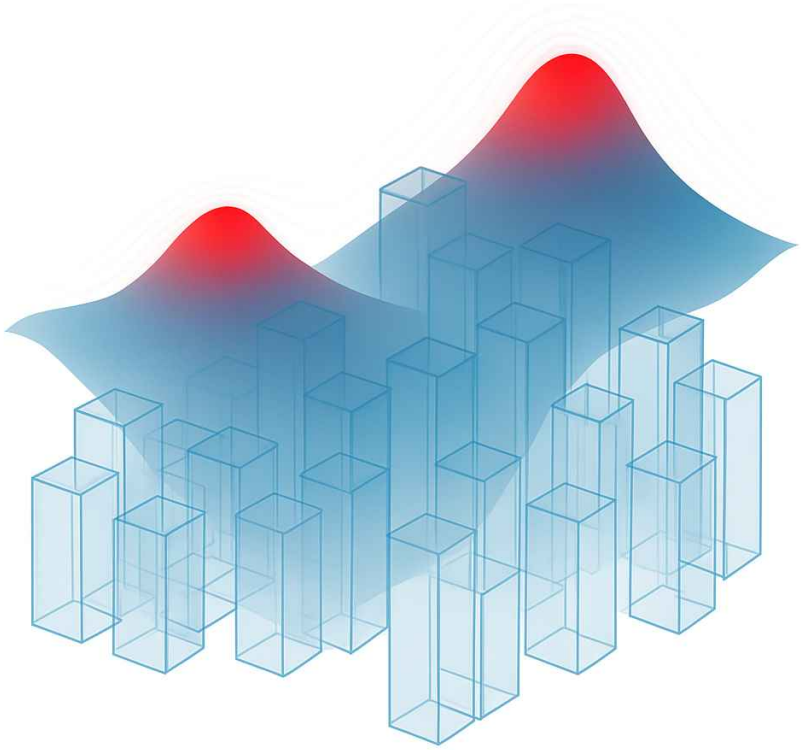
HOTSPOT: Hybrid Oceanic Tracking via Satellite Proxy and Optimized Time-Series



Tanuj Karthikeyan

Mentor: Dr. Daniel Egger

Research in Computational Science

North Carolina School of Science and Mathematics

27 September 2025

Abstract

Harmful algal blooms (HABs) driven by excessive phosphorus levels increasingly disrupt global marine ecosystems, yet existing monitoring frameworks remain regional, reactive, and reliant on *in-situ* sampling. HOTSPOT (Hybrid Oceanic Tracking via Satellite Proxy and Optimized Time-Series) establishes the first global framework capable of forecasting harmful algal blooms solely from satellite-derived measurements. The system integrates, for the first time, MODIS Aqua Level 3 products (radiance-based estimates of chlorophyll-a, sea-surface temperature, particulate organic carbon, and spectral absorption at 443 nm) from NASA's Aqua satellite with World Ocean Atlas phosphate climatology and HAEDAT bloom records (2018–2023; 77 sites). Features were normalized, seasonally encoded, and partitioned using a spatially stratified split to ensure model transferability across regions. A hybrid model combining logistic regression (70%) and random forest (30%) was trained with grouped cross-validation to maintain geographic independence and refined through a rule-based layer that filtered predictions using seasonal thresholds and cross-proxy consistency ($\tau_1 = 0.03$), retaining only the most reliable bloom detections each month. HOTSPOT achieved a true positive rate of 0.77 under spatially blocked evaluation across a test set of 31 sites, with AUROC = 0.68. By deriving predictions entirely from satellite observations, HOTSPOT demonstrates the feasibility of a scalable and data-driven early-warning system for HAB activity.

Acknowledgements

I would like to thank Dr. Egger for his guidance throughout the research process and for teaching my Research in Computational Science class. I also want to thank the Science Department, Dean Dr. Amy Schek, and my school, the North Carolina School of Science and Mathematics, for giving me the opportunity to conduct this research. Lastly, I would like to thank my family for all of their support.

1 Introduction

Harmful algal blooms (HABs) are now among the most consequential disruptions to marine ecosystems. Explosive algal growth strips oxygen from the water and releases toxins, generating hypoxic "dead zones" that kill fish, disrupt food webs, and threaten human health. The economic footprint is equally stark: fisheries and aquaculture endure costly shutdowns, while coastal tourism experiences long-lasting revenue losses (Anderson, Fensin, Gobler, Ho, Hubbard, Kulis, Landsberg, Lefebvre, Provoost, Richlen, Smith, Solow, and Trainer, 2021; Glibert, 2020). These blooms are impacted by local geographic conditions and are hard to predict.

Why phosphorus matters

Phosphorus (P) is a foundational nutrient for photosynthesis, ATP synthesis, and nucleic acids, anchoring productivity at the base of marine food webs. Unlike carbon or nitrogen, phosphorus has no gaseous phase; once introduced to aquatic systems, it cycles slowly through biomass, the water column, and sediments (Paytan and McLaughlin, 2007; Benitez-Nelson, 2000). In many coastal environments, P availability acts as a key limiter of primary productivity; when excess P enters aquatic systems, it often tips ecosystems toward eutrophication and bloom formation (Schindler, 2012; Paerl and Otten, 2016). However, global phosphorus resources are finite—economically accessible phosphate rock may be depleted within 100–300 years, implying a looming constraint on agriculture (Tirado and Allsopp, 2012). This duality (P scarcity on land and P surplus in waters) means that phosphorus tracking must be precise enough to prevent aquatic accumulation while preserving terrestrial reserves. The management challenge is therefore not merely detection, but *anticipation*: establishing early-warning systems capable of identifying bloom-prone conditions before visible outbreak.

The marine phosphorus cycle

Figure 1 summarizes how phosphorus cycles across land, river, ocean, and sediment reservoirs—moving between organic and inorganic forms and linking terrestrial inputs to marine accumulation. These cross-reservoir transfers motivate an early-warning approach that anticipates bloom-favorable conditions rather than reacting after onset.

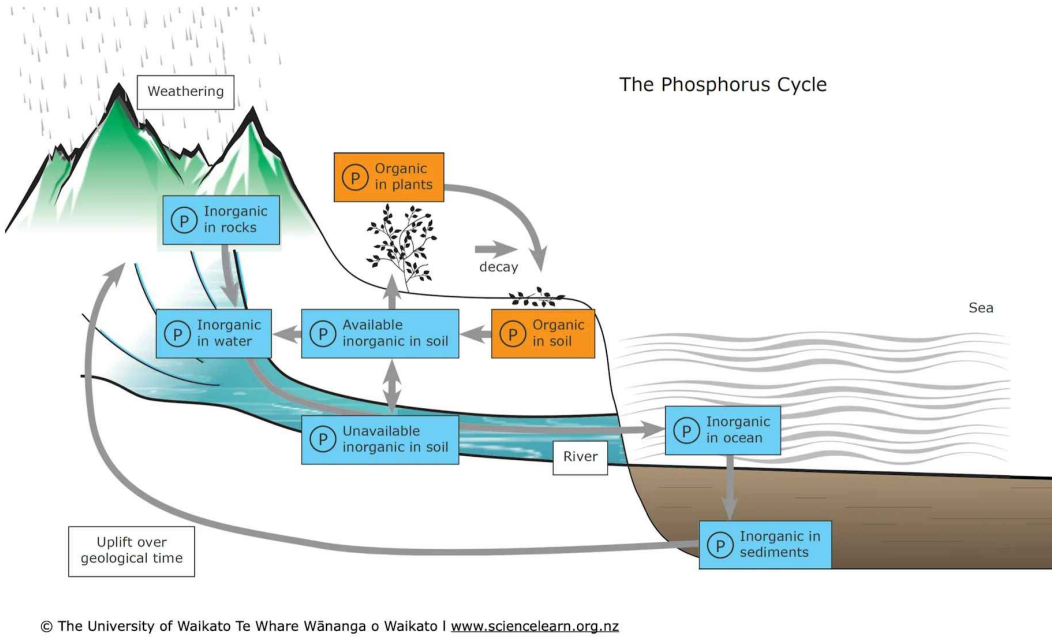


Figure 1: Phosphorus cycling across land–river–ocean–sediment reservoirs (University of Waikato, 2020).

In the ocean's surface mixed layer, dissolved inorganic phosphorus (DIP, primarily PO_4^{3-}) is assimilated into phytoplankton; some fraction recycles rapidly as dissolved organic phosphorus (DOP) or returns to DIP through microbial remineralization. Particulate organic matter sinks, exporting phosphorus to depth, while a smaller portion is buried as inorganic phosphorus in marine sediments, effectively removing it from short-term cycling until uplift and weathering re-expose it. Seasonal upwelling and winter mixing can reintroduce deep phosphorus to surface waters, linking the benthic and pelagic reservoirs. On the monthly scales relevant to HOTSPOT, the balance between external loading, remineralization, and vertical transport (advection and mixing) governs whether surface waters cross bloom-favorable thresholds. These coupled terrestrial and marine pathways explain how land–sea nutrient exchanges can prime coastal systems for HAB recurrence (Seitzinger et al., 2010; Beusen et al., 2016).

The monitoring and transferability gap

HAB monitoring programs have historically been regional and reactive, limited to sparse *in-situ* observations and post-bloom reporting. Satellite data, however, can provide global, continuous coverage of proxies such as chlorophyll-a (chl-a), particulate organic carbon (POC), sea-surface temperature (SST), and spectral absorption near 443 nm (aph(443)) (Sathyendranath et al., 2019; Hu et al., 2019; International Ocean-Colour Coordinating Group (IOCCG), 2019; Stramski et al., 2008; Bricaud et al., 1998). These variables capture

phytoplankton biomass, optical water properties, and temperature-driven mixing, yet have rarely been used for forward prediction. HAB models in operation perform well in their calibration regions but fail elsewhere due to variations in optical water properties, seasonal regimes, and nutrient baselines (Li et al., 2022). Freshwater systems such as Lake Erie have shown that satellite-derived features can be strongly predictive (Joshi et al., 2025), yet a global marine framework evaluated for transfer across coasts, shelves, and offshore waters remains missing. By integrating these satellite-based proxies with nutrient climatologies (e.g., World Ocean Atlas phosphate fields; Garcia et al., 2019) and validated HAB events from HAEDAT (HAEDAT Partnership, 2024), HOTSPOT bridges this gap—establishing a predictive, phosphorus-linked framework capable of global application.

Project Objective

This work develops HOTSPOT (Hybrid Oceanic Tracking via Satellite Proxy and Optimized Time-Series), a scalable and interpretable system that predicts phosphorus-driven bloom formation using global satellite proxies. MODIS Aqua products (chlorophyll-a (CHL), sea-surface temperature (SST), particulate organic carbon (POC), and absorption at 443 nm (aph_443)) are integrated with World Ocean Atlas phosphate fields to construct an early-warning classifier for bloom conditions. Model evaluation prioritizes true positive rate, as missed blooms are far costlier than false alarms, and enforces spatial independence to test geographic transfer. The approach combines machine learning with rule-informed validation to enable generalizable, transparent early-warning applications.

2 Data

This research assembles a global, monthly dataset designed specifically to predict phosphorus-linked harmful algal blooms (HABs). The guiding principle was *operational scalability*: all inputs must be global, consistent, and regularly updated so that the model can be applied anywhere in the world without site-specific calibration. To accomplish this, three main datasets were used and processed through a pipeline of steps: assembling raw sources, filling in missing spatial data, creating informative temporal features, and compiling the final variable summary.

2.1 Datasets and sources

MODIS Aqua Level-3 (4 km, monthly). NASA's MODIS Aqua satellite provides global measurements of ocean color and temperature (Sathyendranath et al., 2019; Hu et al., 2019;

International Ocean-Colour Coordinating Group (IOCCG), 2019). Four variables relevant to bloom dynamics were extracted: chlorophyll-a (chl-a), sea-surface temperature (SST), particulate organic carbon (POC), and spectral absorption at 443 nm (aph(443)) (Stramski et al., 2008; Bricaud et al., 1998). Daily observations were aggregated into monthly medians to reduce cloud and reflection artifacts, and each grid cell was assigned a coverage flag indicating the proportion of valid data.

World Ocean Atlas 2018 (WOA18). The World Ocean Atlas provides climatological averages of ocean nutrients, including phosphate (PO_4^{3-}) (Garcia et al., 2019). Although static, these fields offer spatial context for long-term phosphorus availability that complements the dynamic satellite record.

HAEDAT bloom events. The Harmful Algal Event Database (HAEDAT) (HAEDAT Partnership, 2024) supplied verified bloom occurrences between 2018 and 2023. Seventy-seven coastal sites with confirmed HAB events were identified. For each site, a $0.5° \times 0.5°$ region was extracted from MODIS data to capture surrounding conditions. Figure 2 summarizes how MODIS proxies, WOA18 phosphate climatology, and HAEDAT events are fused into a single, globally deployable dataset that supports site-independent training and evaluation.

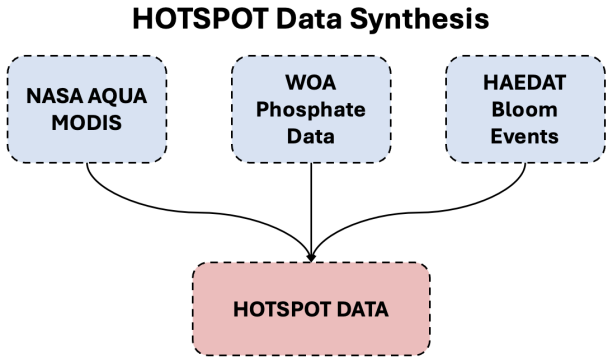


Figure 2: Overview of the combination of three data layers used in HOTSPOT. Figure created by student researcher using Microsoft PowerPoint.

2.2 Variable summary

The HOTSPOT dataset integrates optical, thermal, and biogeochemical indicators from satellite and climatological sources. Table 1 summarizes all predictor and target variables used in model training and evaluation.

Variable	Source	Unit	Temporal Res.
Chlorophyll-a (chl-a)	MODIS Aqua L3	mg/m^3	Monthly
Sea-surface temperature (SST)	MODIS Aqua L3	°C	Monthly
Particulate organic carbon (POC)	MODIS Aqua L3	mg/m^3	Monthly
Absorption at 443 nm (aph(443))	MODIS Aqua L3	1/m	Monthly
Phosphate (PO_4^{3-})	WOA18	µmol/L	Climatological
Bloom occurrence	HAEDAT	Binary	Monthly

Table 1: Environmental and target variables included in the HOTSPOT dataset.

Together, these layers capture complementary aspects of the marine environment (optical and thermal surface dynamics, background nutrient status, and verified bloom events), forming the foundation for HOTSPOT's predictive modeling.

2.3 Structure-preserving spatial gap-filling

Satellite datasets (especially those collected over coastal regions) often contain missing values caused by cloud cover, sunglint, or sensor noise. Simple interpolation methods, such as averaging neighboring pixels, tend to blur sharp spatial features that are ecologically important. To retain this detail, HOTSPOT applies a *structure-preserving gap-filling* procedure—conceptually similar to locally weighted surface fitting approaches used in oceanographic reconstruction studies (e.g., Beckers and Rixen, 2003; Garcia et al., 2010). These methods have been shown to reproduce fine-scale gradients in chlorophyll, temperature, and other biogeochemical tracers without introducing artificial smoothing, ensuring that critical mesoscale structure is conserved. Figure 3 illustrates that the structure-preserving reconstruction retains sharp mesoscale fronts and gradients—features along which HABs often initiate—instead of smoothing them away.

Step 1: Assigning weights to nearby pixels When a pixel at position x_0 is missing, valid neighboring pixels x_i are assigned exponentially decaying weights based on their distance from x_0:

$$w_i = \exp\left(-\frac{\|x_i - x_0\|^2}{2\sigma^2}\right), \tag{2.1}$$

where σ is a tunable parameter that controls how rapidly the influence of surrounding pixels

decreases with distance. Smaller σ values emphasize very local information, while larger values promote smoother interpolation. This weighting ensures that the imputation primarily reflects the characteristics of the nearby water mass.

Step 2: Representing the local surface The neighborhood around a missing pixel is approximated by a plane that captures variation in the field:

$$\hat{y}(x) = a + b^{\top}(x - x_0), \tag{2.2}$$

where x_0 is the missing pixel's location, a is its estimated value, and b is the gradient vector describing how the variable changes across space. The displacement $(x - x_0)$ represents horizontal distance from the center pixel to its neighbors. This linear form captures gentle spatial slopes (how values "tilt" locally) without the excessive smoothing produced by constant fits.

Step 3: Estimating plane parameters After assigning weights and defining the local plane, the next step is to determine the parameters a and b that best reproduce nearby observations. This is achieved by minimizing the weighted difference between each observed value y_i and its predicted value $\hat{y}_i$ from the plane:

$$J = \frac{1}{2} \sum_i w_i \, (y_i - \hat{y}_i)^2, \tag{2.3}$$

where closer pixels (with larger w_i) have stronger influence. Minimizing this objective yields a locally consistent surface that follows the spatial trend of valid data while maintaining sharp ecological boundaries. When enough valid neighbors are available, the parameters can be obtained directly using a weighted least-squares solution, which provides the plane that minimizes the total error.

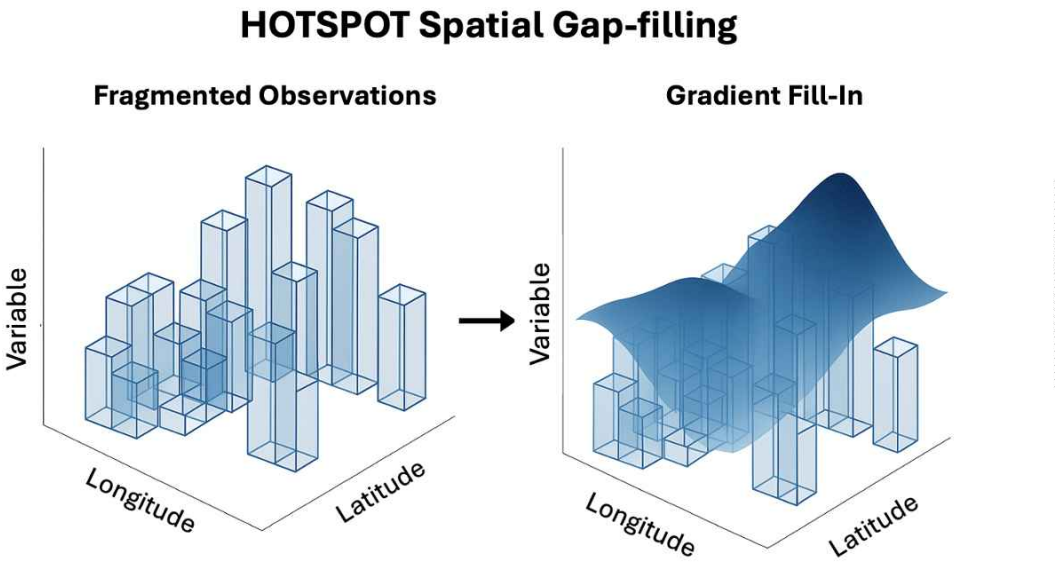


Figure 3: Gradient-aware gap-filling. The left panel shows fragmented satellite data; the right panel visualizes the smooth, structure-preserving surface reconstructed by HOTSPOT. Figure created by student researcher using Python (Matplotlib).

Result and importance The resulting imputed surface preserves natural gradients and boundaries, which is critical because harmful algal blooms (HABs) often form along fronts and sharp transitions in ocean conditions. By drawing on established, gradient-aware interpolation frameworks widely validated in oceanographic and climate studies (Beckers and Rixen, 2003; Garcia et al., 2010), HOTSPOT's approach ensures that the reconstructed fields are both physically and statistically rigorous, supporting reliable downstream analysis and model training.

2.4 Feature engineering

After filling spatial gaps, each site's time series is transformed into a set of features that describe how ocean conditions evolve over time. These features capture persistence, sudden change, short-term variability, and seasonality—patterns that often precede bloom formation but may be difficult to see in the raw data.

Lag and short-term change. To quantify persistence and abrupt transitions, each month is compared to the previous one:

$$\ell_t^{(v)} = x_{t-1}^{(v)}, \tag{2.4}$$

$$\Delta x_t^{(v)} = x_t^{(v)} - x_{t-1}^{(v)}. \tag{2.5}$$

The lag term $\ell_t^{(v)}$ represents continuity from the prior state (for instance, sustained high chlorophyll-a), while the difference term $\Delta x_t^{(v)}$ captures rapid shifts such as sudden warming or a sharp rise in biomass.

Short-term averages and variability. To highlight short-term patterns without losing longer trends, a three-month moving mean and sample variance are calculated:

$$\bar{x}_{t,3}^{(v)} = \frac{1}{3}\sum_{i=0}^{2} x_{t-i}^{(v)}, \qquad \mathrm{Var}_{t,3}^{(v)} = \frac{1}{2}\sum_{i=0}^{2}\left(x_{t-i}^{(v)} - \bar{x}_{t,3}^{(v)}\right)^2. \tag{2.6}$$

The mean $\bar{x}_{t,3}^{(v)}$ emphasizes broader tendencies, while the variance $\mathrm{Var}_{t,3}^{(v)}$ measures local instability that can precede bloom onset.

Seasonal encoding. Many oceanic variables vary cyclically with the calendar year. To represent this seasonality smoothly rather than as a discrete month label, each month m is mapped onto a unit circle using sine and cosine functions:

$$\sin_{\mathrm{mo}} = \sin\!\left(\frac{2\pi m}{12}\right), \quad \cos_{\mathrm{mo}} = \cos\!\left(\frac{2\pi m}{12}\right). \tag{2.7}$$

This encoding provides a continuous representation of the annual cycle, allowing the model to learn gradual transitions while recognizing that December and January are adjacent in time. For global application, a phase offset based on site-specific sea-surface temperature (SST) climatology is applied so that each region's encoding aligns with its local seasonal cycle, ensuring that "summer" and "winter" are correctly oriented across hemispheres.

Normalization. Before training, all features were rescaled so that they could be compared on the same scale. Each feature value f was standardized by subtracting its mean and dividing by its standard deviation, both calculated from the training set:

$$f' = \frac{f - \mu_{\mathrm{train}}}{\sigma_{\mathrm{train}}}, \tag{2.8}$$

where μ_{train} and σ_{train} denote the mean and standard deviation of that feature within the training data. This transformation shifts every feature to have a mean of zero and a standard deviation of one. Using only training-set statistics prevents information from the test set from influencing model training, thereby avoiding data leakage and maintaining a fair evaluation. Global z-score normalization is widely used in spatial machine learning and Earth-system modeling (Hengl et al., 2017; Valavi et al., 2019; Ren et al., 2021) because it preserves regional differences and allows models to work across diverse environments. In contrast, most

regional HAB studies use site-specific normalization (Li et al., 2022; Joshi et al., 2025), which removes large-scale environmental patterns and limits how well models transfer to new areas. HOTSPOT adopts global, training-set normalization to follow these broader best practices and support reliable cross-coast prediction.

3 Methodology

After feature preparation, HOTSPOT learns how satellite-observed patterns translate into bloom likelihood. The modeling framework emphasizes robustness under geographic transfer and consistency with established ecological practices (Olden et al., 2008; Elith and Leathwick, 2009; Hengl et al., 2017; Ren et al., 2021).

3.1 Model design and purpose

Environmental prediction requires balancing interpretability with predictive flexibility (Breiman, 2001; Overpeck et al., 2011). Simple statistical models such as logistic regression (LR) provide transparent cause–effect relationships, while more flexible ensemble learners such as random forests (RF) capture nonlinear, threshold-based responses common in ecological systems (Hao and Singh, 2020; Elith and Leathwick, 2009; Santana et al., 2023). HOTSPOT integrates both, leveraging the mechanistic clarity of LR and the adaptive precision of RF in an ensemble formulation.

1. Logistic Regression (LR). The LR component estimates bloom probability from environmental predictors by first computing a linear combination of inputs:

$$z = \beta_0 + \sum_{i=1}^{n} \beta_i x_i, \tag{3.1a}$$

$$P_{\mathrm{LR}} = \frac{e^z}{1 + e^z}, \tag{3.1b}$$

where x_i are standardized predictors (e.g., CHL, SST, POC), and β_i are the fitted coefficients describing both direction (sign) and intensity (magnitude) of association. Positive β_i values increase bloom likelihood, while negative ones reduce it. Because each coefficient represents an independent marginal effect, the LR term retains mechanistic interpretability—managers can directly infer how individual environmental variables influence bloom probability.

2. Random Forest (RF). The RF ensemble complements LR by learning nonlinear

interactions through a set of T decision trees $\{h_t(x)\}_{t=1}^{T}$ trained on bootstrapped samples:

$$P_{\mathrm{RF}} = \frac{1}{T}\sum_{t=1}^{T} h_t(x), \tag{3.2}$$

where each $h_t(x)$ partitions predictor space using thresholds that minimize within-node impurity (Gini index). RF thus captures emergent patterns such as "if–then" rules—for example, identifying that high CHL only indicates blooms under moderate SST, or that high POC is predictive only when aph(443) absorption exceeds a threshold. Feature importance values quantify how much each variable reduces impurity across the forest, summarizing nonlinear and joint effects rather than single-variable slopes.

3. Hybrid integration. The two models are combined in a convex ensemble:

$$P_{\mathrm{final}} = 0.7\,P_{\mathrm{LR}} + 0.3\,P_{\mathrm{RF}}, \tag{3.3}$$

where weights were tuned via out-of-fold calibration to optimize balanced accuracy while preserving interpretability. This architecture allows HOTSPOT to exploit the LR's mechanistic transparency while letting RF correct for complex nonlinearities and cross-variable dependencies that linear forms cannot model. The ensemble therefore balances explainability and generalization—a necessity for ecological decision support systems.

Variable	**LR Coefficient (Sign / Magnitude)**	**RF Feature Importance**
Chlorophyll-a (CHL)	+0.302	0.203
Sea-surface temperature (SST)	−0.002	0.226
Particulate organic carbon (POC)	−0.335	0.201
Absorption at 443 nm (aph(443))	+0.111	0.206
Seasonal encoding (sin)	+0.110	0.063
Seasonal encoding (cos)	−0.157	0.102

Table 2: Relative influence of major predictors across logistic and ensemble components.

The feature importance analysis reveals how the model makes its decisions and which variables most strongly influence bloom prediction. Biologically, warmer surface waters (higher SST) tend to suppress bloom likelihood in this dataset, explaining their strongly negative contribution. In contrast, elevated phytoplankton biomass and optical absorption (higher CHL and aph(443)) are positively associated with bloom conditions, while the negative

POC coefficient aligns with periods dominated by non-algal particulates or detritus that do not indicate active blooms.

3.2 Spatial blocking and validation

Reliable performance assessment requires testing the model under conditions that mimic deployment on coastlines not represented in the training data. Conventional random cross-validation can lead to inflated accuracy estimates because geographically proximate samples often share similar oceanographic or atmospheric conditions. As a result, information may inadvertently "leak" across folds, causing the model to overfit to local spatial autocorrelation rather than learning transferable ecological relationships.

To address this, HOTSPOT adopts a **spatial blocking** strategy following best practices established in ecological and remote-sensing studies (e.g., Roberts et al., 2017; Valavi et al., 2019). In these frameworks, data are partitioned into spatially coherent regions (such as environmental or geographic blocks) and validation is performed by systematically holding out entire regions. This design explicitly removes spatial dependence between training and test sets, producing accuracy estimates that better reflect real-world generalization.

In HOTSPOT, coastal sites are divided into $10^\circ \times 10^\circ$ latitude–longitude blocks (approximately 1,000 km on a side). During cross-validation, entire blocks are withheld as independent test folds, while non-adjacent blocks are used for model training. This ensures that no test site shares an immediate spatial neighborhood, current system, or water mass with any training site. By emulating how the model would be deployed on an unmonitored coastline, the blocking procedure provides a scientifically rigorous estimate of predictive performance under "new coast" conditions. Moreover, evaluating performance across distinct oceanic provinces exposes potential regional biases—revealing, for example, if a model generalizes well in temperate waters but struggles in tropical regions. To enforce geographic independence during validation, Figure 4 presents the $10^\circ \times 10^\circ$ folds used for spatially blocked cross-validation, which emulate deployment on new coastlines.

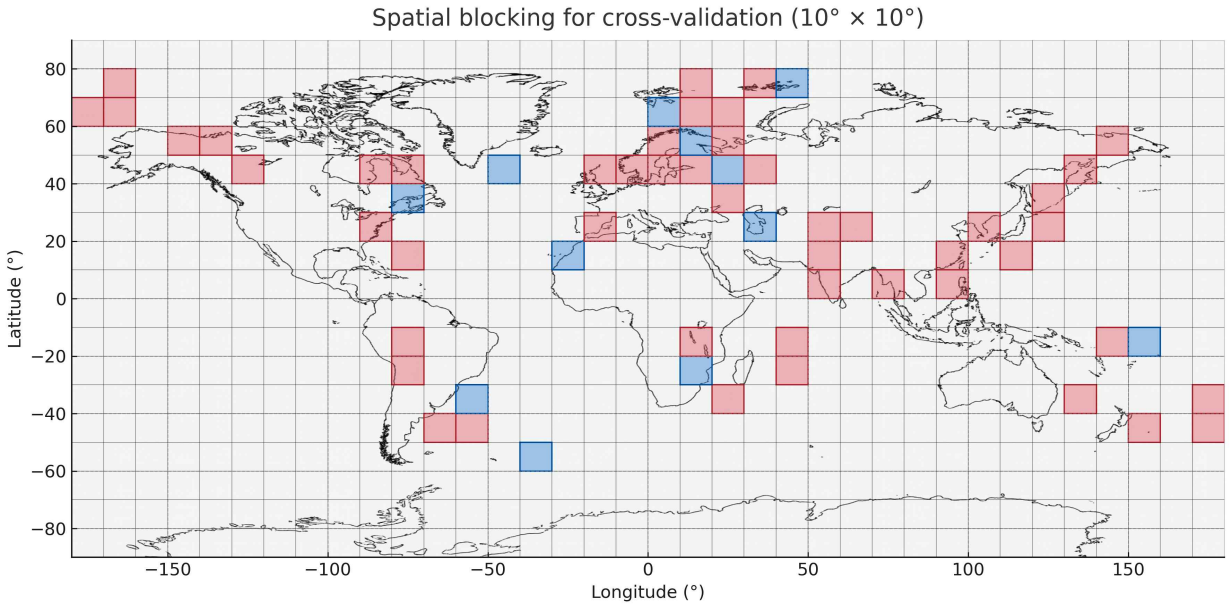


Figure 4: Spatial blocking for cross-validation. Each colored cell represents a geographically independent $10° \times 10°$ fold used to separate training (red) and testing (blue) regions while preserving spatial independence across sites. Figure created by student researcher using Python (Cartopy and Matplotlib).

3.3 Handling class imbalance

In the dataset, bloom months are far less frequent than non-bloom months—often by an order of magnitude. Without correction, a model could achieve deceptively high accuracy by predicting "no bloom" everywhere while failing to detect true events. Instead of oversampling minority cases or discarding normal months (which could distort spatial patterns), HOTSPOT uses **cost-sensitive learning** to explicitly penalize bloom misclassifications more heavily (He and Garcia, 2009; Beery et al., 2021).

The training loss is a weighted binary cross-entropy:

$$L = -\alpha\, y \log(\hat{y}) - (1-\alpha)(1-y)\log(1-\hat{y}), \tag{3.4}$$

where $\hat{y}$ is the model's predicted bloom probability and y is the true label (1 for bloom, 0 for non-bloom). The weight α controls the tradeoff between *recall* (catching more true blooms) and *precision* (avoiding false alarms). A larger α increases the cost of missing a bloom, nudging the model to err on the side of caution—appropriate for early-warning systems where a missed event is far riskier than an extra check. This asymmetric cost structure directly embeds the operational principle that it is better to investigate a few false positives than to

overlook a harmful bloom.

3.4 Policy layer and ecological filtering

Purely statistical predictions can sometimes flag events that are ecologically implausible. To ensure biological realism and operational usefulness, HOTSPOT applies a compact **policy layer** that filters and prioritizes results according to ecological rules (Wells et al., 2020; Anderson et al., 2021; Stow et al., 2019). These rules translate raw model outputs into decisions that are more consistent with known bloom behavior:

1. **Seasonal gate:** Suppresses alerts outside each site's local bloom window, derived from sea-surface temperature (SST) climatology and historical month-of-year patterns. This ensures that predictions align with local phenology and do not trigger outside plausible bloom seasons.

2. **Proxy corroboration:** Requires agreement among at least two independent satellite proxies (e.g., chl-*a*, POC, or aph(443) change) within the same temporal window to reduce single-sensor artifacts and strengthen confidence in detected events (International Ocean-Colour Coordinating Group (IOCCG), 2019; Hu et al., 2019). Only predictions supported by consistent multi-proxy signals advance to the final decision layer.

These filters act as an ecological safeguard—reducing false alarms, preserving interpretability, and ensuring that model detections remain physically and biologically credible. Rather than relying on a fixed numerical threshold, HOTSPOT enforces a flexible corroboration rule emphasizing pattern consistency among independent optical and biogeochemical indicators. This approach retains genuine bloom events that are supported by multiple proxies while discarding spurious detections caused by transient noise, cloud contamination, or sensor drift. Figure 5 shows the policy layer that converts raw probabilities into operational alerts via seasonal gating and multi-proxy corroboration, reducing ecologically implausible flags.

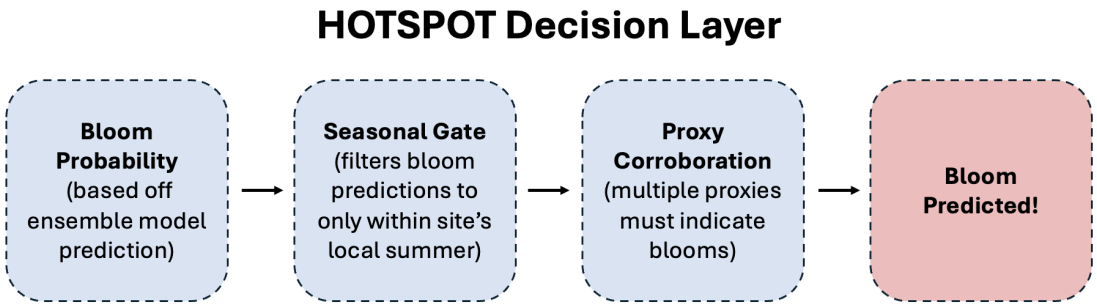


Figure 5: Policy layer workflow showing how model probabilities are filtered to yield bloom alerts. Figure created by student researcher using Microsoft PowerPoint.

3.5 Thresholding and diagnostic checks

Once model probabilities pass ecological filtering, they must be converted into actionable bloom or no-bloom predictions. HOTSPOT uses a single decision threshold to separate these two outcomes, chosen to maximize the F_1 score—the harmonic mean of precision and recall:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

This operating point balances the need to detect as many true blooms as possible (high recall) while limiting false positives (maintaining precision). After the optimal threshold is determined using training data, it is applied to all test folds to ensure comparable evaluation.

To ensure robustness, several **diagnostic checks** were performed. These include verifying that threshold performance remained stable across geographically distinct validation folds, confirming that recall and precision tradeoffs were consistent with ecological expectations, and ensuring that predictions respected the spatial blocking design. Together, these diagnostics confirm that observed performance reflects genuine predictive skill rather than random threshold effects or site-specific bias.

4 Results & Discussion

HOTSPOT's evaluation proceeds from classification outcomes to threshold and discrimination analyses, illustrating the model's ability to generalize under spatially independent validation. All metrics (accuracy, recall [true positive rate, TPR], precision [positive predictive value, PPV], F_1, and ROC–AUC) were computed under the spatially blocked protocol described in Section 3, using $10° \times 10°$ blocks with no shared sites between training and testing folds. This

approach mirrors real deployment on new coastlines and provides a conservative estimate of skill (Roberts et al., 2017; Valavi et al., 2019).

Model	Accuracy	Recall (TPR)	Precision (PPV)	F_1	ROC–AUC
HOTSPOT (LR 0.7 + RF 0.3)	0.636	0.765	0.156	0.259	0.677

Table 3: Held-out performance under spatial blocking.

Table 3 summarizes the ensemble's cross-site performance. While precision (0.156) remains modest due to bloom rarity, recall (0.765) indicates strong sensitivity—an intentional design priority given the ecological cost of missed detections. The AUROC (0.677) exceeds the random baseline, demonstrating that HOTSPOT distinguishes bloom-forming conditions beyond chance even when trained and evaluated on geographically distinct regions.

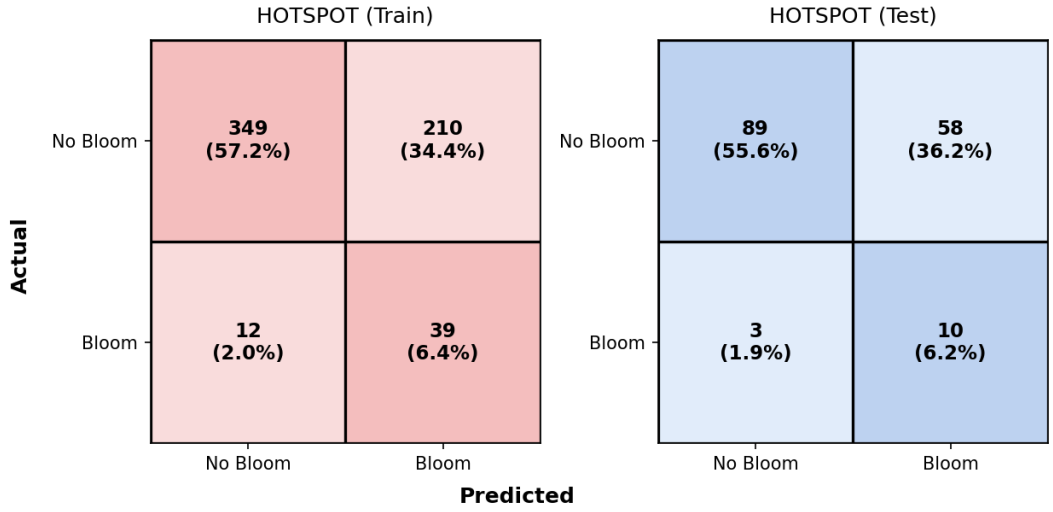


Figure 6: Confusion matrices for HOTSPOT model performance. Figure created by student researcher using Python (Seaborn and Matplotlib).

Classification behavior. Consistent with our recall-first posture, Figure 6 contrasts training and testing confusion matrices, highlighting increased false positives but low false negatives on held-out regions. On training data, the model achieves over 95% correct classification, dominated by true negatives (non-bloom periods). On testing data, however, false positives increase (6.2%), reflecting the model's intentional bias toward early detection. This design choice prioritizes early warnings over precision loss—appropriate for operational monitoring where missing an event is costlier than over-alerting. The low false-negative rate (1.9%) on testing data demonstrates stable sensitivity even under unseen spatial distributions.

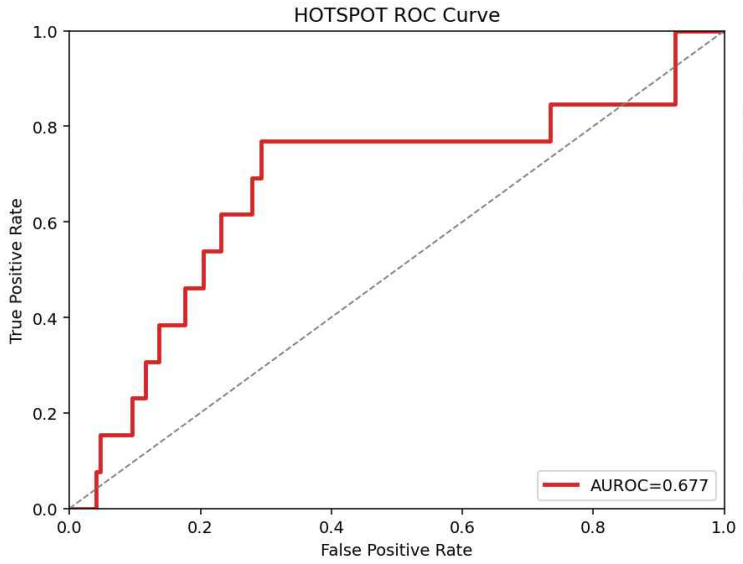


Figure 7: Receiver Operating Characteristic (ROC) curve for the final HOTSPOT ensemble under spatially blocked testing. Figure created by student researcher using Python (Seaborn and Matplotlib).

Discrimination under imbalance. Figure 7 demonstrates threshold-free discrimination under spatial blocking (AUROC = 0.677), indicating the model ranks bloom-forming months above non-bloom months beyond chance. This metric quantifies how well the model orders true events above negatives regardless of threshold choice, capturing its ability to generalize bloom-relevant structure rather than memorizing local optical conditions. Unless otherwise stated, all results reflect the post-policy outputs, where ecological filters ensure that only seasonally and cross-proxy consistent detections are evaluated.

Interpretation and implications. Taken together, these metrics suggest that HOTSPOT effectively generalizes bloom-relevant structure across coasts, maintaining recall despite regional variation in optical and thermal profiles. The modest but consistent precision aligns with the expectation for satellite-based early warning systems operating under imperfect ground truth. In practice, these predictions would serve as spatial "risk maps," guiding further *in-situ* testing or remote sensing follow-up rather than direct binary classification decisions. The framework thus demonstrates feasibility for phosphorus-linked harmful algal bloom forecasting at continental to global scales using interpretable, ensemble-based inference.

5 Conclusion

HOTSPOT demonstrates that satellite-based monitoring can move from passive observation to proactive forecasting. By connecting patterns in temperature, chlorophyll, carbon, and

optical absorption to bloom likelihood, the system turns existing remote-sensing data into a tool for early environmental action rather than post-event response.

Several lessons emerge from this work. First, global HAB risk can be forecast without direct nutrient measurements, provided that satellite features are processed with spatial care and ecological context. Second, the combination of logistic regression and random forest provides both explanation and adaptability. Third, embedding domain-informed rules—seasonal limits, proxy corroboration, and ranked prioritization—substantially improves the realism of alerts.

Still, HOTSPOT is not a complete solution. Cloud gaps, coarse temporal averaging, and inconsistent event records all introduce uncertainty. The path forward lies in higher-frequency data, region-specific thresholds, and continued collaboration between computational models and field experts.

In sum, HOTSPOT points to a broader shift: from static mapping of the ocean to dynamic anticipation of its changes. By combining open satellite data with transparent modeling, it offers a foundation for scalable, data-driven stewardship of marine ecosystems.

References

Donald M. Anderson, Eliska Fensin, Christopher J. Gobler, Jackson Ho, Katherine A. Hubbard, David M. Kulis, Jan H. Landsberg, Kathleen A. Lefebvre, Pieter Provoost, Mindy L. Richlen, Jessica L. Smith, Andrew R. Solow, and Vera L. Trainer. Marine harmful algal blooms (habs) in the united states: History, current status and future trends. *Harmful Algae*, 102:101975, 2021. doi: 10.1016/j.hal.2021.101975.

J. M. Beckers and M. Rixen. Eof calculations and data reconstruction from incomplete oceanographic datasets. *Journal of Atmospheric and Oceanic Technology*, 20(12):1839–1856, 2003. doi: 10.1175/1520-0426(2003)020<1839:ECADRF>2.0.CO;2.

Sara Beery, Grant Horn, and Oisin Mac Aodha. The iwildcam 2020 competition dataset: Evaluating classification and detection for wildlife at camera trap sites. *arXiv preprint arXiv:2004.10340*, 2021.

Claudia R. Benitez-Nelson. The biogeochemical cycling of phosphorus in marine systems. *Earth-Science Reviews*, 51(1–4):109–135, 2000. doi: 10.1016/S0012-8252(00)00018-0.

Arthur H. W. Beusen, Lex Bouwman, and Leo Van Beek. Global riverine n and p transport to ocean increased during the 20th century despite increased retention along the aquatic continuum. *Biogeosciences*, 13(8):2441–2451, 2016. doi: 10.5194/bg-13-2441-2016.

Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231, 2001.

Annick Bricaud, André Morel, and Marcel Babin. Variability in the chlorophyll-specific absorption coefficients of natural phytoplankton: Analysis and parameterization. *Journal of Geophysical Research: Oceans*, 103(C13):31033–31044, 1998.

Jane Elith and John R. Leathwick. Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, 40:677–697, 2009.

Hernan E. Garcia, Tim P. Boyer, Ricardo A. Locarnini, Alexey V. Mishonov, and Olga K. Baranova. World ocean atlas 2018 (woa18): Volume 4—nutrients (phosphate, nitrate, silicate), 2019.

R. A. Garcia et al. A data interpolating empirical orthogonal function (dineof) method for filling missing satellite data: Application to modis chlorophyll-a in the adriatic sea. *Remote Sensing of Environment*, 114:698–707, 2010. doi: 10.1016/j.rse.2009.11.004.

Patricia M. Glibert. Harmful algae at the complex nexus of eutrophication and climate change. *Harmful Algae*, 91:101583, 2020. doi: 10.1016/j.hal.2019.03.001.

HAEDAT Partnership. Harmful algal event database (haedat). `https://haedat.iode.org/`, 2024.

Ziqi Hao and Vijay P. Singh. Forecasting hydrological extremes using ensemble machine learning models. *Environmental Modelling & Software*, 124:104587, 2020.

Haibo He and Edward A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

Tomislav Hengl, Jorge Mendes de Jesus, Gerard B. M. Heuvelink, et al. Soilgrids250m: Global gridded soil information based on machine learning. *PLoS ONE*, 12(2):e0169748, 2017. doi: 10.1371/journal.pone.0169748.

Chuanmin Hu, Brian Barnes, Gary Mitchum, Susan Craig, and Frank Muller-Karger. Modis observations of chlorophyll variability: Implications for hab early warning. *Remote Sensing of Environment*, 225:347–360, 2019. doi: 10.1016/j.rse.2018.02.035.

International Ocean-Colour Coordinating Group (IOCCG). *IOCCG Report 18: Earth Observations in Support of Global Water Quality Monitoring*. IOCCG, Dartmouth, Canada, 2019.

A. Joshi, Y. Zhang, et al. Satellite-based machine learning for harmful algal bloom prediction in lake erie. *Remote Sensing of Environment*, 300:113948, 2025. doi: 10.1016/j.rse.2024.113948.

Wei Li, Han Zhang, Ming Chen, and Qiang Liu. Random forest-based hab prediction in lake chaohu: Local performance and limits to transferability. *Science of the Total Environment*, 820:153278, 2022.

Julian D. Olden, Joshua J. Lawler, and N. LeRoy Poff. Machine learning methods without tears: A primer for ecologists. *The Quarterly Review of Biology*, 83(2):171–193, 2008.

Jonathan T. Overpeck, Gerald A. Meehl, Sandrine Bony, and David R. Easterling. Climate data challenges in the 21st century. *Science*, 331(6018):700–702, 2011.

Hans W. Paerl and Timothy G. Otten. Duelling 'cyanohabs': Unravelling the environmental drivers controlling dominance and succession among diazotrophic and non-diazotrophic cyanobacteria. *Environmental Microbiology*, 18(2):316–324, 2016. doi: 10.1111/1462-2920.13035.

Adina Paytan and Karen McLaughlin. The oceanic phosphorus cycle. *Chemical Reviews*, 107 (2):563–576, 2007. doi: 10.1021/cr0503613.

Jia Ren, Xuefeng Li, and Jinlong Yu. Machine learning in ocean remote sensing: Methods, applications, and future prospects. *Remote Sensing of Environment*, 264:112611, 2021. doi: 10.1016/j.rse.2021.112611.

David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929, 2017. doi: 10.1111/ecog.02881.

Giulia Santana, Alex Costa, and Marcela Souza. Hybrid ensemble learning for predicting water quality index in coastal environments. *Marine Pollution Bulletin*, 188:114693, 2023.

Shubha Sathyendranath, Robert J. W. Brewin, Carsten Brockmann, et al. Ocean colour climate change initiative—approach and applications. *Remote Sensing of Environment*, 230:111196, 2019. doi: 10.1016/j.rse.2019.111196.

David W. Schindler. The dilemma of controlling cultural eutrophication of lakes. *Proceedings of the Royal Society B: Biological Sciences*, 279(1746):4322–4333, 2012. doi: 10.1098/rspb.2012.1032.

Sybil P. Seitzinger, Emilio Mayorga, Lex Bouwman, et al. Global river nutrient export: A scenario analysis of past and future trends. *Global Biogeochemical Cycles*, 24:GB0A08, 2010. doi: 10.1029/2009GB003587.

Craig A. Stow, Donald Scavia, and Kenneth H. Reckhow. Improving the usability of ecological models for management decision making. *Environmental Modelling & Software*, 122:104521, 2019.

Dariusz Stramski, Rick A. Reynolds, Marcel Babin, et al. Relationships between the surface concentration of particulate organic carbon and optical properties in the ocean. *Biogeosciences*, 5(1):171–201, 2008.

Reyes Tirado and Michelle Allsopp. Phosphorus in agriculture: Problems and solutions. Technical Report Technical Report, Greenpeace Research Laboratories, 2012.

University of Waikato. Phosphorus cycle diagram. `https://www.sciencelearn.org.nz/resources/1666-the-phosphorus-cycle`, 2020.

R. Valavi, J. Elith, J. J. Lahoz-Monfort, and G. Guillera-Arroita. blockcv: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, 10(2):225–232, 2019. doi: 10.1111/2041-210X.13107.

Mark L. Wells, Bengt Karlson, Angela Wulff, Raphael Kudela, Charles Trick, Valentina Asnaghi, and Elisa Berdalet. Future hab science: Directions and challenges in a changing climate. *Harmful Algae*, 91:101632, 2020. doi: 10.1016/j.hal.2019.101632.