IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 3, June 2026, pp. 2081~2091
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i3.pp2081-2091

Advanced inferential statistics and data mining for chlorophyll distribution clustering

Felix Reba[1], Toha Saifudin[2,3], Rimuljo Hendradi[4,5]
[1]Doctoral Program of Mathematics and Natural Sciences, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia
[2]Department of Mathematics, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia
[3]Research Group of Statistical Modeling in Life Sciences, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia
[4]Information Systems Study Program, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia
[5]Research Group of Center for Business Intelligence (CenBI), Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia

Article Info

Article history:

Received Nov 8, 2024
Revised Feb 13, 2026
Accepted Apr 20, 2026

Keywords:

Chlorophyll
Clustering
K-means
Marine environment
Probability distribution
Silhouette index

ABSTRACT

This study proposes an integrated statistical framework to analyze chlorophyll distribution in marine environments by combining probability distribution modeling, goodness-of-fit (GoF) evaluation, and machine learning-based clustering. Eight probability distribution models—half-normal, inverse Gaussian, Rician, Birnbaum–Saunders, Nakagami, extreme value, t location-scale, and stable—were evaluated using observational chlorophyll-a data from the Copernicus Marine Service. Model performance was assessed through the Kolmogorov–Smirnov (KS) and Anderson–Darling (AD) GoF tests, along with five statistical information criteria. The results indicate that the inverse Gaussian and extreme value distributions consistently offered the best statistical fit and ecological relevance across varying sample sizes. Clustering analysis, performed using the k-means algorithm and validated via the silhouette index, further confirmed the robustness of these two models in forming stable and well-separated clusters. In contrast, the half-normal distribution showed poor performance and instability, especially with smaller sample sizes. The proposed taxonomy and spatial visualizations enable empirical classification of model behavior and support integration into real-time marine decision support systems (DSS) for ecosystem monitoring. Overall, the study contributes to the development of accurate, data-driven analytical tools that aid sustainable marine resource management, aligned with sustainable development goal (SDG) 14 on marine ecosystem protection.

Corresponding Author:

Toha Saifudin
Department of Mathematics, Faculty of Science and Technology, Universitas Airlangga
Surabaya, Indonesia
Email: tohasaifudin@fst.unair.ac.id

1. INTRODUCTION

Modeling probability distributions in environmental studies, particularly for chlorophyll data, plays a critical role in understanding the variability and dynamics of natural oceanographic phenomena such as productivity, seasonal blooms, and carbon cycling processes [1], [2]. Chlorophyll concentration is a key indicator of marine ecosystem health, especially in monitoring phytoplankton dynamics, which are central to

the biological pump and oceanic carbon flux [3]. This highlights the importance of accurately characterizing chlorophyll variability to support reliable environmental analysis and decision-making.

Although various environmental variables, such as wind and rainfall, have been successfully modeled using classical probability distributions, in-depth studies of the complex and heterogeneous spatial-temporal patterns of chlorophyll distribution remain limited [4], [5]. Many previous approaches have relied on simulated datasets, which often fail to reflect the real spatiotemporal variability of marine chlorophyll [6]. This limitation is particularly problematic in ecological contexts such as harmful algal blooms (HABs), which are highly sensitive to small fluctuations in chlorophyll concentration [7].

To overcome these limitations, observational satellite datasets, such as those from the Copernicus Marine Service, offer high-resolution and more representative alternatives for capturing dynamic marine conditions [8], [9]. These real-world datasets enhance ecological assessment by accurately reflecting spatial and temporal variations in chlorophyll concentration. This study seeks to fill the research gap by evaluating the performance of eight probability distribution models—half-normal, inverse Gaussian, Rician, Birnbaum–Saunders, Nakagami, extreme value, t location-scale, and stable—in representing the variability of chlorophyll-a data [4], [5]. Model evaluation is conducted using two well-established goodness-of-fit (GoF) tests: the Kolmogorov–Smirnov (KS) and Anderson–Darling (AD) tests, which are widely used in environmental statistics and tail modeling [10]–[12].

In addition to the distributional evaluation, this study introduces a novel machine learning–based clustering approach, using the k-means algorithm to group distribution models based on GoF performance across multiple sample sizes. Cluster quality is validated through the silhouette index, which measures the cohesion and separation of resulting clusters [13]. This framework enables the discovery of behavioral patterns among models that may not be apparent through individual evaluation alone. Recent advances in environmental monitoring, including the internet of things (IoT) and intelligent ocean sensors, further enhance the capability to detect hidden patterns in oceanographic datasets [14]. For instance, Zhang *et al.* [15] proposed a cross-domain adaptive deep learning method for improving object detection in coastal remote sensing data, while Ling *et al.* [16] emphasized the relevance of spectral features for monitoring vegetation and chlorophyll dynamics. Such innovations support the growing integration of artificial intelligence (AI) and statistical modeling in marine science.

Unlike prior studies that have primarily focused on satellite-based chlorophyll estimation (e.g., moderate resolution imaging spectroradiometer (MODIS) and sea-viewing wide field-of-view sensor (SeaWiFS)), this study proposes a structure-based statistical framework that combines probability modeling, GoF evaluation, and clustering—advancing beyond estimation toward analysis of statistical distribution behavior and inter-model relationships [17]. The integrated framework offers several methodological contributions. First, it enables comparative analysis of distribution models based on both their individual performance and shared behavioral patterns through clustering—an approach rarely applied in chlorophyll modeling. Second, using the silhouette index improves interpretability by empirically grouping models with similar statistical profiles. This transcends traditional GoF-based ranking by revealing deeper relational patterns among the candidate models [13].

Lastly, chlorophyll's broader ecological significance—as a proxy for primary productivity, nutrient cycling, and climate resilience—underscores the importance of this study. Better understanding of chlorophyll variability supports marine ecological forecasting and contributes to sustainable marine resource management and climate adaptation, in line with sustainable development goal (SDG) 14 on marine ecosystem protection [3]. This further emphasizes the need for robust and reliable analytical approaches to capture its complex variability.

2. METHOD

2.1. Probability distribution models

In this study, eight probability distribution models were selected for their potential to capture the characteristic variations in chlorophyll-a data within marine environments. The selection was based on a comprehensive literature review and empirical evidence supporting their suitability for environmental datasets characterized by high spatial and temporal variability. The selected distributions include: half-normal, inverse Gaussian, Rician, Birnbaum–Saunders, Nakagami, extreme value, t location-scale, and stable [18]–[22].

Each distribution provides specific strengths in modeling distinct features of chlorophyll-a concentrations. For example, Rician and Birnbaum–Saunders distributions are frequently used in contexts involving oceanic dynamics and biological variability. The extreme value distribution is well-suited to capture peak events such as algal blooms. The half-normal distribution is appropriate for modeling strictly positive data, whereas the inverse Gaussian is effective for highly skewed datasets exhibiting significant

variability. Nakagami and stable distributions accommodate asymmetry and heavy-tailed behavior, while the t location-scale distribution is known for its robustness to outliers.

To assess the fitness of these models in representing chlorophyll-a variability, two GoF tests were applied: the KS and AD tests. In addition, two model selection criteria—the Akaike information criterion (AIC) and the Bayesian information criterion (BIC)—were employed to guide the identification of the most appropriate models for characterizing the complex distributional behavior of chlorophyll-a concentrations in marine ecosystems. Together, these evaluation metrics provide a comprehensive framework for comparing model performance and ensuring robust statistical inference.

2.2. Test statistics

To evaluate the fit of probability distribution models to the chlorophyll-a dataset, two GoF tests were applied: the KS test and the AD test [23]–[27]. The KS test measures the maximum distance between the empirical cumulative distribution function $F_n(x)$ and the theoretical distribution $F_0(x)$, as expressed in (1).

$$D = \max_{x \in \mathbb{R}} |F_n(x) - F_0(x)| \quad (1)$$

A smaller D_{count} value indicates a closer alignment between empirical data and the theoretical model, suggesting a better fit [24], [26]. The AD test complements the KS test by placing greater weight on differences in the tails of the distribution—important for detecting rare but ecologically significant events. The AD statistic is calculated using (2).

$$A_n^2 = -n - \sum_{i=1}^{n} \left[\frac{(2i-1)}{n} * (\log F(x_i) + \log(1 - F(x_{n+1-i}))) \right] \quad (2)$$

Lower A_n^2 values indicate a better fit, particularly in capturing extreme distributional behavior, which is vital in environmental and hydrological contexts [25], [27]. In addition to the GoF tests, this study employs five statistical information criteria to assess model fit while penalizing complexity: hannan–quinn criterion (HQC), consistent Akaike information criterion (CAIC), BIC, corrected Akaike information criterion (AICc), AIC. These criteria ensure a balanced evaluation by integrating both fit and parsimony, supporting the identification of the most appropriate probability distribution model to characterize chlorophyll-a behavior in marine ecosystems [23], [27].

2.3. Model selection based on information-theoretic criteria

In statistical model selection, information-theoretic criteria are essential tools for evaluating and comparing candidate models by balancing GoF with model complexity. The most commonly employed criteria include the HQC, CAIC, BIC, AICc, and the AIC [28]. Each criterion imposes a penalty on model complexity to mitigate overfitting, though the severity of these penalties varies. AIC applies a moderate penalty and is generally suitable for smaller sample sizes, while BIC imposes a stronger penalty that scales with sample size, making it more appropriate for larger datasets. AICc adjusts AIC for small-sample bias, improving its reliability in limited-data contexts. CAIC and HQC, meanwhile, act as more conservative metrics, enforcing stricter penalties on overly complex models.

In this study, all five criteria were applied to assess the fit of each probability distribution model across four sample sizes (n =20, 25, 30, 35). Instead of evaluating performance independently at each sample size, the mean value of each criterion was computed across all sample sizes. This averaging method provides a more stable and equitable foundation for comparison, minimizing the impact of variability due to sample size. By identifying models that consistently strike an optimal balance between fit and simplicity, this approach improves generalizability and reduces the risk of overfitting in chlorophyll-a distribution modeling [28].

2.4. Clustering method and evaluation

Unlike traditional studies that cluster chlorophyll data spatially or based solely on satellite-derived observations, this study adopts a novel approach by applying k-means clustering directly to GoF metrics—specifically, the KS and AD statistics. This method enables clustering based on distributional behavior rather than on raw chlorophyll-a values, offering a more robust and informative foundation for statistical pattern recognition across varying sample sizes [29], [30]. This approach provides deeper insight into the comparative performance and relationships among candidate distribution models.

To implement this approach, the k-means algorithm was employed to group the distribution models according to the similarity of their GoF values. The algorithm calculates Euclidean distance between each model's GoF vector and the cluster centroid, and iteratively reassigns models to clusters in order to minimize the total intra-cluster variance, continuing this process until convergence. In (3) quantifies the distance

between a distribution model—based on its GoF values—and the cluster centroid in a multidimensional GoF feature space [29].

$$d_{ik} = \sum_{j=1}^{m} (x_{kj} - c_{ij})^2 \quad (3)$$

To evaluate clustering quality, this study employed the silhouette index, which measures both cohesion (how close a distribution is to others within its own cluster) and separation (how distinct it is from distributions in other clusters). The index is computed as (4).

$$\bar{S} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{b(i) - a(i)}{\max(a(i), b(i))} \right) \quad (4)$$

Where $a(i)$ is the average intra-cluster distance for data point i, and $b(i)$ is the lowest average distance to points in other clusters. Values close to 1 indicate well-separated clusters, while negative values suggest potential misclassification [31], [32].

This integrated method allows for the identification of robust, unstable, or outlier-prone distribution models across varying sample sizes. As such, clustering in this study is no longer limited to spatial analysis but is instead driven by GoF-based behavioral patterns, enabling empirically grounded groupings of statistical models [29], [30], [33]. To enhance transparency and reproducibility, Figure 1 illustrates the overall workflow of the methodology—starting from data retrieval and distribution fitting, followed by GoF testing, clustering, and final model evaluation. This diagram facilitates a clearer understanding of the analytical process and improves interpretability for readers.

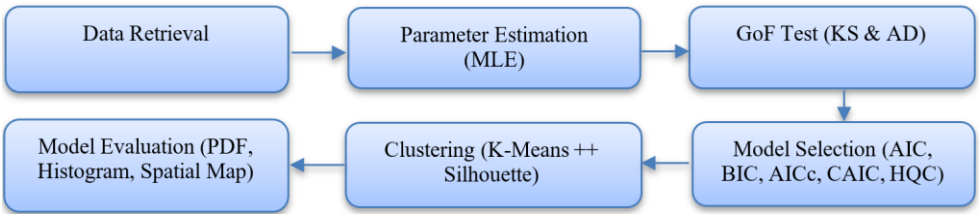

Figure 1. Simplified workflow of the methodological steps

3. RESULTS AND DISCUSSION

3.1. Research data

This study utilizes chlorophyll-an observations obtained from the Copernicus Marine Data Store at https://marine.copernicus.eu/, downloaded in spreadsheet format. The dataset represents a vertical snapshot, meaning chlorophyll-a values were collected at a single timestamp—14 May 2024, 11:46 UTC—from a fixed geographic location in the northern black sea (approximately 31.35727°E, 41.69535°N), covering a vertical depth range from 0 to 19.2 meters. This vertical sampling design intentionally excludes temporal variability, focusing solely on the probabilistic behavior of chlorophyll concentrations under static environmental conditions. Sample sizes (n =20, 25, 30, 35) were derived by drawing subsets from the vertical profile and are justified based on prior remote sensing studies concerning chlorophyll resolution optimization and spatial representation under various observational constraints [7]–[9], [14], [34]. While depth was recorded, it was not treated as an analytical variable. Instead, it served as a structural reference for subsample construction to ensure ecological representativeness. The primary analytical focus remains on how the chlorophyll-a values conform to various univariate probability distribution models.

To reduce bias, only valid observations that satisfied the underlying assumptions of the tested distributions were included. Potential sources of uncertainty—such as atmospheric correction artifacts, cloud masking, or instrument calibration issues—were acknowledged and mitigated through data screening and controlled sampling procedures. Although limited to a single temporal snapshot, this approach provides a robust foundation for probabilistic modeling of chlorophyll-a concentrations in ecologically dynamic coastal waters.

3.2. Test statistic values

In this study, the KS and AD tests were applied to evaluate the GoF of eight distribution models to chlorophyll-a data across various sample sizes (n =15–50). Tables 1 and 2 report the resulting p-values, with

hypothesis decisions (H_0 accepted/rejected) based on a significance threshold α =0.05. These results provide a systematic basis for comparing model performance across different sample conditions.

Table 1. p-values from KS and AD tests for various distributions (n =20, 25, 30, 35)

Distribution	Hypotheses	n =20		n =25		n =30		n =35	
		$p-$KS	$p-$AD	$p-$KS	$p-$AD	$p-$KS	$p-$AD	$p-$KS	$p-$AD
Rician	H_0 accepted	0.378	0.671	0.205	0.317	0.14	0.317	0.486	0.623
Birnbaum–Saunders	H_0 accepted	0.383	0.63	0.277	0.407	0.368	0.588	0.778	0.741
Extreme value	H_0 accepted	0.409	0.788	0.048	0.102	0.088	0.164	0.382	0.417
Half-normal	H_1 (H_0 rejected)	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000
Inverse Gaussian	H_0 accepted	0.383	0.63	0.277	0.407	0.376	0.594	0.779	0.741
Nakagami	H_0 accepted	0.379	0.658	0.226	0.345	0.173	0.386	0.572	0.678
Stable	H_0 accepted	0.379	0.672	0.878	0.877	0.721	0.73	0.477	0.616
T location-scale	H_0 accepted	0.378	0.671	0.818	0.673	0.123	0.289	0.478	0.616

Table 2. Extended p-values from KS and AD tests for sensitivity analysis (n =15, 40, 45, 50)

Distribution	Hypotheses	n =15		n =40		n =45		n =50	
		$p-$KS	$p-$AD	$p-$KS	$p-$AD	$p-$KS	$p-$AD	$p-$KS	$p-$AD
Rician	H_0 accepted	0.745	0.879	0.361	0.226	0.684	0.419	0.372	0.4
Birnbaum–Saunders	H_0 accepted	0.831	0.949	0.061	0.076	0.857	0.743	0.185	0.109
Extreme value	H_0 accepted	0.696	0.609	0.43	0.517	0.107	0.063	0.047	0.101
Half-normal	H_1 (H_0 rejected)	0.001	0.01	0.000	0.000	0.000	0.000	0.000	0.000
Inverse Gaussian	H_0 accepted	0.831	0.949	0.059	0.074	0.857	0.745	0.184	0.105
Nakagami	H_0 accepted	0.771	0.906	0.209	0.163	0.762	0.527	0.63	0.557
Stable	H_0 accepted	0.995	0.999	0.38	0.233	0.769	0.828	0.32	0.349
T location-scale	H_0 accepted	0.751	0.882	0.377	0.233	0.916	0.589	0.319	0.351

The statistical outcomes are visually summarized in Figure 2, which presents a Grouped Visual Taxonomy based on the KS and AD evaluations. This figure retains both the KS statistic values (as horizontal bar lengths) and AD statistic values (numerically displayed on each bar), providing a comprehensive and intuitive quantitative comparison across models. The taxonomy organizes all eight distribution models across four sample sizes (n =20, 25, 30, 35) within a 2×2 matrix layout, facilitating clear visual interpretation.

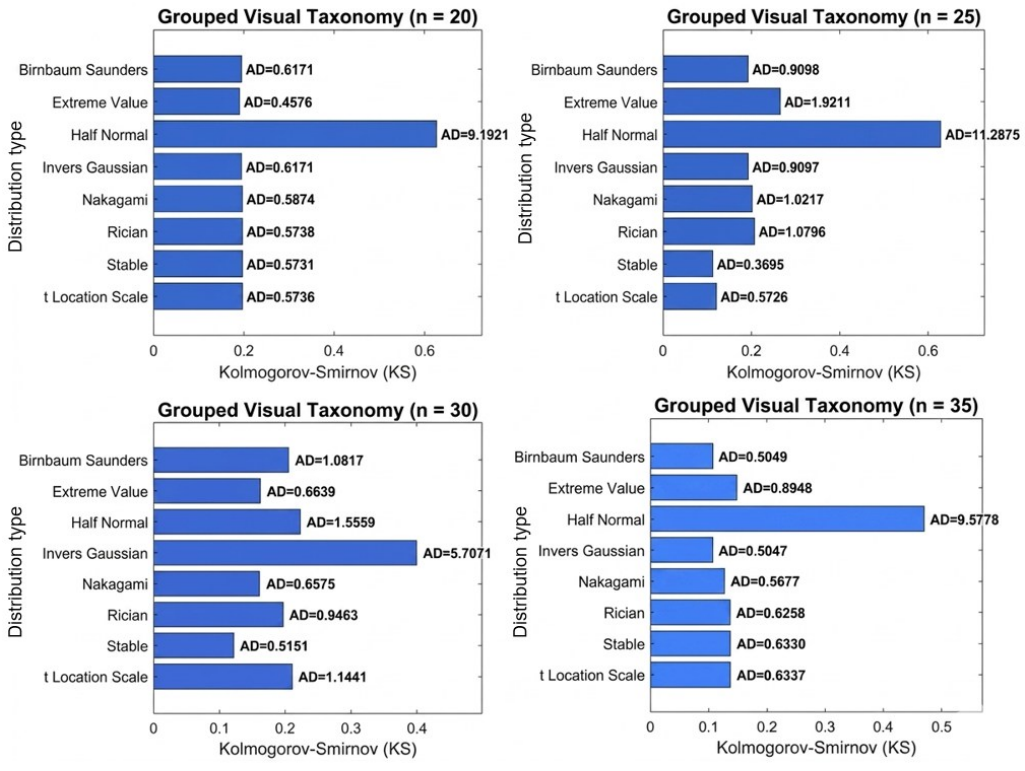

Figure 2. Grouped KS–AD evaluation across sample sizes (n =20, 25, 30, 35)

This visual taxonomy functions not only as a comparative fit quality tool but also as a proposed classification framework for understanding distributional behavior. By explicitly grouping models based on their empirical performance, it enables researchers to identify consistent, unstable, or outlier-prone distributions across varying sample sizes. For instance, the consistently low KS and AD values associated with the inverse Gaussian and extreme value distributions indicate their robustness in representing real-world chlorophyll variability. This is especially important for detecting ecologically significant phenomena such as seasonal algal blooms and environmental stressors.

In contrast, the half-normal distribution exhibits clear signs of instability, particularly at smaller sample sizes, which limits its reliability for ecological interpretation. Its high AD values across multiple sample sizes indicate poor GoF and potential misclassification risks. Hence, the taxonomy contributes not only to statistical model selection based on GoF but also adds ecological relevance—distinguishing between models that capture stable chlorophyll behavior and those that signal high variability or ecological extremes (e.g., potential bloom events).

Figure 3 displays the spatial distribution of chlorophyll sampling points and the resulting distribution-based clusters across selected marine locations. Here, cluster 1 includes only the half-normal distribution and is visually and statistically distinct from cluster 2, which comprises the remaining seven models. Inset maps further highlight two focus regions—the Black Sea and West Africa—emphasizing the distribution of sample sizes used across different marine environments. This spatial clustering visualization reinforces the interpretation of model behavior under varying environmental conditions and supports its application in regional chlorophyll monitoring strategies.

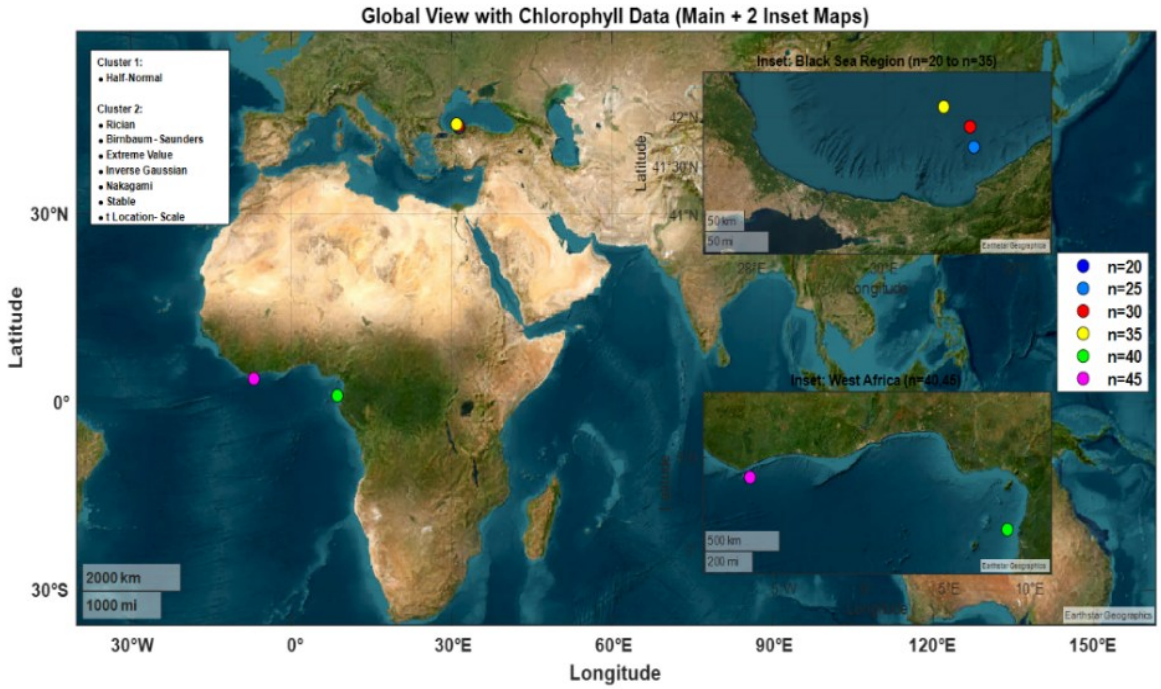

Figure 3. Chlorophyll-a sampling locations and clustering results with regional insets

3.3. Evaluation of model complexity and fit using information criteria

To complement the GoF results, this study employed five information criteria—HQC, CAIC, AICc, BIC, and AIC—to evaluate the trade-off between model fit and complexity. These criteria support the identification of models that achieve both high accuracy and parsimony, thereby minimizing the risk of overfitting. Table 3 presents the average information criterion values across four sample sizes (n =20, 25, 30, 35). The inverse Gaussian and extreme value distributions consistently yielded the lowest scores, indicating strong performance in both statistical fit and model parsimony. These models also performed well under the more stringent penalties imposed by CAIC and HQC, reinforcing their robustness.

In contrast, the half-normal distribution exhibited the highest (i.e., least favorable) values across all criteria, reflecting its limited suitability for modeling chlorophyll-a data—particularly in scenarios involving

outliers or data heterogeneity. These findings are consistent with the earlier GoF analysis and confirm inverse Gaussian and extreme value as the most reliable candidate models. Such consistency further supports their selection in the subsequent clustering and classification stages.

Table 3. Average information criteria values for chlorophyll distribution models

Distribution	Average criteria value (n =20, 25, 30, 35)				
	AIC	BIC	AICc	CAIC	HQC
Rician	-93.729	-91.143	-13.61	-13.324	-92.97
Birnbaum–Saunders	-95.677	-93.091	-94.669	-94.383	-94.918
Extreme value	-88.046	-85.46	-95.035	-94.749	-87.287
Half-normal	-25.924	-23.338	-91.903	-91.617	-25.165
Inverse Gaussian	-95.684	-93.099	-93.576	-93.291	-94.925
Nakagami	-94.63	-92.044	-93.051	-92.766	-93.871
Stable	-91.526	-86.355	-46.481	-45.451	-90.008
T location-scale	-92.076	-88.197	-95.03	-94.436	-90.938

3.4. Clustering analysis

This section applies k-means clustering to group probability distribution models based on their KS and AD statistics. Cluster quality is assessed using the silhouette index, where values approaching 1 indicate well-separated and cohesive clusters. Unlike traditional model-by-model evaluation, this approach emphasizes collective distributional behavior across varying data complexities, offering a more integrative perspective on model performance. Clustering was initially performed using sample sizes of n =20, 25, 30, and 35, and subsequently extended to n =15, 40, 45, and 50 to evaluate the robustness of the results. As shown in Table 4, the inverse Gaussian and extreme value distributions consistently appeared in well-defined clusters across all sample sizes, demonstrating strong stability and high silhouette index scores. In contrast, the Half-normal distribution was repeatedly isolated in cluster 1, exhibiting low SH values, which confirms its instability when evaluated using the combined KS–AD metrics.

Table 4. Silhouette index values from K-means clustering across distribution models

Distribution	K-means cluster evaluation with SH							
	n =15	v =20	n =25	n =30	n =35	n =40	n =45	n =50
Rician	Cluster 2	Cluster 2	Cluster 2	Cluster 2	Cluster 2	Cluster 2	Cluster 2	Cluster 2
Birnbaum–Gaunders	Cluster 2	Cluster 2	Cluster 2	Cluster 2	Cluster 2	Cluster 2	Cluster 2	Cluster 2
Extreme value	Cluster 2	Cluster 2	Cluster 2	Cluster 2	Cluster 2	Cluster 2	Cluster 2	Cluster 2
Half-normal	Cluster 1	Cluster 1	Cluster 1	Cluster 1	Cluster 1	Cluster 1	Cluster 1	Cluster 1
Inverse Gaussian	Cluster 2	Cluster 2	Cluster 2	Cluster 2	Cluster 2	Cluster 2	Cluster 2	Cluster 2
Nakagami	Cluster 2	Cluster 2	Cluster 2	Cluster 2	Cluster 2	Cluster 2	Cluster 2	Cluster 2
Stable	Cluster 2	Cluster 2	Cluster 2	Cluster 2	Cluster 2	Cluster 2	Cluster 2	Cluster 2
T location-scale	Cluster 2	Cluster 2	Cluster 2	Cluster 2	Cluster 2	Cluster 2	Cluster 2	Cluster 2

Distributions such as Birnbaum–Saunders and Nakagami displayed moderate consistency—frequently assigned to cluster 2, but with less distinct separation compared to the top-performing models. These findings underscore the effectiveness of using KS and AD statistics as clustering attributes and provide additional insight into distributional behavior beyond individual model fit. This indicates that these models offer reasonable performance, although they are less stable compared to the best-performing distributions.

To visualize these findings, Figure 4 compares empirical chlorophyll-a data (gray histogram) against fitted distributions for four sample sizes (n =20–35), where each subplot shows the overlay of empirical data and fitted PDFs of eight candidate distributions under different sampling conditions: Figure 4(a) n =20, Figure 4(b) n =25, Figure 4(c) n =30, and Figure 4(d) n =35. Inverse Gaussian and extreme value closely follow the empirical data and capture asymmetry and extremes, while half-normal consistently deviates at the distribution tails, confirming poor ecological fit. Other models like Birnbaum–Saunders and Nakagami perform reasonably but are less precise.

Overall, this clustering framework provides deeper insight into model behavior, improving reliability in selection and enhancing ecological interpretation. Models forming distinct, stable clusters—particularly inverse Gaussian and extreme value—are better suited to capture chlorophyll variability relevant to phenomena like blooms or nutrient upwelling. In contrast, unstable models like half-normal may lack ecological significance.

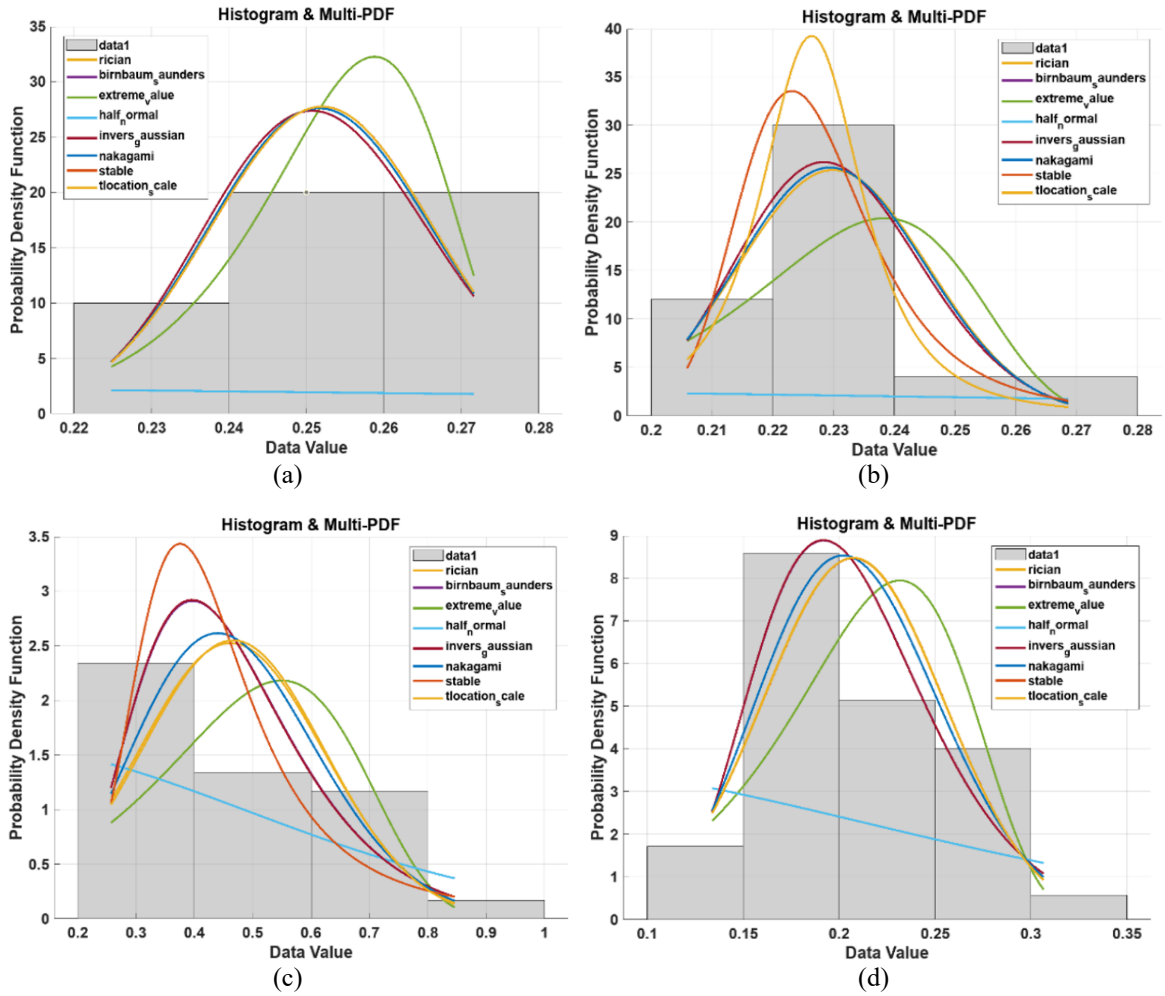

Figure 4. Comparison of empirical chlorophyll-a data (gray histogram) and fitted probability density functions (PDFs) of eight candidate distributions across different sample sizes: (a) n =20, (b) n =25, (c) n =30, and (d) n =35

3.4.1. Methodological considerations and policy integration

A key limitation of this study lies in its reliance on univariate probability distribution models, which do not account for spatial dependencies or regional variability in marine chlorophyll data. While clustering methods such as k-means and the silhouette index provide valuable insights into distributional behavior, they overlook spatial relationships among sampling points—potentially limiting their ecological representativeness. Nonetheless, specific distributions—such as the inverse Gaussian, extreme value, and Birnbaum–Saunders—effectively capture key features like skewness, stability, and heavy tails, which are often associated with environmental stressors (e.g., algal blooms). Integrating these models into predictive frameworks may improve chlorophyll fluctuation estimates and support ecosystem-based strategies, including fisheries planning and eutrophication mitigation.

To enhance both model robustness and ecological relevance, future research should explore multivariate or spatially explicit approaches, such as geostatistical models (e.g., kriging), spatial autoregression, or spatiotemporal machine learning techniques. These methods can account for spatial autocorrelation and facilitate integration with remote sensing and geographic information system layers, supporting real-time monitoring and localized marine ecosystem management. In addition, the clustering framework could be strengthened by incorporating alternative GoF statistics—such as the Cramér–von Mises, Liliefors, or Shapiro–Wilk tests—which provide complementary perspectives on distributional fit. These tests are particularly useful in small-sample contexts or when targeting specific distributional characteristics, thereby enriching the clustering feature space and improving model discrimination.

From an operational perspective, the proposed modeling framework holds strong potential for integration into satellite-based decision support systems (DSS). By combining statistical modeling with

clustering, it enables the delineation of ecological zones, sensor calibration, and early warning system enhancement. Marine stakeholders—such as fisheries agencies, policy makers, and environmental managers—can apply this framework to support data-driven, ecologically informed marine resource governance.

4. CONCLUSION

This study proposed an integrated statistical framework for evaluating and clustering probability distribution models to analyze chlorophyll-a variability in marine environments. Eight candidate distributions were assessed using GoF tests and clustering techniques based on model performance metrics. The results indicate that the inverse Gaussian and extreme value distributions consistently provided the most accurate representations of chlorophyll distribution across varying sample sizes. These models also exhibited strong clustering performance, as reflected by high silhouette index scores, indicating their stability and clear separation from other models. In contrast, the half-normal distribution demonstrated substantial misfit and instability, particularly at smaller sample sizes. Meanwhile, models such as Birnbaum–Saunders and Nakagami showed moderately strong results and may serve as complementary options for future exploration. Beyond methodological insights, the proposed framework offers practical relevance for integration into satellite-based DSS—supporting enhanced marine monitoring and evidence-based policy-making. By emphasizing robust statistical model selection, the findings contribute to data-driven marine resource management, particularly in detecting HABs and other ecologically significant phenomena. This research aligns with global sustainability goals—particularly SDG 14, which emphasizes the protection and sustainable use of marine ecosystems.

FUNDING INFORMATION

This research was supported by the Education Fund Management Agency (LPDP), Ministry of Finance of the Republic of Indonesia, through the Indonesian Endowment Fund for Education scholarship program under contract no. LOG-6603/LPDP/LPDP.3/2023.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Felix Reba	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓
Toha Saifudin	✓	✓			✓	✓	✓			✓		✓	✓	
Rimuljo Hendradi	✓	✓		✓	✓	✓	✓			✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

DATA AVAILABILITY

The data that support the findings of this study were obtained from the Copernicus Marine Service at https://data.marine.copernicus.eu/ and are openly accessible to the public.

REFERENCES

[1] T. Zhang, L. Song, H. Yuan, B. Song, and A. Ngando, "A comparative study on habitat models for adult bigeye tuna in the Indian ocean based on gridded tuna longline fishery data," *Fisheries Oceanography*, vol. 30, no. 5, pp. 584–607, 2021, doi: 10.1111/fog.12539.

[2] J. Zhai, Z. Li, R. Wan, S. Tian, P. Song, and H. Lin, "Effects of estuarine environmental heterogeneity on the habitat of gobiidea species larvae," *Marine and Coastal Fisheries*, vol. 15, no. 3, 2023, doi: 10.1002/mcf2.10241.
[3] A. Wang and H. Su, "Spatio-temporal neighbors adaptive learning with two-point differences for ocean subsurface temperature reconstruction from 1960 to 2022," *International Journal of Digital Earth*, vol. 18, no. 1, 2025, doi: 10.1080/17538947.2025.2500525.
[4] M. Wallner, "A half-normal distribution scheme for generating functions," *European Journal of Combinatorics*, vol. 87, 2020, doi: 10.1016/j.ejc.2020.103138.
[5] H. M. Ali, M. Z. Ahmed, R. E. Khamiss, and A. Karim, "Evaluation of nine two-parameter probability distributions for modeling wind speed data at three sites in Western Mountain-Libya," in *2022 IEEE 2nd International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA),* Sabratha, Libya, 2022, pp. 767-771, doi: 10.1109/MI-STA54861.2022.9837731.
[6] A. Karakus, E. E. Kuruoğlu, and A. Achim, "A generalized gaussian extension to the rician distribution for SAR image modeling," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, 2022, doi: 10.1109/TGRS.2021.3069091.
[7] S. Drozd, N. Kussul, and A. Shelestov, "Improving spatial resolution of aqua MODIS and GCOM-C chlorophyll-a data for cyprus coastal waters monitoring," *European Journal of Remote Sensing*, vol. 58, no. 1, 2025, doi: 10.1080/22797254.2025.2551023.
[8] S. Radfar, P. Galiatsatou, and T. Wahl, "Application of nonstationary extreme value analysis in the coastal environment-a systematic literature review," *Weather and Climate Extremes*, vol. 41, 2023, doi: 10.1016/j.wace.2023.100575.
[9] Y. Xu, L. Lu, L. Yan, and H. K. Zhang, "A review of remote sensing in coastal aquaculture: data, geographic hotspots, methods, and challenges," *GIScience & Remote Sensing*, vol. 62, no. 1, 2025, doi: 10.1080/15481603.2025.2573529.
[10] G. Fallahgoul, H. Hassan, and G. Loeper, "Modelling tail risk with tempered stable distributions: an overview," *Annals of Operations Research*, vol. 299, pp. 1253–1280, 2021, doi: 10.1007/s10479-019-03204-3.
[11] X. Li and J. Ma, "Non-central student-t mixture of student-t processes for robust regression and prediction," in *Intelligent Computing Theories and Application*, Cham, Switzerland: Springer, 2021, pp. 499–511, doi: 10.1007/978-3-030-84522-3_41.
[12] M. A. U. Haq, G. S. Rao, M. Albassam, and M. Aslam, "Marshall–olkin power lomax distribution for modeling of wind speed data," *Energy Reports*, vol. 6, pp. 1118–1123, 2020, doi: 10.1016/j.egyr.2020.04.033.
[13] K. S. Guedes, C. F. D. Andrade, P. A. C. Rocha, R. D. S. Mangueira, and A. D. Moura, "Performance analysis of metaheuristic optimization algorithms in estimating the parameters of several wind speed distributions," *Applied Energy*, vol. 268, 2020, doi: 10.1016/j.apenergy.2020.114952.
[14] X. P. Nguyen, T. D. Nguyen, D. N. Nguyen, R. Hidayat, T. T. Huynh, and D. T. Nguyen, "A review on the internet of thing (IoT) technologies in controlling ocean environment," *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, vol. 47, no. 1, pp. 10064–10082, 2021, doi: 10.1080/15567036.2021.1960932.
[15] Z. Zhang, G. Ren, P. Wu, Y. Hu, S. Liu, and F. Zhang, "SAA-VMamba: cross-domain adaptive remote sensing recognition method of Sonneratia apetala," *International Journal of Digital Earth*, vol. 18, no. 2, 2025, doi: 10.1080/17538947.2025.2564260.
[16] S. W. Ling, H. Wang, and J. Zhang, "A review of mangrove degradation assessment using remote sensing: advances, challenges, and opportunities," *GIScience & Remote Sensing*, vol. 62, no. 1, 2025, doi: 10.1080/15481603.2025.2491920.
[17] M. M. Badr, "Goodness-of-fit tests for the compound rayleigh distribution with application to real data," *Heliyon*, vol. 5, no. 8, 2019, doi: 10.1016/j.heliyon.2019.e02225.
[18] A. M. C. D. Souza, F. Aristone, W. A. Fernandes, A. P. G. Oliveira, Z. Olaofe, and M. Abreu, "Analysis of ozone concentrations using probability distributions," *Ozone: Science & Engineering*, vol. 42, no. 6, pp. 539–550, 2020, doi: 10.1080/01919512.2020.1736987.
[19] W. Zheng, D. Lai, and N. Gould, "A simulation study of a class of nonparametric test statistics: a close look of empirical distribution function-based tests," *Communications in Statistics - Simulation and Computation*, vol. 52, no. 3, pp. 1133–1148, 2023, doi: 10.1080/03610918.2021.1874987.
[20] S. Benchiha, A. I. Al-Omari, and M. Alomani, "Goodness-of-fit tests for weighted generalized quasi-lindley distribution using SRS and RSS with applications to real data," *Axioms*, vol. 11, no. 10, 2022, doi: 10.3390/axioms11100490.
[21] Ž. Lukić and M. Milošević, "Characterization-based approach for construction of goodness-of-fit test for lévy distribution," *Statistics: A Journal of Theoretical and Applied Statistics*, vol. 57, no. 5, pp. 1087–1116, 2023, doi: 10.1080/02331888.2023.2238236.
[22] H. M. Goual and H. Yousof, "Validation of burr XII inverse rayleigh model via a modified chi-squared goodness-of-fit test," in *Journal of Applied Statistics*, 2020, pp. 393–423, doi: 10.1080/02664763.2019.1639642.
[23] H. S. Bakouch, T. Hussain, C. Chesneau, and T. Jónás, "A notable bounded probability distribution for environmental and lifetime data," *Earth Science Informatics*, vol. 15, no. 3, pp. 1607–1620, 2022, doi: 10.1007/s12145-022-00811-w.
[24] M. Mahmoud, A. I. Boghdady, A. R. A. El-Fikky, and M. H. Aly, "Statistical studies using goodness-of-fit techniques with dynamic underwater visible light communication channel modeling," *IEEE Access*, vol. 9, pp. 56378–56391, 2021, doi: 10.1109/ACCESS.2021.3072689.
[25] A. Banibayat, H. G. Kharazi, H. Eslami, S. Khoshnavaz, and B. Dahanzadeh, "Drought monitoring in bivariate probabilistic framework for the maximization of water use efficiency," *Iranian Journal of Science and Technology, Transactions of Civil Engineering*, vol. 46, no. 2, pp. 499–514, 2022, doi: 10.1007/s40996-021-00589-9.
[26] H.-Y. Cho, "Normality test of the water quality monitoring data in harbour," *Journal of the Korean Society of Coastal and Ocean Engineers*, vol. 33, no. 2, pp. 53–62, 2021, doi: 10.9765/kscoe.2021.33.2.53.
[27] C. E.-Sandoval, "Mixture probability distributions for low-flow frequency analysis in mexico: implications for environmental impact assessment, drought management, and regional water policy," *Environments*, vol. 12, no. 12, 2025, doi: 10.3390/environments12120450.
[28] S. Eo, H. Park, and Y. R. Kim, "Predictive & representative composite score," *Quality and Quantity*, vol. 57, no. 1, pp. 1–14, 2023, doi: 10.1007/s11135-022-01345-5.
[29] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data," *Information Sciences*, vol. 630, pp. 119–147, 2023, doi: 10.1016/j.ins.2022.11.139.
[30] B. Sadeghi, "Clustering in geo-data science: navigating uncertainty to select the most reliable method," *Ore Geology Reviews*, vol. 181, 2025, doi: 10.1016/j.oregeorev.2025.106591.
[31] M. N. N. Shutaywi and N. N. Kachouie, "Silhouette analysis for performance evaluation in machine learning with applications to clustering," *Entropy*, vol. 23, no. 6, 2021, doi: 10.3390/e23060759.
[32] A. M. Bagirov, R. M. Aliguliyev, and N. Sultanova, "Finding compact and well-separated clusters: clustering using silhouette coefficients," *Pattern Recognition*, vol. 135, 2023, doi: 10.1016/j.patcog.2022.109144.

[33] A. Singh and P. Ojha, "Stable clustering of offshore downhole data using a combined k-means and Gaussian mixture modelling approach," *Marine Geophysical Research*, vol. 43, no. 1, 2022, doi: 10.1007/s11001-022-09498-6.
[34] B. F. Jönsson *et al.*, "Using probability density functions to evaluate models (PDFEM, v1.0) to compare a biogeochemical model with satellite-derived chlorophyll," *Geoscientific Model Development*, vol. 16, no. 11, pp. 4639–4662, 2023, doi: 10.5194/gmd-16-4639-2023.

BIOGRAPHIES OF AUTHORS

Felix Reba earned his bachelor's degree from Universitas Cenderawasih and obtained his master's degree from Universitas Gadjah Mada. Since 2017, he has been a lecturer at Universitas Cenderawasih. His research interests include computational statistics, data mining, and machine learning. He is currently a doctoral student in the Mathematics and Natural Sciences Program at the Faculty of Science and Technology, Universitas Airlangga, Surabaya. He can be contacted at email: felix.reba-2023@fst.unair.ac.id.

Toha Saifudin completed his undergraduate studies at Universitas Airlangga, pursued his master's degree at Institut Teknologi Sepuluh Nopember, and earned his doctoral degree from Universitas Airlangga. He is currently serving as a lecturer in the Statistics Study Program and Mathematics Study Program, Faculty of Science and Technology, Universitas Airlangga. His research interests focus on computational statistics and spatial statistics, including statistical modeling, data analysis, and nonparametric regression, particularly in developing innovative statistical methods for spatial data analysis and their applications across various fields. He is actively involved in research activities and scientific publications in his area of expertise. He can be contacted at email: tohasaifudin@fst.unair.ac.id.

Rimuljo Hendradi received both his bachelor's and master's degrees from Universitas Gadjah Mada and later pursued his doctoral degree at Institut Teknologi Sepuluh Nopember. He is currently a lecturer in the Information Systems Study Program, Faculty of Science and Technology, Universitas Airlangga. His research interests encompass biomedical signal processing, business intelligence, and decision support systems, with a focus on developing advanced systems for data analysis and decision-making support in healthcare and business contexts. He is actively engaged in research activities and scientific publications within his area of expertise. He can be contacted at email: rimuljohendradi@fst.unair.ac.id.