Shifting paradigms in Ocean Color: Bayesian Inference for Uncertainty-Aware Chlorophyll Estimation

Erdem M. Karaköylü[1], Susanne E. Craig[2]

[1]Consultant,
[2]NASA,

Corresponding author: Erdem M. Karaköylü, erdemk@protonmail.com

Abstract

Placeholder

Plain Language Summary

Placeholder

0.1 Introduction

0.1.1 Historical Context of Chlorophyll Algorithms

Satellite ocean color observations have long been fundamental for monitoring marine ecosystems, as they enable the global estimation of chlorophyll-a—a key indicator of phytoplankton biomass and ocean productivity. Early empirical algorithms, notably developed by O'Reilly et al. (John E. O'Reilly et al., 1998; John E. O'Reilly et al., 2000), established the OCx family (where x denotes the number of bands used) of polynomial regression models. These models relate blue-to-green reflectance ratios (after log-transformation) to in situ chlorophyll-a, employing either straight band ratios (BR) or maximum band ratios (MBR)—the latter selecting the highest available blue-to-green ratio for any given observation as input to a high-order polynomial. These formulations have served as the operational foundation for chlorophyll-a products across a broad range of satellite ocean color sensors—from the pioneering Coastal Zone Color Scanner (CZCS) through SeaWiFS, MODIS, and MERIS to more recent missions—offering a straightforward and robust approach for Case-1 waters. However, their performance is more limited in optically complex Case-2 waters and remains sensitive to atmospheric correction errors.

Subsequent refinements were introduced to address these deficiencies. For example, Hu et al. (Hu et al., 2012) proposed a Color Index (CI) formulation that employs a band-difference approach to reduce sensitivity to residual atmospheric errors and instrument noise, with further improvements enhancing inter-sensor agreement (Hu et al., 2019). The increasing availability of calibration data (e.g., (**Valente2015?**)) and ongoing algorithmic improvements have led to the development of additional variants of the OCx algorithms—specifically, the OC5 and OC6 formulations. O'Reilly and Werdell (John E. O'Reilly & Werdell, 2019) maintain that OC5 extends the spectral basis by incorporating the 412 nm band, thereby exploiting its strong signal in clear, oligotrophic waters, while OC6 replaces the traditional denominator with the mean of the 555 and 670 nm reflectances, improving the dynamic range at low chlorophyll concentrations. In total, (John E. O'Reilly & Werdell, 2019) propose 65 versions of BR/MBR OCx type algorithms for 25 sensors—on average, two or more variants per sensor. With this arsenal, it is hoped, researchers are better equipped to address the wide array of bio-optical environments encountered in global ocean color applications.

0.1.2 Limitations of Existing Approaches

Regrettably, the development of traditional ocean color algorithms relies on a fundamental statistical error. This error originates in the conflation of empirical frequencies — derived from a limited, noisy set of observations — with the theoretical relative frequencies that would be observed over an infinite number of trials, thereby ignoring the intrinsic uncertainty and variability inherent in finite sampling. The fallout has and continues to be extremely detrimental to observational science. Traditional, frequentist approaches, where the variability in sample frequencies is treated as negligible, yield overconfident inferences, and allegedgly significant findings that do not hold water, in time leading to replication crises.

In statistical modeling, the goal is to assess how well a hypothesized model M represents the relationship between independent and dependent variables based on collected data D. Frequentist methods focus exclusively on the likelihood $p(D|M)$, estimating model parameters by maximizing this probability. This approach assumes that, with a sufficiently large dataset, the chosen parameters reliably reflect the true system. However, in practice, any finite dataset carries inherent uncertainty and noise. As a consequence, models developed under this framework tend to be overconfident - yielding point estimates that do not adequately capture the variability of *in-situ* observations - and may not generalize well to new data. Ad hoc attempts to generate uncertainty, such as bootstrapping, further compound the problem by resampling the flawed, finite observations and thus amplifying any existing biases. This overconfidence, coupled with unverified assumptions presented as objectivity, can lead to misleading inferences and contribute to replication challenges. The historical process by which these assumptions have been enshrined is fascinating in its own right, and we invite interested readers to consult works such as McGrayne (**mcgrayne?**) and Clayton (**clayton?**).

As elaborated extensively by Jaynes (**jaynes2003probability?**), and more recently by practitioners like Gelman (**gelman2013bayesian?**), and McElreath (**mcelreath2015statistical?**), probability is best understood as a degree of knowledge rather than as a long-run frequency. Accordingly in this Bayesian framework, the prior $p(M)$ encodes our current state of knowledge about the model's parameters before D is observed. This knowledge can be elicited from experts, or if none exixts weakly informative priors can be used. Note that whether M is based on empirically observed relationships, analytical understanding, or something in between is irrelevant to its integration into this framework. Once the data D are collected, Bayes' theorem combines the likelihood $p(D|M)$ with the prior $p(M)$ to yield the posterior distribution $p(M|D)$, which represents an updated level of knowledge about the model's parameters. Crucially this posterior can then serve as the new prior when new data become available, allowing for sequential integration of new information. The by-product of this approach is uncertainty estimates that more accurately reflect the complexities of the underlying system as currently understood. In effect, rather than relying on single point estimates based on $p(D|M)$, the posterior $p(M|D)$ can be mined to retrieve a range of plausible parameter values through credible intervals offering a more transparent and robust basis for inference. This principled integration of prior knowledge with observed data is central to our approach, as detailed further in the Supplementary Material.

In this paper we recast a number of OCx-like algorithms in the Bayesian framework. We also explore a different kind of

0.1.3 Motivation for Bayesian Tree-Based Models

To address these limitations, we explore Bayesian tree-based modeling – specifically, Bayesian Additive Regression Trees (BART) – as a novel approach for ocean color chlorophyll retrieval. BART offers several advantages for this application:

Direct Bayesian inference: As a Bayesian ensemble method, BART directly targets the posterior distribution $p(Chl|Rrs)$ rather than a purely empirical fit. The model inherently provides a probabilistic mapping from reflectance to chlorophyll. Rather than producing a single best-fit value, BART returns a posterior mean and credible interval for each prediction, naturally quantifying predictive uncertainty.

Uncertainty quantification: The BART framework includes uncertainty in its predictions by design. Every estimated chlorophyll value comes with a confidence range (e.g., 95% credible interval) reflecting model uncertainty and observational noise. This is a key improvement over deterministic algorithms, enabling more reliable use of the data for scientific and operational purposes where understanding uncertainty is crucial e.g. for forecasting trends.

Interpretability: Being a sum-of-trees model, BART can be interrogated to understand the influence of each input variable. Tools such as partial dependence plots allow us to visualize the relationship between each spectral band and the predicted chlorophyll, marginalizing over other bands. In contrast to high-order polynomial coefficients, these tree-based partial dependencies are intuitively interpretable. For instance, one can observe how changes in a particular waveband's reflectance (holding others average) drive the chlorophyll estimate, revealing any thresholds or non-linear responses.

BART's modeling approach is non-parametric and highly flexible, which means it can capture complex, non-linear relationships between multi-spectral Rrs inputs and chlorophyll without imposing a rigid functional form. At the same time, its Bayesian regularization – which shrinks the contribution of each individual tree – helps prevent overfitting even with a rich model structure. This balance between flexibility and regularization is well-suited to global ocean color data, where the true Rrs–Chl relationship is known to be non-linear and context-dependent, yet a model must generalize across diverse water types.

The interpretability of the BART model is illustrated by partial dependence analysis on the six input Rrs bands (centered approximately at 412, 443, 490, 510, 555, and 670 nm). The partial dependence plots reveal nuanced spectral–chlorophyll relationships and clear transitions in the dominant wavelength influence across different chlorophyll regimes. For example, the shorter blue wavelengths (412–443 nm) show a strong inverse influence on predicted $log(Chl_a)$ (higher blue reflectance corresponds to lower chlorophyll) that gradually saturates in the most oligotrophic waters, consistent with the leveling off of blue-to-green ratios at very low Chl. The green band (~555 nm) exhibits a more complex, non-monotonic effect: at low reflectance (clear water) its influence on the Chl estimate is minimal, but it becomes increasingly important through intermediate reflectance ranges – reflecting the band-difference signal exploited by the CI algorithm – and then diminishes for very high chlorophyll where green reflectance tends to flatten. In contrast, the red band (~670 nm) has virtually no effect on the model's predictions until a threshold is reached at elevated chlorophyll concentrations, after which its influence steeply increases. This behavior aligns with optical expectations, since red reflectance is negligible in low-Chl waters but rises sharply once phytoplankton absorption in the blue saturates and biomass is high. These partial dependence results indicate that the BART model automatically learns the piecewise spectral logic that oceanographers often handle via separate algorithms (blue/green ratios for low-to-moderate Chl, red bands for high Chl). The ability to capture such regime-dependent responses in a single unified model, and to visualize them, underscores the interpretability and scientific insight provided by the BART approach.

1 Data Preprocessing

Data for this study were acquired from multiple satellite ocean color sensors and corresponding in situ chlorophyll-(a) measurements obtained from sources such as the NASA Bio-Optical Marine Algorithm Data set (NOMAD) and the compilation by Valente et al. (2015). To ensure consistency across sensors, the spectral reflectance data ((R_{rs})) were interpolated as needed to common wavelength centers.

For the empirical (OCx) formulation, blue-to-green band ratios were computed for each observation. In particular, the maximum band ratio (MBR) was determined by taking the highest value among the available blue-band ratios (e.g., (Rrs(443)/Rrs(555)), (Rrs(490)/Rrs(555)), and (Rrs(510)/Rrs(555))). This maximum value was then log-transformed:

$$\log R = \log_{10}\left(\frac{R_{rs}(\lambda_{\text{blue}})}{R_{rs}(555)}\right).$$

For the Color Index (CI) formulation of Hu et al. (2012), the CI was calculated as:

$$\text{CI} = R_{rs}(555) - \left[R_{rs}(443) + \frac{555-443}{670-443}\Big(R_{rs}(670) - R_{rs}(443)\Big)\right],$$

and the corresponding in situ chlorophyll-(a) concentrations were log-transformed:

$$\log \mathrm{Chl} = \log_{10}(\mathrm{Chl}).$$

These transformations standardize the data to a common scale, ensuring that variability is appropriately captured for subsequent regression and uncertainty quantification. Detailed descriptions of the interpolation methods and quality control procedures are provided in the Supplementary Material.

2 Statement of Contribution

In this study, we develop and demonstrate a new global chlorophyll retrieval model based on Bayesian Additive Regression Trees implemented in PyMC. We train the BART model on a large, standardized dataset of satellite remote-sensing reflectance (Rrs) spectra matched with in situ chlorophyll measurements, using log - transformed Chl-a as the response to stabilize variance. The resulting model is applied globally to produce chlorophyll-a estimates from multi-spectral satellite data, with associated uncertainty estimates for each prediction. We show that this Bayesian tree-based model can serve as a general-purpose ocean color algorithm that is sensor-agnostic (provided reflectances are harmonized to common wavebands), interpretable, and uncertainty-aware. Unlike conventional empirical algorithms, the BART approach allows users to examine the inferred Rrs–Chl relationships and trust the model's performance across regimes, while also quantifying confidence in each retrieval. This work thus contributes a novel methodological advance to satellite ocean color science: a unified chlorophyll retrieval model that marries the strengths of empirical algorithms (global applicability and simplicity) with the benefits of modern Bayesian machine learning (flexibility, interpretability, and rigorous uncertainty quantification). Our introduction of BART for global chlorophyll prediction opens the door for more robust monitoring of ocean biogeochemistry and improved integration of ocean color data into scientific and management applications.

3 Acknowledgments

4 Open research

References

Hu, C., Lee, Z., & Franz, B. (2012). Chlorophyll aalgorithms for oligotrophic oceans: A novel approach based on three-band reflectance difference. *Journal of Geophysical Research: Oceans*, *117*(C1). https://doi.org/https://doi.org/10.1029/2011JC007395

Hu, C., Feng, L., Lee, Z., Franz, B. A., Bailey, S. W., Werdell, P. J., & Proctor, C. W. (2019). Improving satellite global chlorophyll a data products through algorithm refinement and data recovery. *Journal of Geophysical Research: Oceans*, *124*(3), 1524–1543. https://doi.org/https://doi.org/10.1029/2019JC014941

O'Reilly, John E., & Werdell, P. J. (2019). Chlorophyll algorithms for ocean color sensors - OC4, OC5 & OC6. *Remote Sensing of Environment*, *229*, 32–47. https://doi.org/https://doi.org/10.1016/j.rse.2019.04.021

O'Reilly, John E., Maritorena, S., Mitchell, B. G., Siegel, D. A., Carder, K. L., Garver, S. A., et al. (1998). Ocean color chlorophyll algorithms for SeaWiFS. *Journal of Geophysical Research: Oceans*, *103*(C11), 24937–24953. https://doi.org/https://doi.org/10.1029/98JC02160

O'Reilly, John E., Maritorena, S., Siegel, D. A., O'Brien, M. C., Toole, D., Mitchell, B. G., et al. (2000). Ocean color chlorophyll a algorithms for SeaWiFS, OC2, and OC4: Version 4. *SeaWiFS Postlaunch Calibration and Validation Analyses, Part*, *3*, 9–23.