameter sensitivity [25], and depth-resolved profiles paired with high-frequency meteorological forcing are essential for capturing episodic physical–biogeochemical coupling [26,48]. Previously modeled horizontal circulation in Kun Cheng Lake [29] and storm-driven spatial gradients across the Taihu network [31,49] motivate collecting multi-station field observations and testing horizontal transport explicitly. Quantifying their actual contribution to the present TN, TP, and Chl-a residuals will require controlled model comparisons against our 1D baseline.

A correction to dissolved nutrients can alter subsequent algal uptake; clipping plankton or detrital pools can modify internal nutrient storage and mineralization; and

adjusting dissolved oxygen can perturb redox-dependent reaction rates and benthic–pelagic fluxes. Because biogeochemical source–sink terms depend nonlinearly on state concentrations, frequent minor adjustments can accumulate along a multi-year integration, while larger clipping events can directly distort mass inventories. The net ecological impact depends on the affected state pool, the signed magnitude of the adjustment, and subsequent nonlinear feedbacks. As emphasized by Bruggeman et al. [50], numerical state positivity (admissibility) is distinct from strict elemental mass conservation, which requires balanced stoichiometric transfers among coupled pools. Whereas a raw intervention counter weights every trigger equally, the actual mass-balance perturbation depends on clipping magnitude, elemental stoichiometry, and the associated layer volume or sediment area. Accordingly, the 2,686,380 reported state repairs document the frequency and domain distribution of numerical clipping, while our mechanistic analysis identifies the pathways through which those interventions can propagate into simulated nutrient, oxygen, and phytoplankton trajectories.

Spatially resolved alternatives offer specific ways to test the horizontal transport hypothesis. For example, Wang et al. [51] coupled a two-dimensional (2D) depth-averaged hydrodynamic and eutrophication model for shallow Lake Baiyangdian to simulate horizontal circulation alongside nutrient–algal dynamics. Similarly, Podgornyi and Leonov [52] developed a spatially heterogeneous compartment model to resolve nitrogen, phosphorus, and dissolved oxygen biotransformations across Neva Bay, while Medvinsky et al. [53] demonstrated how reaction–diffusion interactions generate spatiotemporal plankton patchiness. Whereas reaction–diffusion theory elucidates intrinsic ecological pattern formation, evaluating site-specific advective transport in Kun Cheng Lake requires measured inflow hydrographs and circulation fields. Where vertical profile observations confirm thorough water-column mixing, a 2D depth-averaged horizontal model provides a parsimonious framework for testing lateral nutrient and Chl-a gradients across Kun Cheng Lake. Conversely, if field surveys reveal concurrent horizontal gradients and transient vertical stratification, a three-dimensional (3D) coupled hydrodynamic–biogeochemical framework—such as the Lake Taihu model of Mao et al. [54], which resolves wind-driven 3D circulation, tributary plumes, and benthic release—would be warranted. Any higher-dimensional model should be benchmarked against the present 1D configuration using identical boundary inputs, calibration/validation splits, and seasonal phase and spatial-gradient metrics.

### *4.6. Benchmarks and the Domain of Applicability*

Together, the statistical and hydrological–nutrient mass-balance benchmarks establish an objective baseline for interpreting concentration errors and defining the domain of model applicability summarized in Table 5.

Among the five evaluated variables, empirical support is strongest for retrospective near-surface thermal diagnosis over the 2021–2023 calibration window (Table 5). High pointwise accuracy (NSE = 0.933), exact July peak-month agreement, and strong climatological correlation ($r$ = 0.989) support this diagnostic role, while the documented warm-season bias and lack of post-2023 or depth-resolved temperature profiles define clear priorities for temporal and vertical validation. Positive calibration and validation DO NSE values (0.368 and 0.290) support conditional interpretation of broad seasonal oxygen patterns, whereas depth-resolved high-frequency profiles are required before evaluating benthic hypoxia or synoptic depletion events. By contrast, prospective nutrient and Chl-a predictions remain unsupported because negative validation NSE, weak temporal correlation, and pronounced seasonal phase and bias errors persist. Extending the model to evaluate nutrient-load reductions, climate warming, algal blooms, regime shifts, or management thresholds would additionally require process-level observations tailored to

the target decision—including submerged macrophytes, underwater light transparency, plankton community structure, and direct benthic nutrient fluxes. These applicability boundaries apply specifically to the evaluated model configuration, boundary dataset, and calibration design.

The monthly climatology benchmark achieved a TN validation NSE of 0.841, compared with −0.156 for the accepted coupled model (Table S10). This stark contrast demonstrates that calendar-month identity alone captures a strong, repeatable seasonal cycle in the Kun Cheng Lake TN record. The coupled model's persistent five-month phase advance prevented it from reproducing that regular seasonal sequence under constant boundary forcing, while higher-frequency event sampling will be needed to characterize synoptic departures from the monthly climatology.

While the coupled GOTM–FABM–WET framework provides an integrated representation of vertical hydrodynamics and food-web biogeochemistry, our parsimonious mass-balance benchmark isolates the baseline loading–flushing–retention response governed by the prescribed constant boundaries and a single first-order net-removal coefficient per nutrient. Because constant boundary inputs drive the one-box analytical solution monotonically toward steady state, it serves as a transparent baseline for judging the net value of resolving internal biogeochemical cycling. Prior lake phosphorus-retention studies demonstrate that complete-mixing assumptions, effective reaction order, and rapid particulate settling strongly influence input–output model skill [55], while explicit benthic storage and internal recycling introduce multi-year recovery lags [56]—explaining why our fitted first-order removal coefficient reflects lumped retention over the observation window. Under the supplied constant boundary inputs, the coupled model reduced out-of-sample TP error relative to both the mass-balance benchmark and all three statistical benchmarks, whereas the simpler mass-balance model achieved lower out-of-sample TN error (Tables S10 and S11). Both nutrient evaluations rely on the exact same primary field measurements from the supplied OBS files and an identical calibration–validation split (Supplementary Methods S6). Even for TP, however, the coupled model's relative RMSE advantage coexisted with a slightly negative validation NSE (−0.025). Together, these benchmark comparisons demonstrate that added mechanistic complexity yielded a variable-specific reduction in concentration error for TP but not for TN under constant forcing; moving from aggregate error reduction to credible process attribution and scenario forecasting will require boundary and process observations matched to the intended management application [5].

Ultimately, the primary contribution of this diagnostic study is to separate incremental gains in in-sample calibration metrics from the empirical evidence required to support ecological prediction. Evaluating seasonal phase fidelity, out-of-sample validation skill, and performance relative to statistical and nutrient mass-balance benchmarks establishes clear, defensible boundaries on current model use while leaving the causal partitioning of remaining errors—and the potential benefits of higher-dimensional or time-varying configurations—open to rigorous empirical testing.

### *4.7. Limitations and Research Priorities*

Advancing from diagnostic evaluation to prognostic lake management requires organizing future work around three complementary priorities. First, collecting synchronized inflow/outflow hydrographs, speciated nutrient loads, multi-station spatial surveys, and in situ process measurements is essential to remove the primary boundary and observational bottlenecks. Second, our controlled fixed-mean seasonal nitrogen-loading experiments (Section 3.9) provide a template for mechanistic hypothesis testing: the persistence of the simulated May TN peak across all four ±50% sinusoidal loading regimes directs subsequent experiments toward the joint interaction of episodic flow-

weighted loading, internal benthic–pelagic turnover parameterizations, and simultaneous multi-objective calibration on a common validation dataset. Similarly, our sample-size and across-design Delta diagnostics separate a robust high-sensitivity parameter core from a sample-responsive secondary ranking, providing a foundation for larger ensemble designs and formal posterior parameter-uncertainty estimation, while the mass-balance benchmark establishes an ongoing baseline for justifying added model complexity. Third, pairing vertical profile deployments with horizontal water-quality surveys will provide the empirical basis needed to compare our 1D water column against 2D depth-averaged and 3D coupled hydrodynamic–ecological models. The Conclusions Section below translates these three priorities into an operational roadmap for lake managers.

## 5. Conclusions

The sensitivity-guided sequential calibration improved selected dynamic diagnostics while leaving substantial out-of-sample ecological prediction errors in the Kun Cheng Lake GOTM–FABM–WET configuration. Near-surface water temperature was reproduced accurately during the 2021–2023 calibration window (NSE = 0.933; post-2023 observations were unavailable for independent validation), and DO achieved modest validation skill during 2024–2025 (NSE = 0.290). Although TN refinement corrected an inverse correlation and increased simulated variance, a five-month seasonal phase advance persisted, and validation NSE remained negative for TN (−0.156), TP (−0.025), and Chl-a (−0.481). In particular, Chl-a calibration was inadequate (calibration NSE = −0.405), with its dedicated optimization stage retaining the incoming parameter seed. On validation records, the coupled model outperformed all three statistical benchmarks for TP and reduced TP mean squared error by 28.68% relative to the nutrient mass-balance benchmark, whereas TN mean squared error was 7.58% higher. Fewer than 1% of candidates satisfied each ecological stage's guardrails, while constant boundary forcing, coarse TP reporting resolution, and frequent numerical state repairs point to testable hypotheses for the remaining residuals.

For lake managers, the current evidence supports retrospective near-surface thermal diagnosis and conditional interpretation of broad seasonal DO patterns. Extending the model to nutrient or algal decision-making requires three operational priorities:

(1) Priority data gaps. Monitor synchronized inflow/outflow discharge (or water-balance-reconstructed hydrographs validated against lake levels), speciated nutrient chemistry, and storm/diversion events alongside complete meteorological forcing and bathymetry. Pair vertical temperature and DO profiles with multi-station TN, TP, and Chl-a surveys across pelagic, inflow, and littoral zones. Record unrounded TP quality-assurance metrics, P fractions, and benthic fluxes, and collect concurrent phytoplankton taxonomy, zooplankton, light attenuation, and bioavailable nutrient data across seasons with persistent Chl-a bias.

(2) Model analysis improvement. Evaluate bias, phase, and amplitude across standardized out-of-sample windows against statistical and mass-balance benchmarks. Apply sensitivity convergence diagnostics, quantify parameter equifinality alongside observation resolution, audit numerical state repairs for mass conservation, and archive all software versions, configurations, and logs in a version-locked repository.

(3) Model structure optimization. Use horizontal surveys and vertical profiles to determine whether a 1D column, 2D depth-averaged, or 3D coupled framework is warranted, and test lateral transport, benthic exchange, light limitation, and algal losses against field measurements before deploying the model for management scenarios.

**Supplementary Materials:** The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/su182111314/s1. Figure S1: Observed versus simulated

concentrations and temperatures across all evaluated state variables; Figure S2: Normalized positions of the 29 calibrated parameters within their declared bounds; Figure S3: One-at-a-time (OAT) Chl-a sensitivity across declared parameter bounds; Table S1: Summary audit of hydrological and nutrient boundary inputs; Table S2: Software versions and documented model configuration; Table S3: Model evaluation counts for sensitivity screening and the final sequential calibration; Table S4: The 29 parameters varied across the sequential calibration stages; Table S5: Analytical design and reproducibility record for the Delta sample-size response assessment; Table S6: Standard Delta stability diagnostics for paired-prefix sample-size comparisons and the independent A75-B75 comparison; Table S7: Repeated cumulative-subsampling diagnostics across adjacent sample sizes; Table S8: Parameter-level stability for the D response at P150 and across 100 cumulative $n$ = 125 subsets; Table S9: Moment-independent Delta sensitivity point estimates, uncentered bootstrap interval bounds, and dummy thresholds from the 256-point Latin-hypercube screening design across four response metrics: normalized TN RMSE (R), correlation loss (C), amplitude loss (A), and combined dynamic loss (D) (Supplementary Methods S2); Table S10: Out-of-sample validation performance of the accepted coupled model alongside three prespecified statistical benchmarks; Table S11: Out-of-sample validation comparison between the parsimonious nutrient mass-balance benchmark and the accepted coupled model across the 2024–2025 evaluation window (24 monthly observations per nutrient); Table S12: Retrospective reclassification of the 336 archived TN candidate runs under alternative shared water-temperature and DO RMSE guardrails, holding all other acceptance criteria and the pre-refinement baseline fixed; Table S13: Empirical candidate feasibility across the sequential calibration stages under all cross-variable and numerical guardrails; Table S14: TN response to controlled seasonal N loading in the version-locked coupled reconstruction (2021–2025). Refs. [57,58].

**Author Contributions:** Investigation, T.Y.; data curation, T.Y., J.W. and Y.S.; methodology, T.Y., Y.S. and D.Z.; software, T.Y. and D.L.; formal analysis, J.W.; writing—original draft preparation, T.Y.; writing—review and editing, P.G., D.Z., D.L. and S.A.; project administration, D.L. and S.A.; supervision, S.A.; funding acquisition, S.A. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the Research on Degradation Modeling and Restoration Technologies of Submerged Aquatic Vegetation in Shallow Lakes—Open Fund Project of the Key Laboratory of Taihu Basin Water Governance, Ministry of Water Resources (Fiscal Year 2024, project Yk924004), the National Natural Science Foundation of China (project 42607441), and the Scientific Research Project of the Department of Education of Hubei Province (project B2025163).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author due to a confidentiality agreement signed with municipal governments.

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

## References

1. United Nations Department of Economic and Social Affairs. Goal 6. Available online: https://sdgs.un.org/goals/goal6 (accessed on 7 October 2026).
2. Frassl, M.A.; Abell, J.M.; Botelho, D.A.; Cinque, K.; Gibbes, B.R.; Jöhnk, K.D.; Muraoka, K.; Robson, B.J.; Wolski, M.; Xiao, M.; et al. A short review of contemporary developments in aquatic ecosystem modelling of lakes and reservoirs. *Environ. Model. Softw.* **2019**, *117*, 181–187. https://doi.org/10.1016/j.envsoft.2019.03.024.
3. Soares, L.M.V.; Calijuri, M.D.C. Deterministic modelling of freshwater lakes and reservoirs: Current trends and recent progress. *Environ. Model. Softw.* **2021**, *144*, 105143. https://doi.org/10.1016/j.envsoft.2021.105143.

4. Uusitalo, L.; Lehikoinen, A.; Helle, I.; Myrberg, K. An overview of methods to evaluate uncertainty of deterministic models in decision support. *Environ. Model. Softw.* **2015**, *63*, 24–31. https://doi.org/10.1016/j.envsoft.2014.09.017.
5. Grüss, A.; Rose, K.A.; Simons, J.; Ainsworth, C.H.; Babcock, E.A.; Chagaris, D.D.; De Mutsert, K.; Froeschke, J.; Himchak, P.; Kaplan, I.C.; et al. Recommendations on the Use of Ecosystem Modeling for Informing Ecosystem-Based Fisheries Management and Restoration Outcomes in the Gulf of Mexico. *Mar. Coast. Fish.* **2017**, *9*, 281–295. https://doi.org/10.1080/19425120.2017.1330786.
6. Nash, J.E.; Sutcliffe, J.V. River flow forecasting through conceptual models part I—A discussion of principles. *J. Hydrol.* **1970**, *10*, 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.
7. Bennett, N.D.; Croke, B.F.W.; Guariso, G.; Guillaume, J.H.A.; Hamilton, S.H.; Jakeman, A.J.; Marsili-Libelli, S.; Newham, L.T.H.; Norton, J.P.; Perrin, C.; et al. Characterising performance of environmental models. *Environ. Model. Softw.* **2013**, *40*, 1–20. https://doi.org/10.1016/j.envsoft.2012.09.011.
8. Beven, K. A manifesto for the equifinality thesis. *J. Hydrol.* **2006**, *320*, 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007.
9. Her, Y.; Chaubey, I. Impact of the numbers of observations and calibration parameters on equifinality, model performance, and output and parameter uncertainty. *Hydrol. Process.* **2015**, *29*, 4220–4237. https://doi.org/10.1002/hyp.10487.
10. Pianosi, F.; Beven, K.; Freer, J.; Hall, J.W.; Rougier, J.; Stephenson, D.B.; Wagener, T. Sensitivity analysis of environmental models: A systematic review with practical workflow. *Environ. Model. Softw.* **2016**, *79*, 214–232. https://doi.org/10.1016/j.envsoft.2016.02.008.
11. Soares, L.M.V.; Calijuri, M.C. Sensitivity and identifiability analyses of parameters for water quality modeling of subtropical reservoirs. *Ecol. Model.* **2021**, *458*, 109720. https://doi.org/10.1016/j.ecolmodel.2021.109720.
12. Borgonovo, E. A new uncertainty importance measure. *Reliab. Eng. Syst. Saf.* **2007**, *92*, 771–784. https://doi.org/10.1016/j.ress.2006.04.015.
13. Plischke, E.; Borgonovo, E.; Smith, C.L. Global sensitivity measures from given data. *Eur. J. Oper. Res.* **2013**, *226*, 536–550. https://doi.org/10.1016/j.ejor.2012.11.047.
14. Hipsey, M.R.; Gal, G.; Arhonditsis, G.B.; Carey, C.C.; Elliott, J.A.; Frassl, M.A.; Janse, J.H.; de Mora, L.; Robson, B.J. A system of metrics for the assessment and improvement of aquatic ecosystem models. *Environ. Model. Softw.* **2020**, *128*, 104697. https://doi.org/10.1016/j.envsoft.2020.104697.
15. Kurucz, K.; Carey, C.C.; Huang, P.; De Sousa, E.R.; White, J.T.; Hipsey, M.R. Using System-Inspired Metrics to Improve Water Quality Prediction in Stratified Lakes. *Water Resour. Res.* **2024**, *60*, e2023WR036570. https://doi.org/10.1029/2023wr036570.
16. Hipsey, M.R.; Hamilton, D.P.; Hanson, P.C.; Carey, C.C.; Coletti, J.Z.; Read, J.S.; Ibelings, B.W.; Valesini, F.J.; Brookes, J.D. Predicting the resilience and recovery of aquatic systems: A framework for model evolution within environmental observatories. *Water Resour. Res.* **2015**, *51*, 7023–7043. https://doi.org/10.1002/2015wr017175.
17. Getz, W.M.; Marshall, C.R.; Carlson, C.J.; Giuggioli, L.; Ryan, S.J.; Romañach, S.S.; Boettiger, C.; Chamberlain, S.D.; Larsen, L.; D'Odorico, P.; et al. Making ecological models adequate. *Ecol. Lett.* **2018**, *21*, 153–166. https://doi.org/10.1111/ele.12893.
18. Bruggeman, J.; Bolding, K. A general framework for aquatic biogeochemical models. *Environ. Model. Softw.* **2014**, *61*, 249–265. https://doi.org/10.1016/j.envsoft.2014.04.002.
19. Hu, F.; Bolding, K.; Bruggeman, J.; Jeppesen, E.; Flindt, M.R.; van Gerven, L.; Janse, J.H.; Janssen, A.B.G.; Kuiper, J.J.; Mooij, W.M.; et al. FABM-PCLake—linking aquatic ecology with hydrodynamics. *Geosci. Model Dev.* **2016**, *9*, 2271–2278. https://doi.org/10.5194/gmd-9-2271-2016.
20. Schnedler-Meyer, N.A.; Andersen, T.K.; Hu, F.R.S.; Bolding, K.; Nielsen, A.; Trolle, D. Water Ecosystems Tool (WET) 1.0—A new generation of flexible aquatic ecosystem model. *Geosci. Model Dev.* **2022**, *15*, 3861–3878. https://doi.org/10.5194/gmd-15-3861-2022.
21. Nielsen, A.; Trolle, D.; Bjerring, R.; Søndergaard, M.; Olesen, J.E.; Janse, J.H.; Mooij, W.M.; Jeppesen, E. Effects of climate and nutrient load on the water quality of shallow lakes assessed through ensemble runs by PCLake. *Ecol. Appl.* **2014**, *24*, 1926–1944. https://doi.org/10.1890/13-0790.1.
22. Andersen, T.K.; Nielsen, A.; Jeppesen, E.; Hu, F.; Bolding, K.; Liu, Z.; Søndergaard, M.; Johansson, L.S.; Trolle, D. Predicting ecosystem state changes in shallow lakes using an aquatic ecosystem model: Lake Hinge, Denmark, an example. *Ecol. Appl.* **2020**, *30*, e02160. https://doi.org/10.1002/eap.2160.
23. Chou, Q.; Nielsen, A.; Andersen, T.K.; Hu, F.; Chen, W.; Zhang, X.; Cao, T.; Ni, L.; Jeppesen, E.; Trolle, D. Assessing Impacts of Changes in External Nutrient Loadings on a Temperate Chinese Drinking Water Reservoir. *Front. Environ. Sci.* **2021**, *9*, 632778. https://doi.org/10.3389/fenvs.2021.632778.

24. Yue, L.; Yang, Y.; Zhuang, X.; Chen, W.; Kong, X.; Deng, J.; Zhao, Z.; Lu, Y.; Zhu, G.; Qin, B. Simulation study on the response of chlorophyll-a concentration to extreme heatwaves in a large shallow lake. *J. Lake Sci.* **2026**, *38*, 65–77. https://doi.org/10.18307/2026.0110.
25. Andersen, T.K.; Bolding, K.; Nielsen, A.; Bruggeman, J.; Jeppesen, E.; Trolle, D. How morphology shapes the parameter sensitivity of lake ecosystem models. *Environ. Model. Softw.* **2021**, *136*, 104945. https://doi.org/10.1016/j.envsoft.2020.104945.
26. Mesman, J.P.; Ayala, A.I.; Goyette, S.; Kasparian, J.; Marcé, R.; Markensten, H.; Stelzer, J.A.A.; Thayne, M.W.; Thomas, M.K.; Pierson, D.C.; et al. Drivers of phytoplankton responses to summer wind events in a stratified lake: A modeling study. *Limnol. Oceanogr.* **2022**, *67*, 856–873. https://doi.org/10.1002/lno.12040.
27. Changshu Municipal People's Government. Kun Cheng Lake Water Quality Reaches Class III for the First Time in Nine Years. Available online: https://www.changshu.gov.cn/zgcs/c100293/202501/c86ca1a488d346a2ac3977bbaf0602a8.shtml (accessed on 25 July 2026). (In Chinese)
28. Cao, J.J.; Li, R.J.; Qin, Y. Water-quality simulation for an urban shallow lake: A case study of Kun Cheng Lake. *Water Resour. Prot.* **2009**, *25*, 12–15. (In Chinese)
29. Qin, Y.; Li, R.; Zhu, W.; Cao, J. Numerical analysis of hydrodynamics and water quality in Kun Cheng Lake. *Trans. Oceanol. Limnol.* **2009**, *2*, 7–13. (In Chinese)
30. Yang, X.; Anderson, N.J.; Dong, X.; Shen, J. Surface sediment diatom assemblages and epilimnetic total phosphorus in large, shallow lakes of the Yangtze floodplain: Their relationships and implications for assessing long-term eutrophication. *Freshw. Biol.* **2008**, *53*, 1273–1290. https://doi.org/10.1111/j.1365-2427.2007.01921.x.
31. Li, Q.; Zhu, Y.; Chen, Q.; Li, Y.; Chen, J.; Gao, Y.; Lin, P. Spatio-temporal dynamics of water quality and eutrophication in Lake Taihu, China. *Ecohydrology* **2021**, *14*, e2291. https://doi.org/10.1002/eco.2291.
32. *GB 13195-91*; Water Quality—Determination of Water Temperature—Thermometer or Reversing Thermometer Method. National Environmental Protection Agency of China: Beijing, China, 1991. (In Chinese)
33. *HJ 506-2009*; Water Quality—Determination of Dissolved Oxygen—Electrochemical Probe Method. Ministry of Environmental Protection of the People's Republic of China: Beijing, China, 2009. (In Chinese)
34. *HJ 636-2012*; Water Quality—Determination of Total Nitrogen—Alkaline Potassium Persulfate Digestion UV Spectrophotometric Method. Ministry of Environmental Protection of the People's Republic of China: Beijing, China, 2012. (In Chinese)
35. *GB 11893-89*; Water Quality—Determination of Total Phosphorus—Ammonium Molybdate Spectrophotometric Method. National Environmental Protection Agency of China: Beijing, China, 1989. (In Chinese)
36. *HJ 897-2017*; Water Quality—Determination of Chlorophyll a—Spectrophotometric Method. Ministry of Environmental Protection of the People's Republic of China: Beijing, China, 2017. (In Chinese)
37. *HJ/T 91-2002*; Technical Specifications Requirements for Monitoring of Surface Water and Waste Water. State Environmental Protection Administration of China: Beijing, China, 2002. (In Chinese)
38. *HJ 91.2-2022*; Technical Specification for Surface Water Environmental Quality Monitoring. Ministry of Ecology and Environment of the People's Republic of China: Beijing, China, 2022. (In Chinese)
39. Janse, J.H.; Scheffer, M.; Lijklema, L.; Van Liere, L.; Sloot, J.S.; Mooij, W.M. Estimating the critical phosphorus loading of shallow lakes with the ecosystem model PCLake: Sensitivity, calibration and uncertainty. *Ecol. Model.* **2010**, *221*, 654–665. https://doi.org/10.1016/j.ecolmodel.2009.07.023.
40. Storn, R.; Price, K. Differential Evolution—A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. *J. Glob. Optim.* **1997**, *11*, 341–359. https://doi.org/10.1023/a:1008202821328.
41. Hyndman, R.J.; Athanasopoulos, G. *Forecasting: Principles and Practice*, 3rd ed.; OTexts: Melbourne, Australia, 2021. Available online: https://otexts.com/fpp3/ (accessed on 30 September 2026).
42. Vollenweider, R.A. Input-output models: With special reference to the phosphorus loading concept in limnology. *Schweiz. Z. Hydrol.* **1975**, *37*, 53–84. https://doi.org/10.1007/BF02505178.
43. Brett, M.T.; Benjamin, M.M. A review and reassessment of lake phosphorus retention and the nutrient loading concept. *Freshw. Biol.* **2008**, *53*, 194–211. https://doi.org/10.1111/j.1365-2427.2007.01862.x.
44. Harrison, J.A.; Maranger, R.J.; Alexander, R.B.; Giblin, A.E.; Jacinthe, P.-A.; Mayorga, E.; Seitzinger, S.P.; Sobota, D.J.; Wollheim, W.M. The regional and global significance of nitrogen removal in lakes and reservoirs. *Biogeochemistry* **2009**, *93*, 143–157. https://doi.org/10.1007/s10533-008-9272-x.
45. Stow, C.A.; Reckhow, K.H. Estimator bias in a lake phosphorus model with observation error. *Water Resour. Res.* **1996**, *32*, 165–170. https://doi.org/10.1029/95WR03109.

46. Anagnostou, E.; Gianni, A.; Zacharias, I. Ecological modeling and eutrophication—A review. *Nat. Resour. Model.* **2017**, *30*, e12130. https://doi.org/10.1111/nrm.12130.
47. Zou, W.; Cao, Z.; Wang, X.; Huang, Q.; Xu, H.; Zhu, M.; Zhang, Y.; Qin, B.; Zhu, G. The Hydrological and Nutrient Conditions Mediate Algal Biomass Response to Compound Heatwave and Drought Events in Large Shallow Lakes. *Freshw. Biol.* **2025**, *70*, e70061. https://doi.org/10.1111/fwb.70061.
48. Chen, W.; Nielsen, A.; Andersen, T.K.; Hu, F.; Chou, Q.; Søndergaard, M.; Jeppesen, E.; Trolle, D. Modeling the Ecological Response of a Temporarily Summer-Stratified Lake to Extreme Heatwaves. *Water* **2020**, *12*, 94. https://doi.org/10.3390/w12010094.
49. Ding, Y.; Qin, B.; Zhu, G.; Wu, T.; Wang, Y.; Luo, L. Effects of typhoon Morakot on a large shallow lake ecosystem, Lake Taihu, China. *Ecohydrology* **2012**, *5*, 798–807. https://doi.org/10.1002/eco.270.
50. Bruggeman, J.; Burchard, H.; Kooi, B.W.; Sommeijer, B. A second-order, unconditionally positive, mass-conserving integration scheme for biochemical systems. *Appl. Numer. Math.* **2007**, *57*, 36–58. https://doi.org/10.1016/j.apnum.2005.12.001.
51. Wang, X.; Zhang, S.; Liu, S.; Chen, J. A two-dimensional numerical model for eutrophication in Baiyangdian Lake. *Front. Environ. Sci. Eng.* **2012**, *6*, 815–824. https://doi.org/10.1007/s11783-011-0383-6.
52. Podgornyi, K.A.; Leonov, A.V. The use of a spatially heterogeneous simulation model for studying biotransformation processes of nitrogen and phosphorus compounds and the dynamics of oxygen dissolved in water in the ecosystem of Neva Bay, the Gulf of Finland: 1. Model description. *Water Resour.* **2013**, *40*, 170–180. https://doi.org/10.1134/S0097807813020061.
53. Medvinsky, A.B.; Petrovskii, S.V.; Tikhonova, I.A.; Malchow, H.; Li, B.-L. Spatiotemporal complexity of plankton and fish dynamics. *SIAM Rev.* **2002**, *44*, 311–370. https://doi.org/10.1137/S0036144502404442.
54. Mao, J.; Chen, Q.; Chen, Y. Three-dimensional eutrophication model and application to Taihu Lake, China. *J. Environ. Sci.* **2008**, *20*, 278–284. https://doi.org/10.1016/S1001-0742(08)60044-3.
55. Khorasani, H.; Zhu, Z. Phosphorus retention in lakes: A critical reassessment of hypotheses and static models. *J. Hydrol.* **2021**, *603*, 126886. https://doi.org/10.1016/j.jhydrol.2021.126886.
56. Carleton, J.N.; Lee, S.S. Modeling lake recovery lag times following influent phosphorus loading reduction. *Environ. Model. Softw.* **2023**, *162*, 105642. https://doi.org/10.1016/j.envsoft.2023.105642.
57. Kondo, J. Air-sea bulk transfer coefficients in diabatic conditions. *Bound.-Layer Meteorol.* **1975**, *9*, 91–112. https://doi.org/10.1007/BF00232256.
58. Clark, N.E.; Eber, L.; Laurs, R.M.; Renner, J.A.; Saur, J.F.T. *Heat Exchange Between Ocean and Atmosphere in the Eastern North Pacific for 1961–71*; NOAA Technical Report NMFS SSRF-682; U.S. Department of Commerce: Seattle, WA, USA, 1974. Available online: https://spo.nmfs.noaa.gov/SSRF/SSRF682.pdf (accessed on 7 October 2026).


**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.