WATER QUALITY ESTIMATION AND INTERPRETATION USING HYPERSPECTRAL DIMENSIONALITY REDUCTION

Erlend Tveiten Øien, Aria Alinejad, Joseph L. Garrett

Department of Engineering Cybernetics
Norwegian University of Science and Technology, 7034 Trondheim, Norway

ABSTRACT

Hyperspectral remote sensing enables estimation of water quality parameters (WQPs) across aquatic ecosystems, but its high dimensionality motivates dimensionality reduction (DR) as pre-processing. DR is also useful for representing the data in a way that is favorable for interpretation. Existing work on DR for WQP estimation is dominated by PCA, whose components are not constructed for interpretability, while NMF and AA remain mostly underexplored. In this paper we compare these three methods on the hyperspectral GLORIA dataset. We investigate how interpretable their components are and how comparable they are to common WQPs (Chl-a, CDOM, TSS, and SD), and how well each performs as pre-processing prior to WQP estimation, using linear regression and KNN. The results show that AA gives the components most comparable to the WQPs of interest, while NMF gives the best estimation, with SD the most accurately predicted ($R^2 = 0.73$) and KNN outperforming linear regression.

***Index Terms*—** Archetypal analysis, Dimensionality reduction, Hyperspectral remote sensing, Interpretable machine learning, Spectral unmixing, Water quality parameters

1. INTRODUCTION

Light drives photosynthesis in aquatic ecosystems and is indicative of their state, which makes water-color remote sensing (RS) a powerful tool for the systematic, large-scale observation and estimation of water quality parameters (WQPs) across oceans, coasts, and inland waters, a role it has filled for 60 years [1, 2]. The traditional alternative, in-situ sampling, is spatially and temporally sparse, costly, and labor intensive [3]. Historically, RS has relied on multispectral sensors that record the response at a fixed set of bands, but there is now a growing availability of hyperspectral sensors that capture a near-continuous spectrum, with satellites such as EnMAP, PRISMA, HYPSO, and PACE [4, 5, 6, 7]. This enables more advanced analyses, such as spectral unmixing [8], and allows the data to be used adaptively for a wide range of applications, including ones not foreseen at the time of instrument design. Moreover, hyperspectral data has been specifically called for to advance applications such as phytoplankton diversity estimation [9] and harmful algal bloom detection [10].

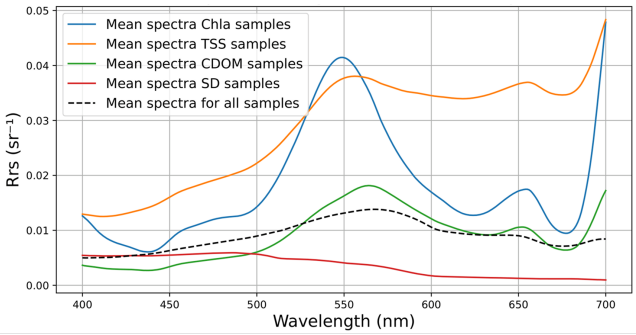


Fig. 1. Mean of top-10 spectra for each WQP in GLORIA.

The benefits of hyperspectral data come at a cost. Its high dimensionality inflates computational, storage, and transmission demands, and because a very large number of samples is needed to populate such a high-dimensional feature space, it leaves datasets sparse and susceptible to the curse of dimensionality [11], resulting in poor generalization, an effect amplified in the optically complex, turbid waters typical of coasts and inland lakes. Dimensionality reduction (DR) mitigates these issues while retaining information from across the full spectrum. Importantly, DR serves a dual purpose: beyond compressing the data for downstream estimation, different DR objectives yield components with different structure, some of which can be interpreted physically and related to the constituents that color the water.

Using DR-preprocessed hyperspectral data for target detection and classification in ocean RS is well studied and has shown good results [12, 13]. DR for WQP estimation has received far less attention, and the work that exists is dominated by principal component analysis (PCA) and partial least squares [14, 15, 16, 17, 18, 19], which use it purely as pre-processing, without interpreting the resulting components against the water constituents that shaped them. Non-negative matrix factorization (NMF) and archetypal analysis (AA), whose non-negativity and convexity constraints tend to produce more physically meaningful parts, remain largely unexplored for WQP work, although they are commonly used for terrestrial hyperspectral analysis.

We study four widely measured WQPs: the optically active constituents (OACs) chlorophyll-a (Chl-a), colored dis-

Thanks to the Research Council of Norway and industry partners for funding through the projects SoS (Grant No. 337495) and HYPSCI (Grant No. 325961).

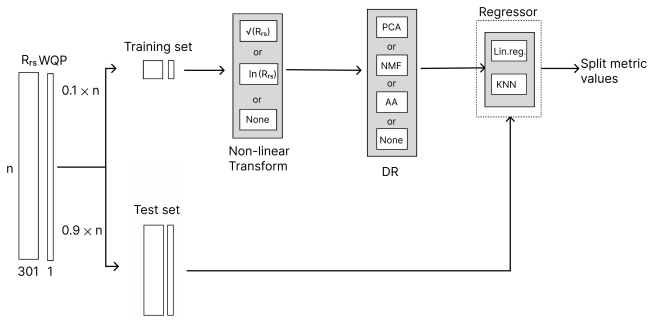


Fig. 2. Processing pipeline for WQP estimation.

solved organic matter (CDOM), and total suspended solids (TSS), together with Secchi depth (SD), an apparent optical property (AOP) that indicates water clarity. First, we interpret the components produced by PCA, NMF, and AA on the GLORIA dataset [20] and assess how closely they correspond to Chl-a, CDOM, TSS, and SD. Second, we benchmark the three methods as pre-processing for WQP estimation using linear regression and k-nearest-neighbor (KNN) regression.

2. MATERIALS AND METHODS

2.1. The GLORIA dataset

The hyperspectral GLObal Reflectance community dataset for Imaging and optical sensing of Aquatic environments (GLORIA) [20] open-access dataset is used for this study. The dataset contains 7500 curated hyperspectral remote sensing reflectance (Rrs) spectra between 250-900 nm at 1 nm spectral resolution, from coastal and inland waters with co-located in-situ WQP measurements. In this analysis, we only analyze samples between 400-700 nm, which is the region where the WQPs of interest are primarily active, and is where all of the sensors in the dataset have samples. Figure 1 shows the characteristic spectrum of each WQP, taken as the mean of the 10 samples with the highest measured value of that WQP.

2.2. Dimensionality Reduction Methods

We compare three unsupervised DR methods: PCA, NMF, and AA. Let X be the $n \times p$ data matrix with n reflectance spectra over p bands. Each method approximates X by a low-rank factorization

$$X \approx AS, \tag{1}$$

where the k rows of the $k \times p$ matrix S are the components and the rows of the $n \times k$ matrix A the corresponding coefficients, representing each spectrum by $k < p$ values. We use *components* for what is otherwise called principal axes or loadings (PCA), basis spectra (NMF), and archetypes (AA), and *coefficients* for the rows of A (often called scores in the context of PCA). The components are plotted and interpreted, while the coefficients are correlated with the WQPs and used as regressor inputs; e.g. "AA3 correlates with Chl-a" refers to the coefficient of AA3. The three methods share this form, but differ in the constraints they place on A and S.

PCA finds the components that explain the most variance in the data, with the first component explaining the most and each successive component explaining less. It places no sign constraint on A or S, so PCA components can take both positive and negative values. The components can be obtained from the singular value decomposition of the mean-centered data,

$$X = U\Sigma V^T, \tag{2}$$

by taking $S = V_k^T$, where V_k^T contains the first k columns of the $p \times k$ matrix V^T, whose rows are orthonormal and ordered by decreasing variance.

NMF constrains both factors to be non-negative, so that each spectrum is reconstructed by a purely additive combination of non-negative components. The factors are found by solving

$$\min_{A,S} \|X - AS\|_F \text{ s.t. } A \geq 0 \text{ and } S \geq 0. \tag{3}$$

AA constrains the components to be convex combinations of the samples themselves, $S = BX$, and each spectrum to be a convex combination of those components,

$$\min_{A,B} \|X - ABX\|_F \text{ s.t. } A, B \geq 0,\ A\mathbb{1} = \mathbb{1},\ B\mathbb{1} = \mathbb{1}, \tag{4}$$

where $A\mathbb{1} = \mathbb{1}$ and $B\mathbb{1} = \mathbb{1}$ are constraints ensuring that the rows of A and B sum to 1. Named archetypes in the original paper [21], these components represent the extreme, or pure, types of the dataset, which when mixed can approximate all remaining samples. In practice, B is sparse, and each component is typically formed from fewer than five samples. This makes AA a natural formulation for spectral unmixing [8], where each spectral sample is assumed to be a convex combination of relatively few phenomena that the components are expected to capture.

Going from PCA to NMF to AA, the constraints on the factorization tighten, which is expected to yield more interpretable components at the cost of reconstruction accuracy. For PCA and NMF scikit-learn [22] implementations were used, while for AA a scikit-learn compatible implementation was used [23].

2.3. Training and Evaluation

To assess interpretability, all DR methods are trained on 20% of the dataset, and 5 components are extracted from each DR method. The shape and overall size of the components are evaluated visually and compared to the WQPs of interest, and Pearson's correlation coefficient is used to quantify the relationship between the component coefficients and the WQPs.

For water quality estimation, 10% of the data is used for training, and the remaining data is used as a test set. Both

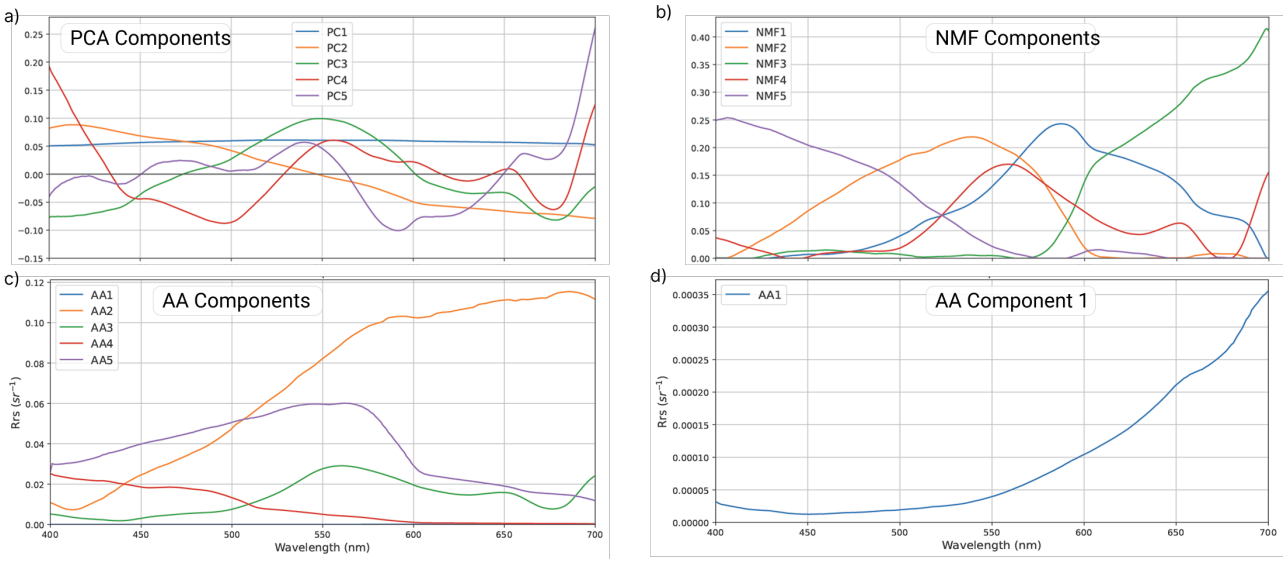


Fig. 3. a)–c) PCA, NMF, and AA components plotted against wavelength. d) shows a magnified plot of AA1.

KNN and linear regression are tested with two different non-linear transforms applied to the spectra as pre-processing, namely the log-transform and the square-root-transform.

Evaluation is performed using the R^2-score and the root mean square error (RMSE). The R^2-score measures the amount of variance explained by the model compared to the variance in the data, with a score of 1 indicating perfect prediction and a score of 0 indicating that the model does not explain any variance. The RMSE is a measure of the average magnitude of the errors in the predictions, with lower values indicating better performance. Figure 2 shows the regression framework for each train-test split.

The processing pipeline has two hyperparameters. A grid search, optimized for R^2, was tested for each DR-regressor pair to find the optimal number of DR components, with 100 being the maximum number of components tested for any pair. For Chl-a, CDOM, and TSS the maximum number of components was set to 50, with the AA Chl-a pairs in particular set to 40. For KNN-pipelines, a grid search of 4 to 12 neighbors was tested across all of the component values. Scikit-learn was used for KNN and linear regression implementations, and also for building the testing regime and pipeline framework.

3. RESULTS

3.1. Component interpretation

Figure 3 shows the PCA, NMF, and AA components plotted against wavelength, and Table 1 shows the Pearson correlation between the component coefficients and the WQPs.

PC1 captures the overall brightness of a sample and PC2 how blue it is; under the log-transform PC2 changes sign and instead captures how red a sample is. Consistent with this, Table 1 shows PC1 and PC2 correlating negatively with SD and PC2 positively with CDOM. PC3 to PC5 have more complex shapes, are harder to interpret, and correlate weakly with the WQPs. Because the components can take both positive and negative values, unlike real reflectances, they are less physically interpretable than those of the other methods.

		Component				
		1	2	3	4	5
PCA	Chl-a	0.06	0.04	-0.03	0.15	**0.21**
	CDOM	-0.20	**0.41**	0.19	-0.12	0.19
	SD	-0.48	**-0.60**	-0.15	-0.11	-0.08
	TSS	0.16	0.05	-0.02	0.08	**0.21**
NMF	Chl-a	-0.01	-0.04	0.09	**0.51**	-0.02
	CDOM	0.10	-0.09	**0.18**	**0.18**	-0.11
	SD	**-0.36**	-0.25	-0.35	-0.34	-0.24
	TSS	0.27	0.19	**0.40**	0.07	0.17
AA	Chl-a	-0.24	0.03	**0.43**	0.00	-0.08
	CDOM	-0.03	**0.18**	0.13	-0.09	-0.17
	SD	**0.43**	-0.32	-0.35	-0.13	-0.13
	TSS	-0.22	**0.38**	-0.02	0.25	-0.02

Table 1. Pearson correlation between the per-sample component coefficients of each DR method and the WQPs. PCA is computed on log-transformed Rrs; NMF and AA on Rrs.

NMF3 is a red-reflection component resembling high-TSS samples; the 40 samples with the highest NMF3 coefficients have a mean TSS of 434.9 $\mathrm{g\,m^{-3}}$, far above the dataset mean, consistent with its 0.40 correlation with TSS in Table 1. NMF4 resembles the spectral shape of Chl-a and has the strongest correlation with it (0.51), although, unlike real

water rich in Chl-a, it dips to zero reflectance around 670 nm and in parts of the blue. NMF5 is a blue-reflection component whose high-valued samples tend to have low CDOM and Chl-a; its weak correlation with these WQPs likely reflects the complex interactions between CDOM and Chl-a in this part of the spectrum. Because NMF reconstructs each spectrum as a positive linear combination of its components, and coefficients can only add and never subtract, brighter samples receive larger coefficients across all components. A high coefficient value might therefore simply indicate overall reflectance magnitude rather than an elevated concentration of a specific constituent, as seen in the uniformly negative correlations with SD. Lastly, NMF1 and NMF2 have complex shapes and are hard to interpret.

Each AA component is a convex combination of only 1 to 4 samples, so the components keep the Rrs unit and can be analyzed directly with existing water-spectra tools. AA1 is a dim attenuation component, built from the darkest sample, that dominates the reconstruction of most samples; AA2 has high red reflectance characteristic of high TSS; AA3 has the spectral shape of water dominated by Chl-a, its samples having a mean Chl-a nearly twice the dataset mean; and AA4 and AA5 both resemble low-constituent water, close to pure and oligotrophic water respectively. These readings are supported by Table 1, where AA3 correlates most strongly with Chl-a (0.43), AA2 with TSS (0.38), and AA1 with SD (0.43). One should note that the overlapping spectral impact of the OACs makes the inversion of WQPs from Rrs non-unique, so a high coefficient value does not necessarily imply the presence of the associated constituent. Because the components are extreme samples, the AA coefficient matrix is sparse (24% zeros versus 3.5% for NMF), so each sample is represented by only a few components, making AA the most interpretable of the three methods.

3.2. Estimation results

Figure 4 shows the R^2-scores for each WQP using the best-performing hyperparameters for each pipeline, and Table 2 reports the corresponding best R^2 and RMSE values for every combination of WQP, DR method, and regressor.

NMF achieves the best performance across all WQPs, with AA a close second and PCA the weakest. Performance depends strongly on the regressor, with the non-linear KNN consistently outperforming linear regression. We find no clear relationship between a method's reconstruction error, measured by the normalized Frobenius error, and its regression performance: PCA has the lowest reconstruction error yet is outperformed by AA. Varying the number of components further shows that AA performs poorly with few components, reaches the other methods at around 10 to 15 components, and still appears to be underfitting at the maximum number tested.

SD was the most accurately predicted WQP, with an R^2 of 0.73, while Chl-a, CDOM, and TSS were predicted relatively poorly. This likely reflects the large spectral diversity of water types across regions on a global scale. An optical water type (OWT) framework that selects or blends pipelines per water type could better capture the dynamics of each WQP.

		Linear Regression			KNN		
		PCA	NMF	AA	PCA	NMF	AA
R^2	Chl-a	**0.46**	0.41	0.43	0.34	**0.46**	0.45
	CDOM	0.31	0.25	0.31	0.32	**0.39**	0.38
	TSS	0.34	0.33	0.32	0.33	**0.39**	0.36
	SD	0.66	0.49	0.70	0.70	**0.73**	0.71
RMSE	Chl-a	42.54	44.07	43.56	46.95	**42.52**	43.56
	CDOM	0.98	1.02	0.98	0.97	**0.92**	0.93
	TSS	24.56	24.70	24.88	24.64	**23.63**	24.22
	SD	1.10	1.35	1.03	1.02	**0.97**	1.01

Table 2. Best R^2 and RMSE of each DR-method and regressor pair for the four WQPs.

4. CONCLUSION

We compared PCA, NMF, and AA on the GLORIA dataset, both for interpretation and as pre-processing for WQP estimation. AA produced the most interpretable components and PCA the least. Since each AA component is a convex combination of a small set of measured spectra, the components retain the Rrs unit and can be analyzed directly with existing water-spectra tools and optical theory. Several AA components matched the spectral signatures of individual WQPs, and a sample's similarity to such a component was often a good indication of that WQP, although this interpretation weakened for spectra that multiple, spectrally overlapping WQPs could explain.

As pre-processing for estimation, NMF performed best and AA nearly as well. The non-linear KNN regressor consistently outperformed linear regression, and SD was the most accurately predicted WQP. The lack of relationship between reconstruction error and estimation accuracy motivates optimization constraints chosen for task relevance rather than reconstruction fidelity, such as AA's convexity constraint. Since the combination of pre-processing transform, DR method, and regressor strongly affected the results, an OWT framework that selects or blends pipelines per water type is a promising direction for future work.

5. REFERENCES

[1] G. E. Adjovu *et al.*, "Overview of the Application of Remote Sensing in Effective Monitoring of Water Quality Parameters," *Remote Sensing*," 2023.

[2] S. N. Topp *et al.*, "Research Trends in the Use of Remote Sensing for Inland Water Quality Science: Moving Towards Multidisciplinary Applications," *Water*," 2020.

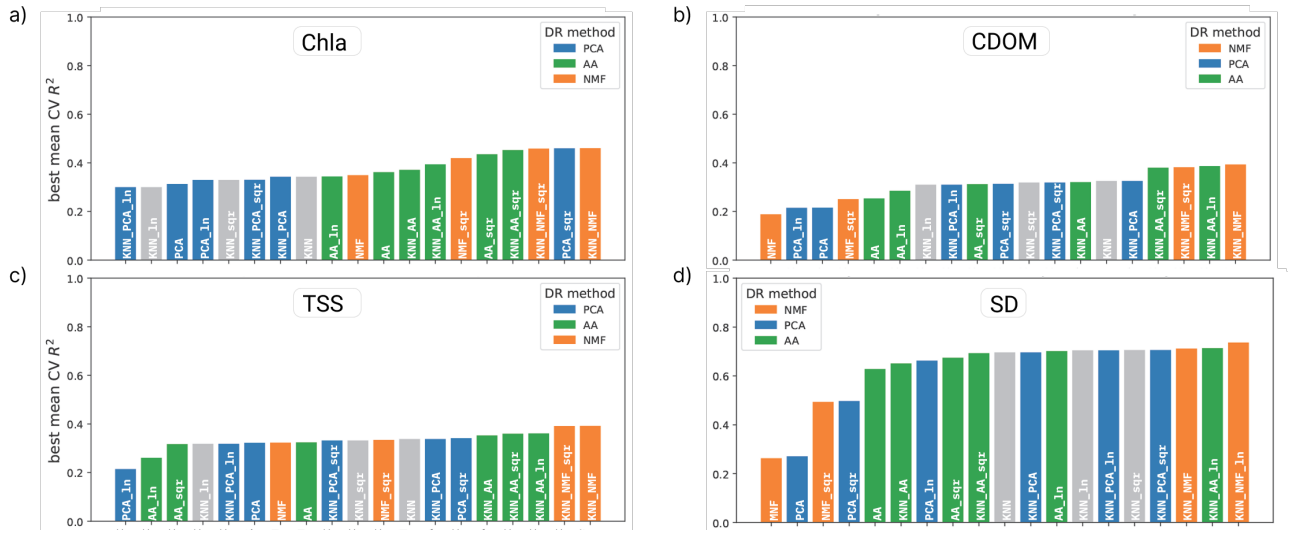


Fig. 4. Best-performing model for each DR-method and regressor pipeline, for a)–d) Chl-a, CDOM, TSS, and SD. In the model names, "ln" and "sqr" denote log- and square-root-transformed reflectance. "KNN" denotes KNN regression, and all other models use linear regression.

[3] H. Yang *et al.*, "A Review of Remote Sensing for Water Quality Retrieval: Progress and Challenges," *Remote Sensing*," 2022.

[4] L. Guanter *et al.*, "The EnMAP Spaceborne Imaging Spectroscopy Mission for Earth Observation," *Remote Sensing*," 2015.

[5] S. Cogliati *et al.*, "The PRISMA imaging spectroscopy mission: overview and first performance analysis," *Remote Sensing of Environment*," 2021.

[6] S. Bakken *et al.*, "HYPSO-1 CubeSat: First Images and In-Orbit Characterization," *Remote Sensing*," 2023.

[7] E. T. Gorman *et al.*, "The NASA Plankton, Aerosol, Cloud, ocean Ecosystem (PACE) mission: an emerging era of global, hyperspectral Earth system remote sensing," in *Sensors, Systems, and Next-Generation Satellites XXIII*, vol. 11151. SPIE, 2019.

[8] A. Zouaoui *et al.*, "Entropic Descent Archetypal Analysis for Blind Hyperspectral Unmixing," *IEEE Transactions on Image Processing*," 2023.

[9] A. Bracher *et al.*, "Obtaining Phytoplankton Diversity from Ocean Color: A Scientific Roadmap for Future Development," *Frontiers in Marine Science*," 2017.

[10] R. M. Khan *et al.*, "A Meta-Analysis on Harmful Algal Bloom (HAB) Detection and Monitoring: A Remote Sensing Perspective," *Remote Sensing*," 2021.

[11] L. van der Maaten *et al.*, "Dimensionality Reduction: A Comparative Review," *Journal of Machine Learning Research - JMLR*," 2007.

[12] D. Lupu *et al.*, "Quick Unsupervised Hyperspectral Dimensionality Reduction for Earth Observation: A Comparison," *IEEE Transactions on Computational Imaging*," 2025.

[13] S. Bakken *et al.*, "The effect of dimensionality reduction on signature-based target detection for hyperspectral remote sensing," in *CubeSats and SmallSats for Remote Sensing III*, vol. 11131. SPIE, 2019.

[14] E. Batur and D. Maktav, "Assessment of Surface Water Quality by Using Satellite Images Fusion Based on PCA Method in the Lake Gala, Turkey," *IEEE Transactions on Geoscience and Remote Sensing*," 2019.

[15] K. A. Ali and J. D. Ortiz, "Multivariate approach for chlorophyll-a and suspended matter retrievals in Case II type waters using hyperspectral data," *Hydrological Sciences Journal*," 2016.

[16] Y. Hou *et al.*, "A study on water quality parameters estimation for urban rivers based on ground hyperspectral remote sensing technology," *Environmental Science and Pollution Research*," 2022.

[17] G. Im *et al.*, "Estimating Chlorophyll-a Concentration from Hyperspectral Data Using Various Machine Learning Techniques: A Case Study at Paldang Dam, South Korea," *Water*," 2022.

[18] D. Kim *et al.*, "Real-time chlorophyll-*a* forecasting using machine learning framework with dimension reduction and hyperspectral data," *Environmental Research*," 2024.

[19] K. Ryan and K. Ali, "Application of a partial least-squares regression model to retrieve chlorophyll-a concentrations in coastal waters using hyper-spectral data," *Ocean Science Journal*," 2016.

[20] M. K. Lehmann *et al.*, "GLORIA - A globally representative hyperspectral in situ dataset for optical sensing of water quality," *Scientific Data*," 2023.

[21] A. Cutler and L. Breiman, "Archetypal Analysis," *Technometrics*," 1994.

[22] "Scikit-learn: Machine learning in Python — scikit-learn 1.9.0 documentation," https://scikit-learn.org/stable/.

[23] A. Alcacer, "Scikit-learn compatible package for archetypal analysis." https://github.com/aleixalcacer/archetypes.