

## Research article

---

# Clustering Analysis of Univariate Probability Distribution Model

Felix Reba<sup>1,2</sup>, Toha Saifudin<sup>3,4\*</sup> and Rimuljo Hendradi<sup>5,6</sup>

<sup>1</sup>Doctoral Program of Mathematics and Natural Sciences, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia

<sup>2</sup>Mathematics Department, Faculty of Mathematics and Natural Sciences, Universitas Cenderawasih, Jayapura, Indonesia

<sup>3</sup>Mathematics Department, Faculty of Sciences and Technology, Universitas Airlangga, Surabaya, Indonesia

<sup>4</sup>Research Group of Statistical Modeling in Life Sciences, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia

<sup>5</sup>Information Systems Study Program, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia

<sup>6</sup>Research Group of Center for Business Intelligence (CenBI), Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia

Received: 16 December 2024, Revised: 16 September 2025, Accepted: 18 September 2025, Published: 20 July 2026

## Abstract

Accurately modeling chlorophyll distribution is essential for understanding plankton dynamics and assessing marine ecosystem health. This study evaluates the suitability of various univariate probability distribution models for chlorophyll data, including the exponential, gamma, log-normal, logistic, log-logistic, normal, Rayleigh, Generalized Extreme Value (GEV), and Weibull distributions. The model selection process incorporates Goodness-of-Fit (GoF) tests (Kolmogorov-Smirnov and Anderson-Darling) and information criteria (AIC, BIC, AICc, CAIC, and HQC) to assess both statistical fit and model complexity. Additionally, the k-means clustering algorithm is applied to classify distributions based on their GoF test performance, with the Calinski-Harabasz Index (CHI) used to determine the optimal number of clusters. The results indicate that the gamma, log-normal, logistic, log-logistic, normal, and GEV distributions consistently demonstrate superior fit across multiple sample sizes, making them the most appropriate models for chlorophyll data. Conversely, the Exponential and Rayleigh distributions exhibit poor performance, particularly for larger datasets, due to their inability to capture data heterogeneity. The clustering analysis reveals that the GEV distribution tends to group with the exponential distribution in small samples ( $n = 8$ ) but aligns with the gamma and log-normal distributions as the sample size increases ( $n = 17, 18, 56$ ). This finding suggests that sample size significantly influences distribution stability and clustering behavior. This study contributes to probability distribution modeling by integrating GoF-based clustering approaches, reducing ambiguity in model selection for heterogeneous marine datasets.

---

\*Corresponding author: E-mail: tohasaifudin@fst.unair.ac.id

<https://doi.org/10.55003/cast.2026.265686>

Copyright © 2026 by King Mongkut's Institute of Technology Ladkrabang, Thailand. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

The findings emphasize the importance of selecting flexible distributions, particularly in ecological and oceanographic applications, where data variability is high. Future research should explore alternative GoF tests (e.g., Cramer-von Mises, Kuiper) and advanced clustering methods (e.g., DBSCAN, Gaussian Mixture Models) to enhance distribution selection for complex marine environments, supporting Marine Life-related SDGs.

**Keywords:** probability distribution; goodness of fit; k-means clustering; information criteria; chlorophyll data; Marine Life

## 1. Introduction

Univariate probability distributions play a critical role in various fields, including wind speed analysis, extreme weather prediction, environmental risk modeling, and public health monitoring. A proper understanding of probability distributions helps researchers predict and interpret patterns of individual variables within complex datasets, thereby supporting informed decision-making across disciplines (Ozay & Celiktas, 2016; Guedes et al., 2020; ul Haq et al., 2020; Lima et al., 2021).

In marine analysis, probability distributions are used to model data such as chlorophyll concentration, which is an important indicator of plankton presence in the ocean, with direct implications for marine ecosystem studies (Indrayani et al., 2018). However, heterogeneous empirical data often do not follow a single distribution but may represent a combination of several distributions. While distributions such as normal, Poisson, and binomial are commonly used as basic models, the complexity of heterogeneous data requires a more flexible approach in selecting probability distributions (Ouarda et al., 2016; Guedes et al., 2020).

This research focuses on evaluating the fit of various probability distributions to chlorophyll data, which serves as an indicator of plankton presence (Bishop, 2016; Rajakaruna et al., 2018). The tested distributions include the exponential, gamma, log-normal, logistic, log-logistic, normal, Rayleigh, Generalized Extreme Value (GEV), and Weibull distributions. Specifically, the gamma, log-normal, log-logistic, normal, Rayleigh, and Weibull distributions were evaluated in research that reviewed and compared the popular models for wind speed probability distribution (Wang et al., 2016; Guedes et al., 2020). Similarly, logistic distributions were examined in the context of the Goodness of Fit Test by Gulati and Shapiro (2009) and extreme rainfall events in Rio de Janeiro State, Brazil were investigated by Lima et al. (2021). Each distribution was evaluated using Goodness of Fit (GoF) tests such as the Kolmogorov–Smirnov (KS) and Anderson–Darling (AD) tests, which helped assess how well these distributions fitted the empirical data being analyzed. These tests were extensively discussed by Yazici and Yolacan (2007), Kumar et al. (2017), and de Souza et al. (2020). Additionally, information criteria such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Corrected Akaike Information Criterion (AICc), Consistent Akaike Information Criterion (CAIC), and Hannan–Quinn Criterion (HQC) were used to assess the balance between data fit and model complexity (Mohammadi et al., 2017; Pobočková et al., 2017; Basheer, 2019; Mansoor et al., 2019; Goual & Yousof, 2020).

This research also aims to address the gap in previous literature regarding more flexible methods for modeling probability distributions for heterogeneous data, particularly in marine studies (Indrayani et al., 2018). The conventional approach of selecting the best probability distribution based on GoF tests often results in inconsistencies due to variations in test statistics. Different GoF tests may produce conflicting rankings, making it

challenging for researchers to determine the most suitable distribution to represent actual data. For instance, while one distribution may yield the lowest KS statistic, indicating a good overall fit, the same distribution may have a high AD statistic, suggesting poor tail behaviour, as demonstrated in comparative studies of GoF tests by Yazici and Yolacan (2007).

This discrepancy raises concerns about the reliability of single-model selection, as previous studies have demonstrated the sensitivity of GoF tests to sample size and tail characteristics (Zheng et al., 2023). Therefore, clustering-based approaches have emerged as a viable alternative, grouping distributions with similar fitting characteristics and providing a more holistic interpretation of the data. By clustering distributions based on their statistical behaviour, this method enhances the robustness of the analysis and mitigates uncertainty in model selection (Zheng et al., 2023). Thus, this study aims to identify the most appropriate distributions for modeling highly variable data such as chlorophyll data using a clustering-based strategy. The primary contribution of this research is the evaluation and clustering of univariate probability distributions for chlorophyll data. This approach addresses challenges in modeling heterogeneous marine data and improves the understanding of appropriate distribution selection, ultimately enhancing the accuracy of models predicting plankton presence patterns.

## 2. Materials and Methods

### 2.1 Probability distribution models

Chlorophyll distribution in marine ecosystems often exhibits positively skewed patterns, characterized by high variability and heavy tails, which are difficult to model using discrete distributions such as Poisson or Binomial (Siegel et al., 2002; Behrenfeld & Boss, 2014). Previous studies have shown that chlorophyll data not only exhibit spatial and temporal variability but also often follow continuous distributions—such as Gamma, Log-Normal, and Weibull—which are better suited to capture high asymmetry in data patterns (Campbell, 1995; Devred et al., 2011; Lombard et al., 2019).

Moreover, differences in distributional characteristics between plankton and chlorophyll have been widely reported. While plankton tends to follow discrete distributions, chlorophyll concentration typically adheres to continuous distributions, which are more capable of representing skewness and large-scale variability (Irwin et al., 2006; Lombard et al., 2019).

Adopting continuous distributions is not only more accurate for representing chlorophyll variability but is also highly relevant for ecological modelling and prediction. These models provide more robust statistical analyses and are better suited for tracking ecosystem changes driven by climate variability (Dutkiewicz et al., 2015).

In addition, continuous distributions have been extensively used in marine biogeochemical models, where the distribution of ecosystem variables is influenced by complex environmental conditions and observational scales (Irwin et al., 2006; Dutkiewicz et al., 2015). Specifically, the gamma and log-normal distributions have been shown to effectively capture asymmetry in marine chlorophyll data (Campbell, 1995; Siegel et al., 2002). Therefore, although discrete models are common in plankton studies, continuous probability distributions such as the gamma, log-normal, or Weibull are recommended for better representation of chlorophyll variability in marine ecosystems (Lombard et al., 2019).

The exponential distribution is commonly used to model waiting times between independent events and was included in this study as a simple baseline model. The gamma

distribution, as a generalization of exponential distribution, provides greater flexibility for modeling positive and right-skewed data. Meanwhile, the Log-Normal distribution is appropriate when the logarithm of the data is approximately normally distributed, a characteristic often observed in chlorophyll data. The candidate probability distributions evaluated in this study are summarized in Table 1.

**Table 1.** Candidate probability distribution models and their parameters

| Distribution | Distribution Model                                                                                                                                                                              | Parameter                                                                  |     |
|--------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------|-----|
| Exponential  | $f(x) = \frac{1}{\theta} \exp\left(-\frac{x}{\theta}\right)$                                                                                                                                    | $\theta = \text{Scale}$                                                    | (1) |
| Gamma        | $f(x) = \frac{x^{a-1}}{b^a \Gamma(a)} \exp\left(-\frac{x}{b}\right)$                                                                                                                            | $a = \text{Shape}$<br>$b = \text{Scale}$                                   | (2) |
| Log-Normal   | $f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\log x - \mu)^2}{2\sigma^2}\right\}$                                                                                                    | $\mu = \text{Location}$<br>$\sigma = \text{Scale}$                         | (3) |
| Logistic     | $f(x) = \frac{\exp\left(\frac{x-\mu}{\sigma}\right)}{\sigma(1 + \exp\left(\frac{x-\mu}{\sigma}\right))^2}$                                                                                      | $\mu = \text{Location}$<br>$\sigma = \text{Scale}$                         | (4) |
| Log-Logistic | $f(x) = \frac{1}{\sigma x} \frac{\exp\left(\frac{\ln x - \mu}{\sigma}\right)}{\left(1 + \exp\left(\frac{\ln x - \mu}{\sigma}\right)\right)^2}$                                                  | $\mu = \text{Location}$<br>$\sigma = \text{Scale}$                         | (5) |
| Normal       | $f(x) = \frac{\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)}{\sigma\sqrt{2\pi}}$                                                                                             | $\mu = \text{Location}$<br>$\sigma = \text{Scale}$                         | (6) |
| Rayleigh     | $f(x) = \frac{x}{b^2} \exp\left(-\frac{1}{2}\frac{x^2}{b^2}\right)$                                                                                                                             | $b = \text{Scale}$                                                         | (7) |
| GEV          | $f(x) = \frac{1}{\sigma} \left[1 + \xi \left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}-1} \exp\left\{-\left[1 + \xi \left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}$ | $\mu = \text{Location}$<br>$\sigma = \text{Scale}$<br>$\xi = \text{Shape}$ | (8) |
| Weibull      | $f(x) = \frac{b}{a} \left(\frac{x}{a}\right)^{b-1} \exp\left(-\left(\frac{x}{a}\right)^b\right)$                                                                                                | $a = \text{Scale}$<br>$b = \text{Shape}$                                   | (9) |

Note: All distributions in the Table are univariate, as they involve only a single random variable, and parametric, as they are defined by specific parameters.

The candidate distributions in Table 1 were selected to represent a range of continuous model families for positive-valued data. Gamma, Log-Normal, and Weibull distributions are commonly considered for right-skewed environmental variables, whereas Exponential and Rayleigh were included as simpler comparative models. Logistic and Log-Logistic distributions were added to broaden the comparison to alternative symmetric and skewed forms, while the GEV distribution was considered to evaluate possible tail-extreme behavior in the observed data. This comparative framework was intended to identify the most suitable distribution for the chlorophyll data rather than to assume that all listed models are equally appropriate a priori.

## 2.2 Parameter estimation

The parameters of the candidate distributions were estimated using maximum likelihood estimation. For models without closed-form solutions, numerical optimization based on the Newton–Raphson iterative method was applied. The update formula is given as:

$$\hat{\theta}_{k+1} = \hat{\theta}_k - H(\theta_{(k)})^{-1} g(\theta_{(k)}) \quad (10)$$

This method is a widely used optimization approach for finding parameter estimates that maximize the likelihood function. The Hessian matrix,  $H(\theta_{(k)})$ , representing the second derivative, provides information on the curvature of the likelihood function, while the gradient vector,  $g(\theta_{(k)})$ , indicates the direction of change needed to reach the optimal value.

In this study, this method is applied to estimate the parameters of the probability distribution models that best fit the observed data. The Newton–Raphson method was selected due to its fast convergence and high accuracy in estimating parameters, especially for complex distributions such as gamma, log-normal, and Weibull (Jiang et al., 2017; Jung & Schindler, 2017; Guedes et al., 2020). Therefore, this approach provides optimal and precise parameter estimates for modeling the research data.

## 2.3 Test statistic value

To evaluate the fit of each probability distribution to the chlorophyll data, this study employs two main Goodness of Fit (GoF) tests. The Kolmogorov-Smirnov (KS) test measures the maximum distance between the empirical cumulative distribution function (CDF) of the sample and the hypothetical CDF. This statistic provides a robust measure of how closely the selected distribution matches the observed data. The KS statistic is formulated as follows (Balakrishnan et al., 2015; Mbah & Paothong, 2015; Badr, 2019; de Souza et al., 2020):

$$D_{calculated} = \sup_x |F(x) - H(x)| \quad (11)$$

where  $D_{calculated}$  represents the maximum vertical distance between  $F(x)$  and  $H(x)$ . The null hypothesis  $H_0$  asserts that the data follows the assumed distribution, while the alternative hypothesis  $H_1$  states that the data does not follow that distribution. The Anderson-Darling (AD) test provides greater sensitivity to the tails of the distribution compared to the KS test, which is particularly important for capturing extreme events in marine data. The AD statistic is formulated as:

$$A_n^2 = -n - \frac{1}{n} \sum_{i=1}^n (2i - 1) [\log F(x_i) + \log(1 - F(x_{n+1-i}))] \quad (12)$$

where  $A_n^2$  is the Anderson-Darling test statistic,  $F(x)$  is the cumulative distribution function, and  $n$  is the number of data points. Both tests provide robust methods for examining the fit of the distributions, and the results determine whether the selected distribution models are valid for the analyzed data.

In addition to these tests, various information criteria such as the AIC, BIC, and AICc are employed to balance model complexity with data fit. Each information criterion is formulated to consider both the complexity of the model and the log-likelihood of the data, aiming to select the optimal distribution model without overfitting.

## 2.4 Information criteria for model selection: AIC, BIC, AICc, CAIC, and HQC

The selection of a probability distribution model for chlorophyll data in this study is based on various Information Criteria (IC) to balance model fit with complexity. The criteria used include the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Corrected Akaike Information Criterion (AICc), Consistent Akaike Information Criterion (CAIC), and Hannan–Quinn Criterion (HQC) (Wasserman, 2000; Burnham & Anderson, 2002; Haggag, 2014; Ding et al., 2018; Zheng et al., 2023). Table 2 presents the Information Criteria used.

**Table 2.** Formulas for calculating information criteria for statistical models

| Information Criterion  | Formula for Calculating Information Criterion                                                            |
|------------------------|----------------------------------------------------------------------------------------------------------|
| <i>AIC</i>             | Tend to favor models with higher data fit, even if they are more complex.                                |
| <i>BIC and CAIC</i>    | Apply stronger penalties on model complexity, making them more conservative in selecting simpler models. |
| <i>AIC<sub>c</sub></i> | A corrected version of AIC for small sample sizes, reducing selection bias.                              |
| <i>HQC</i>             | Positioned between AIC and BIC, offering a balanced selection approach.                                  |

In this study, IC values were calculated for each candidate probability distribution to determine the best-fitting model. If all criteria indicate the same distribution as the best model, the selection process is straightforward. However, when IC rankings vary across different criteria, a clustering-based analysis is applied to group distributions with similar fitting characteristics (Zheng et al., 2023). This strategy helps resolve inconsistencies commonly found in conventional model selection approaches.

This issue is particularly relevant because traditional methods often lead to different conclusions depending on the criterion used, such as AIC or BIC (Burnham & Anderson, 2002; Ding et al., 2018). By integrating multiple information criteria with a clustering method, this study ensures that the selected probability distributions more accurately capture the variability patterns in chlorophyll concentrations in marine environments (Zheng et al., 2023).

## 2.5 Clustering method and evaluation

This study employs the k-means algorithm to partition data into several clusters based on the results of the Goodness-of-Fit (GoF) tests. Each iteration involves calculating the Euclidean distance between each data point and the cluster centroid. The k-means algorithm is chosen for its efficiency in handling large datasets, although it has known limitations in detecting non-linear or complex-shaped clusters. To ensure the quality of clustering results, the Calinski-Harabasz Index (CHI) is employed as an evaluation metric

to assess how well each cluster is separated from others (Jain, 2010; Van & Pham-Gia, 2010).

The integration of GoF test statistics into the k-means clustering framework addresses a critical issue in distribution selection: inconsistencies in GoF test results, where different distributions can receive conflicting rankings depending on the test used (e.g., KS vs AD). By using GoF values as clustering features, the study systematically groups probability distributions with similar fitting characteristics, reducing subjectivity and ambiguity in the model selection process. This method enables clustering of distributions based on statistical similarity, providing a more robust and reproducible way to interpret the data.

In this study, the CHI index validates the clustering results, ensuring that the groups formed reflect meaningful patterns. This clustering-based framework enhances the robustness of distribution selection by combining GoF tests with clustering analysis, as opposed to treating them separately as in traditional approaches. Ultimately, this integration allows for a systematic and data-driven solution to model uncertainty in probability distribution analysis and provides a practical approach to address the long-standing challenges in marine ecosystem modeling.

$$d_{ik} = \sum_{j=1}^m (x_{kj} - c_{ij})^2 \quad (13)$$

This method is considered efficient in clustering large and varied marine data, but to validate the quality of the clusters, the CHI is employed. The evaluation of clustering results is conducted by calculating the Between-Group Sum of Squares (BGSS) and Within-Group Sum of Squares (WGSS). The Calinski-Harabasz Index (CHI) is used to evaluate clustering quality with the formula (Wang & Xu, 2019; Ashari et al., 2023):

$$CHI = \frac{\frac{BGSS}{k-1}}{\frac{WGSS}{N-k}} \quad (14)$$

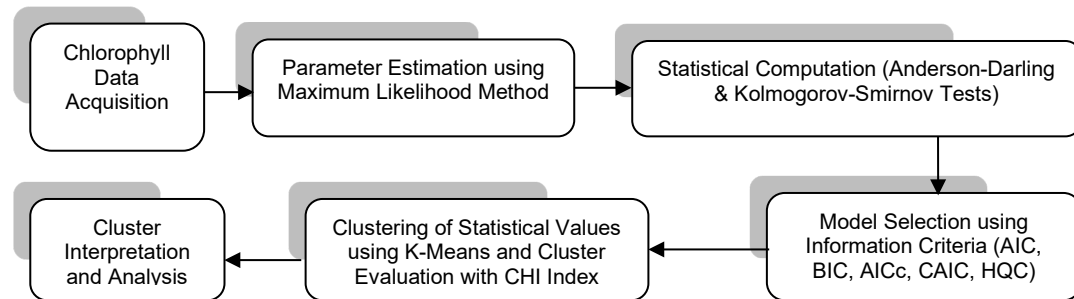
where  $k$  is the number of clusters and  $N$  is the total number of observations. The higher the CHI value, the better the clustering quality.

Figure 1 illustrates the structured workflow of this study, ensuring a systematic approach to chlorophyll data clustering. The process consists of:

1. Chlorophyll data acquisition – Collecting and preprocessing chlorophyll-related data.
2. Parameter estimation using Maximum Likelihood Estimation (MLE) – Estimating statistical distribution parameters.
3. Statistical computation (GoF Tests: Anderson-Darling & Kolmogorov-Smirnov) – Evaluating the goodness-of-fit for probability distributions.
4. Model selection (AIC, BIC, AICc, CAIC, HQC) – Identifying the best-fitting distribution model.
5. Clustering of statistical values (K-Means & CHI Index Evaluation) – Grouping chlorophyll data based on statistical patterns.

6. Cluster interpretation and analysis – Extracting insights for further environmental and ecological assessments.

The detailed methodology is visualized in Figure 1.



**Figure 1.** Structured workflow for chlorophyll data clustering and model selection

As shown in Figure 1, the integration of probability modeling, clustering, and validation ensures a robust chlorophyll data analysis. The application of MLE and GoF tests enhances model selection, ensuring that the probability distribution best represents the chlorophyll variation. The clustering process using K-Means and validation through the CHI Index further refines the classification of chlorophyll distribution. This methodological framework provides a solid basis for analyzing chlorophyll concentration variations, which can be applied to study marine productivity, water quality assessment, and ecological monitoring.

### 3. Results and Discussion

#### 3.1 Research data

The data used in this study were obtained from the Copernicus Marine Service, with a specific location in the Black Sea at coordinates 31.14015 (longitude) and 41.69411 (latitude). Sampling was conducted by considering variations in depth and pressure to capture the vertical distribution patterns in the water column. This approach allows for better representation of the spatial and temporal dynamics of chlorophyll distribution. Before being used in the analysis, the data were first examined using outlier detection based on probability domains. At this stage, the samples were compared against the domain  $X$  of the probability distributions used in modeling, namely exponential, gamma, log-normal, logistic, log-logistic, Rayleigh, GEV, and Weibull.

Samples that did not fall within the appropriate distribution domain were excluded to avoid bias and preserve the validity of the statistical models. As a result of the selection process, the sample sizes varied depending on the considered depth and pressure. The first sample set that met the criteria consisted of 8 samples, followed by sets of 17, 18, and finally 56 samples. These sample sizes remain within a reasonable range and comply with the principles of inferential statistics, as discussed in the study by Yazici & Yolacan (2007). Their research, titled *A Comparison of Various Tests of Normality*, compared 12 normality test methods using the Monte Carlo Simulation approach to evaluate the effectiveness of statistical tests in selecting probability distributions.

Their findings indicated that evaluating probability distributions requires an approach that considers data variation and sample size. In this study, the implementation of clustering-based probability distribution methods offers an alternative to address the inconsistencies in selecting the best-fitting distribution model based on GoF tests. While this approach does not necessarily result in more flexible modeling, it reduces subjectivity in model selection and mitigates reliability issues when determining the most appropriate probability distribution.

### 3.2 Test statistic values

The evaluation results presented in Table 3, using the GoF tests, reveal varying degrees of fit among different probability distributions applied to chlorophyll data. Inconsistencies in probability distribution selection—stemming from variations in GoF test values—pose significant challenges for researchers in identifying the most appropriate distribution to accurately represent empirical data. This dilemma becomes evident when different GoF tests yield contradictory rankings. For example, a distribution may exhibit a low KS statistic, indicating strong agreement with the data, yet simultaneously present a high AD statistic, suggesting poor fit in the tails. Such discrepancies introduce uncertainty in selecting the optimal probabilistic model.

To address this issue, a clustering-based approach is applied to group distributions with similar fitting behavior, reducing subjectivity and ambiguity. This method enhances the interpretability of statistical evaluation by classifying distributions that share comparable goodness-of-fit characteristics. As a result, the statistical modeling becomes more robust, ensuring greater reliability in analytical outcomes. Based on the KS and AD test results, the gamma, log-normal, logistic, log-logistic, normal, and GEV distributions consistently demonstrate high stability and superior performance in modeling chlorophyll data across a wide range of sample sizes.

A detailed examination of Table 3 reveals that gamma, log-normal, logistic, log-logistic, normal, and GEV distributions exhibit consistency across all sample sizes ( $n = 8, 17, 18$ , and  $56$ ). Although the GEV distribution is rejected in the AD test for  $n = 8$ , it maintains a consistent and robust fit for  $n = 17, 18$ , and  $56$ . This stability is evident from the high p-values and relatively low test statistics, confirming that these models adequately represent the chlorophyll dataset. Conversely, the Exponential and Rayleigh distributions exhibit significant misfit, particularly for larger sample sizes. This inadequacy arises due to their inability to capture the intricate variations present in chlorophyll data, as evidenced by their consistently low p-values and high test statistics across all sample sizes. The poor performance of these distributions underscores their lack of suitability for modeling highly heterogeneous chlorophyll data. Overall, the findings suggest that the gamma, log-normal, logistic, log-logistic, normal, and GEV distributions are the most appropriate probabilistic models for representing chlorophyll data characterized by high variability. Selecting an appropriate distribution is crucial for ensuring robust analytical results, particularly in marine studies, where data often exhibit significant heterogeneity and dynamic patterns.

### 3.3 Information criteria for model selection: AIC, BIC, AICc, CAIC, and HQC

In addition to Goodness of Fit (GoF) tests, several information criteria—including the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Corrected Akaike Information Criterion (AICc)—are employed to assess the adequacy of statistical

**Table 3.** Results of KS and AD tests for various distributions

| $n = 8$      |                         |         |        |       |                       |         |        |       |
|--------------|-------------------------|---------|--------|-------|-----------------------|---------|--------|-------|
| Distribution | Kolmogorov-Smirnov (KS) |         |        |       | Anderson-Darling (AD) |         |        |       |
|              | $H_0/H_1$               | p-Value | KSStat | CV    | $H_0/H_1$             | p-Value | ADStat | CV    |
| Exponential  | $H_1$                   | 0.009   | 0.546  | 0.454 | $H_1$                 | 0.024   | 3.116  | 2.516 |
| Gamma        | $H_0$                   | 0.746   | 0.222  | 0.454 | $H_0$                 | 0.595   | 0.651  | 2.516 |
| Log Normal   | $H_0$                   | 0.692   | 0.233  | 0.454 | $H_0$                 | 0.569   | 0.680  | 2.516 |
| Logistic     | $H_0$                   | 0.964   | 0.161  | 0.454 | $H_0$                 | 0.86    | 0.385  | 2.516 |
| Log Logistic | $H_0$                   | 0.927   | 0.177  | 0.454 | $H_0$                 | 0.779   | 0.463  | 2.516 |
| Normal       | $H_0$                   | 0.800   | 0.211  | 0.454 | $H_0$                 | 0.676   | 0.564  | 2.516 |
| Reyleigh     | $H_1$                   | 0.036   | 0.473  | 0.454 | $H_1$                 | 0.043   | 2.643  | 2.516 |
| GEV          | $H_0$                   | 0.487   | 0.277  | 0.454 | $H_1$                 | 0.031   | 2.927  | 2.516 |
| Weibull      | $H_0$                   | 0.997   | 0.126  | 0.454 | $H_0$                 | 0.929   | 0.309  | 2.516 |
| $n = 17$     |                         |         |        |       |                       |         |        |       |
| Distribution | Kolmogorov-Smirnov (KS) |         |        |       | Anderson-Darling (AD) |         |        |       |
|              | $H_0/H_1$               | p-Value | KSStat | CV    | $H_0/H_1$             | p-Value | ADStat | CV    |
| Exponential  | $H_1$                   | 0.0     | 0.582  | 0.318 | $H_1$                 | 0.000   | 6.409  | 2.503 |
| Gamma        | $H_0$                   | 0.409   | 0.206  | 0.318 | $H_0$                 | 0.364   | 0.984  | 2.503 |
| Log Normal   | $H_0$                   | 0.462   | 0.197  | 0.318 | $H_0$                 | 0.414   | 0.897  | 2.503 |
| Logistic     | $H_0$                   | 0.518   | 0.189  | 0.318 | $H_0$                 | 0.382   | 0.951  | 2.503 |
| Log Logistic | $H_0$                   | 0.547   | 0.184  | 0.318 | $H_0$                 | 0.408   | 0.907  | 2.503 |
| Normal       | $H_0$                   | 0.420   | 0.204  | 0.318 | $H_0$                 | 0.385   | 0.945  | 2.503 |
| Reyleigh     | $H_1$                   | 0.000   | 0.531  | 0.318 | $H_1$                 | 0.002   | 5.165  | 2.503 |
| GEV          | $H_0$                   | 0.552   | 0.183  | 0.318 | $H_0$                 | 0.559   | 0.696  | 2.503 |
| Weibull      | $H_0$                   | 0.309   | 0.224  | 0.318 | $H_0$                 | 0.329   | 1.053  | 2.503 |
| $n = 18$     |                         |         |        |       |                       |         |        |       |
| Distribution | Kolmogorov-Smirnov (KS) |         |        |       | Anderson-Darling (AD) |         |        |       |
|              | $H_0/H_1$               | p-Value | KSStat | CV    | $H_0/H_1$             | p-Value | ADStat | CV    |
| Exponential  | $H_1$                   | 0.0     | 0.572  | 0.309 | $H_1$                 | 0.000   | 6.734  | 2.503 |
| Gamma        | $H_0$                   | 0.929   | 0.120  | 0.309 | $H_0$                 | 0.885   | 0.360  | 2.503 |
| Log Normal   | $H_0$                   | 0.955   | 0.113  | 0.309 | $H_0$                 | 0.914   | 0.327  | 2.503 |
| Logistic     | $H_0$                   | 0.943   | 0.116  | 0.309 | $H_0$                 | 0.889   | 0.356  | 2.503 |
| Log Logistic | $H_0$                   | 0.946   | 0.115  | 0.309 | $H_0$                 | 0.882   | 0.363  | 2.503 |
| Normal       | $H_0$                   | 0.952   | 0.113  | 0.309 | $H_0$                 | 0.923   | 0.317  | 2.503 |
| Reyleigh     | $H_1$                   | 0.000   | 0.51   | 0.309 | $H_1$                 | 0.002   | 5.380  | 2.503 |
| GEV          | $H_0$                   | 0.945   | 0.116  | 0.309 | $H_0$                 | 0.859   | 0.387  | 2.503 |
| Weibull      | $H_0$                   | 0.96    | 0.111  | 0.309 | $H_0$                 | 0.878   | 0.368  | 2.503 |
| $n = 56$     |                         |         |        |       |                       |         |        |       |
| Distribution | Kolmogorov-Smirnov (KS) |         |        |       | Anderson-Darling (AD) |         |        |       |
|              | $H_0/H_1$               | p-Value | KSStat | CV    | $H_0/H_1$             | p-Value | ADStat | CV    |
| Exponential  | $H_1$                   | 0.0     | 0.587  | 0.178 | $H_1$                 | 0.0     | 21.458 | 2.496 |
| Gamma        | $H_0$                   | 0.338   | 0.122  | 0.178 | $H_0$                 | 0.342   | 1.028  | 2.496 |
| Log Normal   | $H_0$                   | 0.358   | 0.120  | 0.178 | $H_0$                 | 0.371   | 0.972  | 2.496 |
| Logistic     | $H_0$                   | 0.439   | 0.113  | 0.178 | $H_0$                 | 0.337   | 1.038  | 2.496 |
| Log Logistic | $H_0$                   | 0.421   | 0.114  | 0.178 | $H_0$                 | 0.361   | 0.991  | 2.496 |
| Normal       | $H_0$                   | 0.362   | 0.120  | 0.178 | $H_0$                 | 0.312   | 1.092  | 2.496 |
| Reyleigh     | $H_1$                   | 0.0     | 0.540  | 0.178 | $H_1$                 | 0.0     | 17.603 | 2.496 |
| GEV          | $H_0$                   | 0.456   | 0.111  | 0.178 | $H_0$                 | 0.3198  | 1.075  | 2.496 |
| Weibull      | $H_0$                   | 0.123   | 0.154  | 0.178 | $H_0$                 | 0.1442  | 1.651  | 2.496 |

Note:  $H_0$  indicates that the distribution fits the data, while  $H_1$  suggests otherwise. A p-value > 0.05 means  $H_0$  is accepted. KSStat and ADStat are test statistics from the KS and AD methods, which measure how much the data distribution deviates from the theoretical distribution. These values are compared to the *Critical Value* (CV), which serves as the threshold for accepting or rejecting  $H_0$ . If KSStat or ADStat is smaller than CV, the distribution is considered a good fit for the data.

models. The evaluation results presented in Table 4 indicate that the gamma, log-normal, logistic, log-logistic, normal, and GEV distributions exhibit lower AIC, BIC, and AICc values compared to other distributions. This suggests that these distributions offer an optimal balance between model complexity and goodness of fit, aligning with the findings from the GoF tests in Table 3. Conversely, the Exponential and Rayleigh distributions demonstrate significantly higher information criterion values, indicating their inability to effectively model chlorophyll data. This observation is consistent with the KS and AD test results in Table 3, where these distributions exhibited substantial misfit.

Higher information criterion values imply that these models lack the flexibility to capture the complex variability within the chlorophyll dataset, rendering them unsuitable for accurate statistical modeling. Although the GEV distribution exhibits relatively high AIC, BIC, and AICc values for  $n=8$ , it demonstrates substantial improvements in performance as the sample size increases to  $n=17$ , 18, and 56, where its information criterion values decrease significantly. This decline suggests that the GEV distribution becomes more stable and appropriate for larger sample sizes, reinforcing its suitability as one of the best models for chlorophyll data representation.

The combined evaluation of GoF tests and information criteria further strengthens the conclusion that the gamma, log-normal, logistic, log-logistic, normal, and GEV distributions are the most appropriate models for chlorophyll data, particularly for larger sample sizes. These distributions demonstrate enhanced performance in capturing data heterogeneity, making them more reliable for statistical modeling. In contrast, the Exponential and Rayleigh distributions exhibit significant limitations in flexibility, making them unsuitable for representing the complex structure of chlorophyll data. The consistently higher values of AIC, BIC, and AICc further confirm that these models fail to adequately fit the dataset. Therefore, selecting the appropriate distribution based on both GoF tests and information criteria is crucial for ensuring robust and accurate analytical results.

### 3.4 Clustering analysis

The test statistic values obtained in Table 3 for each sample are used in the clustering process utilizing the K-Means Clustering method. This process aims to identify distributions that exhibit similarities based on the KS and AD test statistics. By employing this method, distributions that demonstrate similar goodness-of-fit patterns will be grouped into the same cluster. The iteration in the K-Means Clustering method is performed progressively until a convergent value is achieved, meaning that changes in cluster assignment between iterations become minimal or no longer significant. This convergence process is crucial to ensure that the clustering structure accurately reflects the relationships among distributions based on KS and AD statistical measures. Once the clustering process reaches a stable state, the results are then evaluated using the CHI to determine the optimal number of clusters. Table 5 presents the clustering results using the K-Means algorithm:

Following Table 5, the clustering results reveal a clear pattern in how distributions are grouped based on their statistical characteristics. For  $n = 8$ , the GEV distribution is placed in Cluster 2, alongside the Exponential distribution. However, as the sample size increases to  $n = 17$ , 18, and 56, the GEV distribution shifts to Cluster 1, joining the gamma, log-normal, logistic, log-logistic, normal, and Weibull distributions. This transition indicates that the GEV distribution exhibits greater similarity with the distributions in Cluster 1 as the sample size increases, aligning with the pattern observed in the other distributions within this group. Meanwhile, the gamma, log-normal, logistic, log-logistic, normal, and

**Table 4.** Information criteria for various distributions on chlorophyll samples

| Distribution | Information Criteria | Chlorophyll Sample |               |               |               |
|--------------|----------------------|--------------------|---------------|---------------|---------------|
|              |                      | <i>n</i> = 8       | <i>n</i> = 17 | <i>n</i> = 18 | <i>n</i> = 56 |
| Exponential  | AIC                  | -2.282             | -10.184       | -9.693        | -40.692       |
|              | BIC                  | -2.202             | -9.3510       | -8.803        | -38.666       |
|              | AICc                 | -1.615             | -9.9175       | -9.443        | -40.618       |
|              | CAIC                 | -1.202             | -8.3510       | -7.803        | -37.666       |
|              | HQC                  | -2.817             | -10.101       | -9.570        | -39.906       |
| Gamma        | AIC                  | -31.597            | -73.330       | -76.353       | -263.563      |
|              | BIC                  | -31.438            | -71.664       | -74.572       | -259.512      |
|              | AICc                 | -29.197            | -72.473       | -75.553       | -263.336      |
|              | CAIC                 | -29.438            | -69.664       | -72.572       | -257.512      |
|              | HQC                  | -32.669            | -73.164       | -76.107       | -261.992      |
| Log Normal   | AIC                  | -31.144            | -73.437       | -76.271       | -264.367      |
|              | BIC                  | -30.985            | -71.771       | -74.491       | -260.316      |
|              | AICc                 | -28.744            | -72.580       | -75.471       | -264.141      |
|              | CAIC                 | -28.985            | -69.771       | -72.491       | -258.316      |
|              | HQC                  | -32.215            | -73.272       | -76.026       | -262.797      |
| Logistic     | AIC                  | -33.147            | -70.962       | -74.650       | -259.346      |
|              | BIC                  | -32.988            | -69.296       | -72.869       | -255.295      |
|              | AICc                 | -30.747            | -70.105       | -73.850       | -259.119      |
|              | CAIC                 | -30.988            | -67.296       | -70.869       | -253.295      |
|              | HQC                  | -34.219            | -70.797       | -74.405       | -257.775      |
| Log Logistic | AIC                  | -32.309            | -71.428       | -74.609       | -261.128      |
|              | BIC                  | -32.150            | -69.762       | -72.828       | -257.077      |
|              | AICc                 | -29.909            | -70.571       | -73.809       | -260.902      |
|              | CAIC                 | -30.150            | -67.762       | -70.828       | -255.077      |
|              | HQC                  | -33.380            | -71.262       | -74.364       | -259.558      |
| Normal       | AIC                  | -32.251            | -72.952       | -76.328       | -261.615      |
|              | BIC                  | -32.092            | -71.285       | -74.548       | -257.565      |
|              | AICc                 | -29.851            | -69.285       | -75.528       | -261.389      |
|              | CAIC                 | -30.092            | -71.818       | -72.548       | -255.565      |
|              | HQC                  | -33.323            | -72.786       | -76.083       | -260.045      |
| Rayleigh     | AIC                  | -13.177            | -33.266       | -34.132       | -116.983      |
|              | BIC                  | -13.097            | -32.433       | -33.241       | -114.958      |
|              | AICc                 | -12.510            | -33.000       | -33.882       | -116.909      |
|              | CAIC                 | -12.097            | -31.433       | -32.241       | -113.958      |
|              | HQC                  | -13.712            | -33.184       | -34.009       | -116.198      |
| GEV          | AIC                  | -44.827            | -72.677       | -75.417       | -266.959      |
|              | BIC                  | -44.588            | -70.178       | -72.745       | -260.883      |
|              | AICc                 | -38.827            | -70.831       | -73.702       | -266.497      |
|              | CAIC                 | -41.588            | -67.178       | -69.745       | -257.883      |
|              | HQC                  | -46.434            | -72.429       | -75.048       | -264.603      |
| Weibull      | AIC                  | -35.098            | -71.744       | -75.950       | -251.027      |
|              | BIC                  | -34.939            | -70.078       | -74.169       | -246.977      |
|              | AICc                 | -32.698            | -70.887       | -75.150       | -250.801      |
|              | CAIC                 | -32.939            | -68.078       | -72.169       | -244.977      |
|              | HQC                  | -36.170            | -71.578       | -75.704       | -249.457      |

Note: This Table presents information criteria (AIC, BIC, CAIC, HQC) for various distributions. Lower values indicate the best model, with AIC focusing on data fit, BIC and CAIC penalizing complexity, and HQC providing a balance.

**Table 5.** K-means evaluation with CHI for various distributions

| Distribution | Evaluation of Cluster K-Means with CHI |           |           |           |
|--------------|----------------------------------------|-----------|-----------|-----------|
|              | $n = 8$                                | $n = 17$  | $n = 18$  | $n = 56$  |
| Exponential  | Cluster 2                              | Cluster 2 | Cluster 2 | Cluster 2 |
| Gamma        | Cluster 1                              | Cluster 1 | Cluster 1 | Cluster 1 |
| Log Normal   | Cluster 1                              | Cluster 1 | Cluster 1 | Cluster 1 |
| Logistic     | Cluster 1                              | Cluster 1 | Cluster 1 | Cluster 1 |
| Log Logistic | Cluster 1                              | Cluster 1 | Cluster 1 | Cluster 1 |
| Normal       | Cluster 1                              | Cluster 1 | Cluster 1 | Cluster 1 |
| Rayleigh     | Cluster 3                              | Cluster 3 | Cluster 3 | Cluster 3 |
| GEV          | Cluster 2                              | Cluster 1 | Cluster 1 | Cluster 1 |
| Weibull      | Cluster 1                              | Cluster 1 | Cluster 1 | Cluster 1 |

Note: Clustering results of K-Means with CHI for samples  $n = 8, 17, 18$ , and  $56$ . Cluster 1 contains distributions with the best fit, while Cluster 2 and 3 represent patterns different from the main data.

Weibull distributions consistently remain in Cluster 1 across all sample sizes. This stability suggests that these distributions share similar statistical goodness-of-fit characteristics, leading to a clustering pattern that remains unchanged despite variations in sample size. These findings reinforce the conclusion that the distributions in Cluster 1 are more stable and reliable for modeling chlorophyll data. On the other hand, the Exponential and Rayleigh distributions remain in separate clusters, specifically Cluster 2 and Cluster 3, across all sample sizes. This consistent separation indicates that these distributions exhibit significant differences compared to those in Cluster 1, further supporting the earlier findings that Exponential and Rayleigh distributions are inadequate for accurately representing chlorophyll data. The clustering results demonstrate that the iteration has reached convergence at  $k = 3$ , where Cluster 1 comprises distributions with the best fit, while Cluster 2 and Cluster 3 contain distributions with distinct characteristics that deviate from the main data pattern.

### 3.5 Evaluation cluster with CHI

To determine the optimal number of clusters in this analysis, an evaluation was conducted using the CHI method. CHI is a clustering evaluation method that measures the ratio between inter-cluster and intra-cluster variability, where higher values indicate better clustering performance. To obtain the results presented in Table 6, the clustering process was performed using the K-Means Clustering algorithm for each sample size ( $n = 8, 17, 18$ , and  $56$ ) with various cluster numbers ( $k = 2, k = 3, k = 4$ , and  $k = 5$ ). Once the iteration reached convergence, CHI values were computed for each cluster configuration. The results in Table 6 indicate that the CHI value for  $k = 3$  is higher than for  $k = 2$ , but does not exceed the values for  $k = 4$  and  $k = 5$ . This suggests that increasing the number of clusters beyond  $k = 3$  enhances inter-cluster separation; however, it also increases model complexity, which may lead to overfitting. Therefore,  $k = 3$  was selected as the optimal number of clusters, as it provides a balance between good cluster separation and manageable complexity.

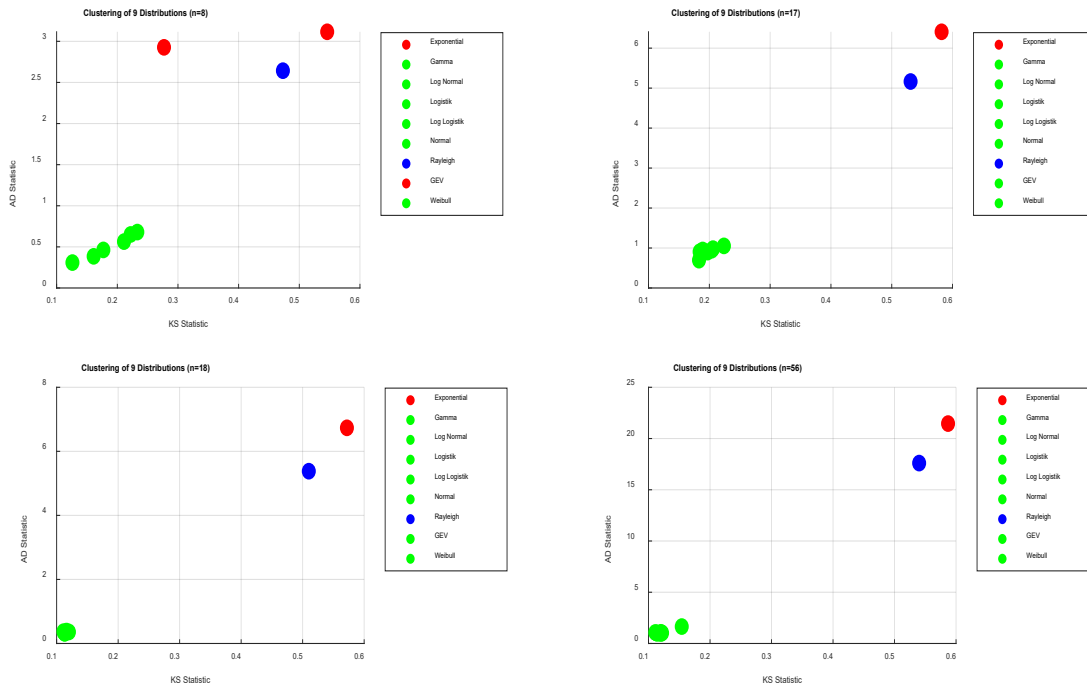
**Table 6.** Calinski-Harabasz Index (CHI) value

| Numbers of Clusters | CHI Value |          |          |          |
|---------------------|-----------|----------|----------|----------|
|                     | $n = 8$   | $n = 17$ | $n = 18$ | $n = 56$ |
| 2                   | 298.0     | 305.0    | 386.2    | 475.1    |
| 3                   | 202.1     | 1503.1   | 43710.4  | 47213.0  |
| 4                   | 260.9     | 3665.2   | 127330.0 | 83778.0  |
| 5                   | 566.8     | 7643.5   | 308780.0 | 184820.0 |

Note: CHI values for different cluster numbers across samples  $n = 8, 17, 18$ , and  $56$ . The cluster with the highest CHI value is considered the most optimal, with  $k = 3$  having the highest value across all sample sizes.

The clustering results indicate that Cluster 2 and Cluster 3 exhibit significant differences based on CHI values. Cluster 2 consists of the Exponential distribution, which consistently demonstrates poor goodness-of-fit to chlorophyll data based on both GoF tests and information criteria. Meanwhile, Cluster 3 consists of the Rayleigh distribution, which, despite exhibiting better fit patterns compared to the Exponential distribution, still presents distinct characteristics that differentiate it from the distributions in Cluster 1. The obtained CHI values confirm that Cluster 1 contains a more stable composition of distributions, whereas Cluster 2 and Cluster 3 consist of distributions with weaker fit to empirical data. The CHI evaluation further reinforces that  $k = 3$  is the optimal number of clusters. For  $k = 2$ , distributions with dissimilar characteristics remain grouped within the same cluster, reducing segmentation clarity. Conversely, for  $k = 4$  and  $k = 5$ , although inter-cluster separation increases, the additional clusters do not provide substantial benefits and merely increase model complexity without improving interpretability. Therefore,  $k = 3$  was selected as the optimal number of clusters, ensuring that the clustering structure remains representative of chlorophyll data patterns without unnecessarily increasing model complexity.

Figure 2 presents a visual representation of the clustering results for probability distributions based on KS and AD test statistics for various sample sizes ( $n = 8, 17, 18$ , and  $56$ ). Each point in the plot represents a specific distribution, where its position is determined by its KS value (x-axis) and AD value (y-axis). Meanwhile, the color of the points indicates clustering results based on the K-Means Clustering method. Distributions categorized into Cluster 1 (green)—such as the gamma, log-normal, logistic, log-logistic, normal, and Weibull—exhibit lower KS and AD values, indicating better goodness-of-fit to empirical data. Conversely, the Exponential (red) and Rayleigh (blue) distributions have higher KS and AD values, suggesting poorer fit to the data. The position of the GEV distribution (black) varies depending on sample size. Specifically, for  $n = 8$ , GEV is grouped in Cluster 2 alongside the Exponential distribution. However, for  $n = 17, 18$ , and  $56$ , GEV shifts to Cluster 1, indicating that its fit to the data improves as the sample size increases. These results support the conclusion that  $k = 3$  is the optimal number of clusters, as the distributions in Cluster 1 exhibit the best fit, while the Exponential and Rayleigh distributions remain in separate clusters, reflecting their distinct goodness-of-fit patterns.

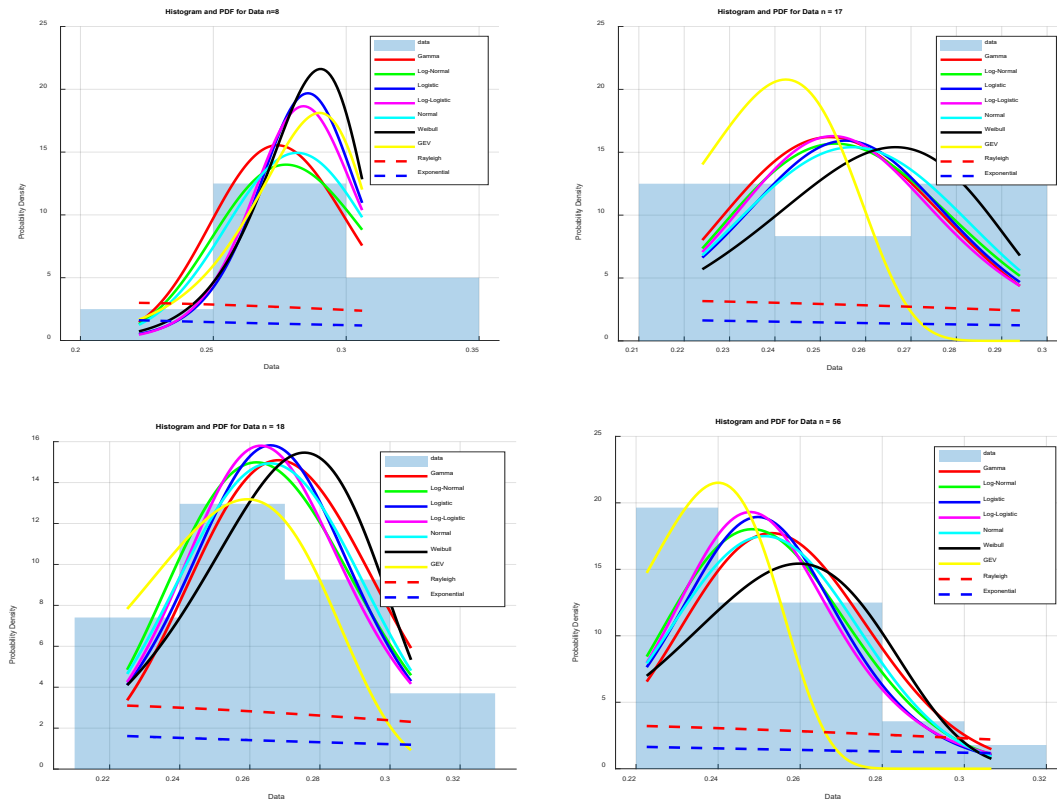


**Figure 2.** Clustering of distributions based on KS and AD for varying sample sizes

Figure 3 presents a comparison between fitted distributions and empirical data, using probability curves for each distribution evaluated against the histogram of actual data. Distributions in Cluster 1, such as the gamma, log-normal, logistic, log-logistic, normal, Weibull, and GEV, exhibit probability curves that better match the shape of the empirical data histogram, indicating their ability to capture data peaks and asymmetric distribution patterns. Meanwhile, the Exponential distribution (dashed red line) and the Rayleigh distribution (dashed blue line) show significant differences from the empirical data, as they fail to capture the main probability patterns in the data. The Exponential distribution appears overly simplified, while the Rayleigh distribution poorly represents the actual data distribution. Additionally, the GEV distribution, which initially resembled the Exponential distribution for small sample sizes, begins to show better fit for larger sample sizes, aligning with the clustering results where GEV shifted to Cluster 1. This visualization supports the findings from Figure 2 and the clustering analysis, where Cluster 1 contains the best-fitting distributions, whereas the Exponential and Rayleigh distributions primarily capture overall data complexity, which explains why they remain in a different cluster in the clustering analysis.

### 3.6 Discussion

The findings of this study highlight the critical role of selecting an appropriate probability distribution when modelling heterogeneous chlorophyll data. The gamma, log-normal, logistic, log-logistic, normal, and GEV distributions demonstrate superior performance in



**Figure 3.** Comparison of probability density functions for various distributions

terms of both GoF test results and their clustering behavior. These distributions consistently provide better adaptability across different sample sizes, making them more suitable for accurately representing the variability of chlorophyll concentration in marine environments. This finding is particularly relevant for marine and ecological research that relies on precise distribution models to understand complex oceanic dynamics.

Conversely, the Exponential and Rayleigh distributions are found to be less suitable for modelling chlorophyll data due to their limitations in capturing data heterogeneity. These distributions are primarily suited for data with simpler structures and do not adequately describe the variability observed in chlorophyll measurements. The selection of appropriate probability distribution is crucial to ensuring robust and reliable statistical analyses, which in turn enhances the accuracy of decision-making processes in marine ecosystem management and environmental monitoring. Furthermore, the clustering analysis based on GoF test statistics indicates that distributional behavior may vary depending on the sample size. For smaller datasets, GEV exhibits clustering patterns closer to the exponential distribution, whereas for larger sample sizes, GEV aligns with more flexible distributions such as gamma and log-normal. This suggests that sample size plays a significant role in determining the stability of distribution fitting, reinforcing the need for careful model selection in statistical applications related to oceanographic studies.

### 3.7 Limitations and recommendations

This study has several limitations that should be acknowledged. First, the sample size used in some cases is relatively small, which may impact the robustness and generalizability of the results, particularly for complex distributions such as GEV. Second, the GoF analysis is limited to the KS and AD tests, which primarily assess overall fit and sensitivity to tail behavior, respectively. These tests may not fully capture the goodness of fit for all probability distributions considered, especially for datasets with extreme values. Third, the clustering method employed (k-means) assumes a spherical cluster structure and does not effectively handle distributions with more complex relationships or overlapping characteristics.

For future research, it is recommended that a larger and more diverse sample set be used to improve the stability of parameter estimation and distribution fitting. Additionally, incorporating alternative GoF tests such as the Cramer–von Mises or Kuiper test could provide a more comprehensive assessment of distributional fit. To enhance clustering performance, more flexible clustering techniques, such as DBSCAN or Gaussian Mixture Models (GMM), should be considered, as they can better accommodate complex and non-spherical distributional structures. Lastly, integrating additional model selection criteria and machine learning-based clustering approaches may further refine the identification of optimal probability distributions for heterogeneous datasets.

## 4. Conclusions

This study evaluates univariate probability distribution models for chlorophyll data as an indicator of plankton presence in the ocean. Using GoF tests, such as the KS and AD tests, along with information criteria (AIC, BIC, AICc, CAIC, and HQC), this study assesses the suitability of various probability distributions, including the exponential, gamma, log-normal, logistic, log-logistic, normal, Rayleigh, GEV, and Weibull. The results indicate that the gamma, log-normal, logistic, log-logistic, normal, and GEV distributions exhibit a good fit to chlorophyll data across different sample sizes, with lower GoF test values and better information criteria compared to other distributions. In contrast, the Exponential and Rayleigh distributions are less suitable, as they fail to capture the heterogeneity of chlorophyll data, especially in larger sample sizes. The k-means clustering method was employed to group distributions based on KS and AD test statistics. The clustering results show that GEV in small samples ( $n = 8$ ) is closer to the exponential distribution, but for larger samples ( $n = 17, 18$ , and  $56$ ), GEV shifts to a more stable group alongside gamma, log-normal, logistic, log-logistic, normal, and Weibull. This suggests that GEV is more appropriate for larger sample sizes in modeling chlorophyll distribution.

The key conclusion of this study is that the gamma, log-normal, logistic, log-logistic, normal, and GEV distributions are more effective in modeling chlorophyll data, particularly for larger datasets. The clustering method based on GoF tests provides a systematic approach for selecting the most suitable probability distributions, reducing uncertainty in model selection. However, limitations in the clustering method (k-means) and the relatively small sample sizes indicate that further studies using more flexible approaches, such as DBSCAN or Gaussian Mixture Models (GMM), as well as additional tests like Cramér–von Mises, are necessary to better address the complexity of chlorophyll distribution.

## 5. Acknowledgements

The authors would like to extend their sincere appreciation to the Education Fund Management Agency (LPDP) of the Republic of Indonesia for their exceptional financial support in this research. The contribution from LPDP has not only enabled them to explore this topic in depth and publish our findings in internationally recognized journals, but has also made a significant impact in the field of education.

## 6. Authors' Contributions

Felix Reba: designed research, analyzed data, and wrote the paper; Toha Saifudin: coordinated research and reviewed the manuscript; Rimuljo Hendradi: contributed analytic tools and validated results.

### ORCID

Felix Reba  <https://orcid.org/0000-0002-5008-4770>

Toha Saifudin  <https://orcid.org/0000-0002-3745-0801>

Rimuljo Hendradi  <https://orcid.org/0000-0002-8679-0220>

## 7. Conflicts of Interest

We declare that there is no conflict of interest regarding the publication of this article.

## 8. AI Declaration

During the preparation of this manuscript, the author(s) used AI tools such as ChatGPT and ScholarGPT to assist with language editing, grammar improvement, and table formatting. These tools were not used to generate original scientific content or perform data analysis. All AI-assisted output has been critically reviewed, revised, and approved by the author(s) to ensure accuracy and scientific integrity.

## References

- Ashari, I. F., Dwi Nugroho, E., Baraku, R., Novri Yanda, I., & Liwardana, R. (2023). Analysis of elbow, silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-index evaluation on K-means algorithm for classifying flood-affected areas in Jakarta. *Journal of Applied Informatics and Computing*, 7(1), 95-103. <https://doi.org/10.30871/jaic.v7i1.4947>
- Badr, M. M. (2019). Goodness-of-fit tests for the Compound Rayleigh distribution with application to real data. *Heliyon*, 5(8), Article e02225. <https://doi.org/10.1016/j.heliyon.2019.e02225>
- Balakrishnan, N., Chimitova, E., & Vedernikova, M. (2015). An empirical analysis of some nonparametric goodness-of-fit tests for censored data. *Communications in Statistics: Simulation and Computation*, 44(4), 1101-1115. <https://doi.org/10.1080/03610918.2013.796982>
- Basheer, A. M. (2019). Alpha power inverse Weibull distribution with reliability application. *Journal of Taibah University for Science*, 13(1), 423-432. <https://doi.org/10.1080/16583655.2019.1588488>

- Behrenfeld, M. J., & Boss, E. S. (2014). Resurrecting the ecological underpinnings of ocean plankton blooms. *Annual Review of Marine Science*, 6(1), 167-194. <https://doi.org/10.1146/annurev-marine-052913-021325>
- Bishop, C. H. (2016). The GIGG-EnKF: Ensemble Kalman filtering for highly skewed non-negative uncertainty distributions. *Quarterly Journal of the Royal Meteorological Society*, 142, 1395-1412. <https://doi.org/10.1002/qj.2742>
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2<sup>nd</sup> ed.). Springer.
- Campbell, J. W. (1995). The lognormal distribution as a model for bio-optical variability in the sea. *Journal of Geophysical Research*, 100(C7), 13237-13254. <https://doi.org/10.1029/95jc00458>
- de Souza, A., Aristone, F., Fernandes, W. A., Oliveira, A. P. G., Olaofe, Z., Abreu, M. C., Junior, J. F. de O., Cavazzana, G., dos Santos, C. M., & Pobocikova, I. (2020). Analysis of ozone concentrations using probability distributions. *Ozone: Science and Engineering*, 42(6), 539-550. <https://doi.org/10.1080/01919512.2020.1736987>
- Devred, E., Sathyendranath, S., Stuart, V., & Platt, T. (2011). A three-component classification of phytoplankton absorption spectra: Application to ocean-color data. *Remote Sensing of Environment*, 115(9), 2255-2266. <https://doi.org/10.1016/j.rse.2011.04.025>
- Ding, J., Tarokh, V., & Yang, Y. (2018). Model selection techniques: An overview. *IEEE Signal Processing Magazine*, 35(6), 16-34. <https://doi.org/10.1109/MSP.2018.2867638>
- Dutkiewicz, S., Hickman, A. E., Jahn, O., Gregg, W. W., Mouw, C. B., & Follows, M. J. (2015). Capturing optically important constituents and properties in a marine biogeochemical and ecosystem model. *Biogeosciences*, 12(14), 4447-4481. <https://doi.org/10.5194/bg-12-4447-2015>
- Goual, H., & Yousof, H. M. (2020). Validation of Burr XII inverse Rayleigh model via a modified chi-squared goodness-of-fit test. *Journal of Applied Statistics*, 47(3), 393-423. <https://doi.org/10.1080/02664763.2019.1639642>
- Guedes, K. S., de Andrade, C. F., Rocha, P. A. C., Manguiera, R. dos S., & de Moura, E. P. (2020). Performance analysis of metaheuristic optimization algorithms in estimating the parameters of several wind speed distributions. *Applied Energy*, 268, Article 114952. <https://doi.org/10.1016/j.apenergy.2020.114952>
- Gulati, S., & Shapiro, S. (2009). A new goodness of fit test for the logistic distribution. *Journal of Statistical Theory and Practice*, 3(3), 567-576. <https://doi.org/10.1080/15598608.2009.10411947>
- Haggag, M. M. M. (2014). New criteria of model selection and model averaging in linear regression models. *American Journal of Theoretical and Applied Statistics*, 3(5), 148-166. <https://doi.org/10.11648/j.ajtas.20140305.15>
- Indrayani, E., Dimara, L., Paiki, K., & Reba, F. (2018). The analysis of phytoplankton abundance using Weibull distribution (a case study in the coastal area of East Yapan in the regency of Yapen Islands, Papua). *Journal of Education and Learning*, 7(3), 251-258. <https://doi.org/10.5539/jel.v7n3p251>
- Irwin, A. J., Finkel, Z. V., Schofield, O. M. E., & Falkowski, P. G. (2006). Scaling-up from nutrient physiology to the size-structure of phytoplankton communities. *Journal of Plankton Research*, 28(5), 459-471. <https://doi.org/10.1093/plankt/fbi148>
- Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651-666. <https://doi.org/10.1016/j.patrec.2009.09.011>
- Jiang, H., Wang, J., Wu, J., & Geng, W. (2017). Comparison of numerical methods and metaheuristic optimization algorithms for estimating parameters for wind energy

- potential assessment in low wind regions. *Renewable and Sustainable Energy Reviews*, 69, 1199-1217. <https://doi.org/10.1016/j.rser.2016.11.241>
- Jung, C., & Schindler, D. (2017). Global comparison of the goodness-of-fit of wind speed distributions. *Energy Conversion and Management*, 133, 216-234. <https://doi.org/10.1016/j.enconman.2016.12.006>
- Kumar, V., Shanu, S., & Jahangeer, J. (2017). Statistical distribution of rainfall in Uttarakhand, India. *Applied Water Science*, 7(8), 4765-4776. <https://doi.org/10.1007/s13201-017-0586-5>
- Lima, A. O., Lyra, G. B., Abreu, M. C., Oliveira-Júnior, J. F., Zeri, M., & Cunha-Zeri, G. (2021). Extreme rainfall events over Rio de Janeiro State, Brazil: Characterization using probability distribution functions and clustering analysis. *Atmospheric Research*, 247, Article 105221. <https://doi.org/10.1016/j.atmosres.2020.105221>
- Lombard, F., Boss, E., Waite, A. M., Vogt, M., Uitz, J., Stemann, L., Sosik, H. M., Schulz, J., Romagnan, J.-B., Picheral, M., Pearlman, J., Ohman, M. D., Niehoff, B., Möller, K. O., Miloslavich, P., Lara-Lopez, A., Kudela, R. M., Lopes, R. M., Kiko, R., Karp-Boss, L., Jaffe, J. S.,... Appeltans, W. (2019). Globally consistent quantitative observations of planktonic ecosystems. *Frontiers in Marine Science*, 6, Article 196. <https://doi.org/10.3389/fmars.2019.00196>
- Mansoor, M., Tahir, M. H., Cordeiro, G. M., Provost, S. B., & Alzaatreh, A. (2019). The Marshall-Olkin logistic-exponential distribution. *Communications in Statistics - Theory and Methods*, 48(2), 220-234. <https://doi.org/10.1080/03610926.2017.1414254>
- Mbah, A. K., & Paothong, A. (2015). Shapiro–Francia test compared to other normality test using expected p-value. *Journal of Statistical Computation and Simulation*, 85(15), 3002-3016. <https://doi.org/10.1080/00949655.2014.947986>
- Mohammadi, K., Alavi, O., & McGowan, J. G. (2017). Use of Birnbaum-Saunders distribution for estimating wind speed and wind power probability distributions: A review. *Energy Conversion and Management*, 143, 109-122. <https://doi.org/10.1016/j.enconman.2017.03.083>
- Ouarda, T. B. M. J., Charron, C., & Chebana, F. (2016). Review of criteria for the selection of probability distributions for wind speed data and introduction of the moment and L-moment ratio diagram methods, with a case study. *Energy Conversion and Management*, 124, 247-265. <https://doi.org/10.1016/j.enconman.2016.07.012>
- Ozay, C., & Celik, M. S. (2016). Statistical analysis of wind speed using two-parameter Weibull distribution in Alaçati region. *Energy Conversion and Management*, 121, 49-54. <https://doi.org/10.1016/j.enconman.2016.05.026>
- Pobočíková, I., Sedláčková, Z., & Michalková, M. (2017). Application of Four Probability Distributions for Wind Speed Modeling. *Procedia Engineering*, 192, 713-718. <https://doi.org/10.1016/j.proeng.2017.06.123>
- Rajakaruna, H., VandenByllaardt, J., Kydd, J., & Bailey, S. (2018). Modeling the distribution of colonial species to improve estimation of plankton concentration in ballast water. *Journal of Sea Research*, 133, 166-176. <https://doi.org/10.1016/j.seares.2017.08.005>
- Siegel, D. A., Doney, S. C., & Yoder, J. A. (2002). The North Atlantic spring phytoplankton bloom and Sverdrup's critical depth hypothesis. *Science*, 296(5568), 730-733. <https://doi.org/10.1126/science.1069174>
- ul Haq, M. A., Rao, G. S., Albassam, M., & Aslam, M. (2020). Marshall–Olkin Power Lomax distribution for modeling of wind speed data. *Energy Reports*, 6, 1118-1123. <https://doi.org/10.1016/j.egyr.2020.04.033>
- Van, T. V., & Pham-Gia, T. (2010). Clustering probability distributions. *Journal of Applied Statistics*, 37(11), 1891-1910. <https://doi.org/10.1080/02664760903186049>
- Wang, J., Hu, J., & Ma, K. (2016). Wind speed probability distribution estimation and wind

- energy assessment. *Renewable and Sustainable Energy Reviews*, 60, 881-899. <https://doi.org/10.1016/j.rser.2016.01.057>
- Wang, X., & Xu, Y. (2019). An improved index for clustering validation based on silhouette index and Calinski-Harabasz index. *IOP Conference Series: Materials Science and Engineering*, 569(5), Article 052024. <https://doi.org/10.1088/1757-899X/569/5/052024>
- Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44(1), 92-107. <https://doi.org/10.1006/jmps.1999.1278>
- Yazici, B., & Yolacan, S. (2007). A comparison of various tests of normality. *Journal of Statistical Computation and Simulation*, 77(2), 175-183. <https://doi.org/10.1080/10629360600678310>
- Zheng, S., Fan, K., Hou, Y., Feng, J., & Fu, Y. (2023). Clustering by the probability distributions from extreme value theory. *IEEE Transactions on Artificial Intelligence*, 4(2), 292-303. <https://doi.org/10.1109/TAI.2022.3153592>