JPPIPA 12(4) (2026)

Jurnal Penelitian Pendidikan IPA

http://jppipa.unram.ac.id/index.php/jppipa/index

# Bayesian Tobit Model for Handling Left-Censored in Water Quality Assessment: A Case Study of Sunda Strait, Indonesia

Iis Afrianti[1*], Sjaifuddin Sjaifuddin[1], Najmi Firdaus[1]

[1] Department of Master of Environmental Studies, Sultan Ageng Tirtayasa University, Indonesia.

Received: November 11, 2025
Revised: March 30, 2026
Accepted: April 25, 2026
Published: April 30, 2026

Corresponding Author:
Iis Afrianti
iis.afrianti09@gmail.com

DOI: 10.29303/jppipa.v12i4.13432



**Abstract:** As an archipelagic nation, Indonesia relies heavily on coastal activities that may affect marine water quality, yet studies addressing this issue remain limited, particularly in the presence of left-censored data. This study aims to evaluate an appropriate method for handling left-censored data in water quality assessment using the CCME-WQI, with a case study in the Sunda Strait. A Bayesian Tobit model was applied to account for left-censored observations and integrated with the CCME-WQI framework. For comparison, conventional substitution methods and exclude left-censored were also used. The performance of approaches was assessed based on their ability to produce reliable water quality index estimates. The results indicate that the Bayesian Tobit model provides more robust estimates than substitution methods, as it incorporates uncertainty through credible intervals and reduces potential bias. The estimated water quality index ranged from 83.8 to 92.1, classifying the water quality as "good." In conclusion, the Bayesian Tobit model is a more reliable approach for handling left-censored environmental data and improving water quality assessment. This method is particularly relevant for routine monitoring and can be extended to other fields with similar data characteristics.

**Keywords:** Bayesian-Tobit Model; CCME-WQI; Left-Censored; Sunda Strait; Water Quality Assessment

## Introduction

The quality of seawater is important both marine life and humans living on land. The quality of seawater can have a detrimental effect on human health through bioaccumulation in the consumption of contaminated marine food (Haeruddin et al., 2021; Tolkou et al., 2023). Seawater quality can be characterized by an index based on quality standard, such as National sanitation Foundation Water Quality Index (NSF-WQI), Canadian Council of Ministers of the Environment Water Quality Index (CCME-WQI), Malaysian Marine Water Quality Index (MMWQI) and another modified index (Al-Qadami et al., 2025; Ma et al., 2020; Ramazanova et al., 2022; Ristanto et al., 2021). One particular method that was frequently used was CCME-WQI, due to its simplicity and flexibility in selecting the water quality parameters that could be employed in the equation (Uddin et al., 2021).

Since its endorsement in 2001, the CCME WQI has been used for reporting on the state of water quality in Canada and other country. The CCME-WQI offers a convenient means of summarizing complicated water quality data and making it easier for audience. The CCME-WQI's specific guidelines, parameters, and time period are not defined, and they may vary from region to region based on local conditions, the purpose of the index's use, and problems with water quality. This index is calculated using metrics that surpass standards or goals and can be applied to river, lake, and marine water (Canadian Council of Ministers of the Environment, 2017; Uddin et al., 2021).

The Sunda Strait is located between Java and Sumatra Island. This strait connects the Java Sea in the north with the Indian Ocean in the south. With an emphasis on water mass characteristics, the Sunda Strait is predominantly influenced by Indian Ocean water masses (Fahlevi et al., 2022). Coral reefs, seagrass meadows and mangrove ecosystems are all part of the rich marine ecology found in this area (Munandar et al., 2022; Nugroho et al., 2024). The Sunda Strait, a region of significant ecological importance, is a well-known

**How to Cite:**
Afrianti, I., Sjaifuddin, S., & Firdaus, N. (2026). Bayesian Tobit Model for Handling Left-Censored in Water Quality Assessment: A Case Study of Sunda Strait, Indonesia. *Jurnal Penelitian Pendidikan IPA*, *12*(4), 176–187. https://doi.org/10.29303/jppipa.v12i4.13432

fishing area for local fishermen. It has been reported that there are an overexploitation and disease found in marine biota that are living in this region (Kartini et al., 2017; Subekti et al., 2021). Furthermore, the volcanic activity of Anak Krakatau, which is located in Sunda Strait, has the potential to significantly impact the surrounding marine and coastal ecosystems (Chasanah et al., 2020; Lestari et al., 2020).

The Sunda Strait also plays as a crucial maritime corridor, functioning as one of the strategic international shipping routes frequented by cargo vessels, with the Merak port as a key transit point. This area holds significant importance for Indonesia's maritime security. In its capacity as a transportation route, it is essential that the strait is protected against threats such as foreign military actions, piracy and smuggling. In consequence, the Traffic Separation Scheme (TSS) model may be applicable in the Sunda Strait, in order to ensure territorial resilience (Sobaruddin et al., 2017).

The ecological quality of the Sunda Strait, particularly the quality of its seawater, can be impacted by a numerous of activities, including natural processes, industrial operations, shipping and anthropogenic sources (Elgendy et al., 2024; Liu et al., 2021). The industrial presence along the coastline has the potential to contaminate marine ecosystem with pollutants. Consequently, it is essential to conduct regular assessments of the water quality in the Sunda Strait to conserve both human and marine ecosystems.

A number of studies have been conducted in Sunda Strait (Iqbal et al., 2023; Li et al., 2018; Susanto et al., 2023; Yonvitner et al., 2020), but the literature on the quality of the seawater in this area is limited. The environmental data, which is used in the CCME-WQI calculation, frequently contains numerical values that are below the limit of detection (LOD), and referred as left-censored or non-detects (P. C. Bürkner, 2017; Wood et al., 2011). Before censored data is used in the CCME-WQI computation, it must be managed. Five popular approaches for handling non-detects are substitute with zero, substitute with half of LOD, substitute with LOD, Bayesian model, and exclude LOD value. Therefore, we conducted this study to determine the CCME-WQI at the Sunda Strait with various comparisons of non-detect handling methods, so we can address these knowledge gaps and propose a Bayesian-Tobit approach to estimate left-censored value for determining the CCME-WQI in the study area.

## Method

The initial step in this investigation was to collect data and identify left-censored data. Following that, data was processed using replacement methods, the Bayesian-Tobit model, and removing left-censored data to determine the CCME-WQI and the state of seawater quality, which were then compared. To put it simply, Figure 1**Error! Reference source not found.** illustrates the flow chart of this study.

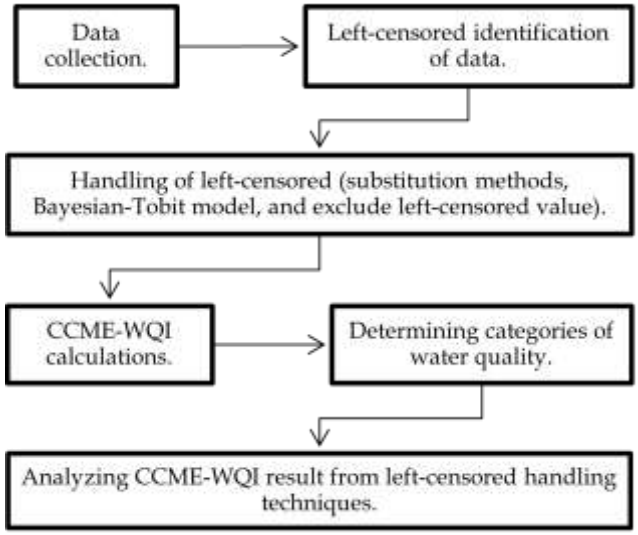



**Figure 1.** Flow chart of the study.

*Study Area and Data Collections***.**

This study was conducted in the northern part of Sunda Strait, boarded by the Cilegon city. Study map Figure 2 and geographical information Table 1 are provided. The study was conducted at six points located in the area that lies from 5°52′23″ S to 5°54′39″ S latitude and 106°00′01″ E to 106°02′26″ E longitude. The water temperature ranges during the sampling were 27.2 - 33.4 °Celsius and the average total precipitation were 77.5 – 114.2 mm/days (Giovanni, 2025). Data were obtained from environmental monitoring reports at SIMPEL (Kementerian Lingkungan Hidup, 2025), with frequency twice a year from 2021 to 2024. After the data obtained, left-censored identified as value below detection limit which mark with '<'.

*The CCME-WQI Calculations*

The CCME-WQI was computed using method from Canadian Council of Ministers of the Environment (Canadian Council of Ministers of the Environment, 2017). To calculate the CCME-WQI, twenty-seven parameters were selected with the standard used in this study was Indonesia Marine Standard as we could see at **Error! Reference source not found.**. The calculation of CCME-WQI is based on the parameters that exceeded the standard or objectives as follows with equations.

$$WQI = 100 - \left[\frac{\sqrt{F_1^2+F_2^2+F_3^2}}{1.732}\right] \quad (1)$$

$$F_1 = \left[\frac{number\ of\ failed\ parameters}{total\ number\ of\ parameters}\right] \times 100 \quad (2)$$

$$F_2 = \left[\frac{number\ of\ failed\ tests}{total\ number\ of\ tests}\right] \times 100 \quad (3)$$

$$F_3 = \left[\frac{nse}{0.01(nse)+0.01}\right] \quad (4)$$

If a value falls below the objective,

$$excursion_i = \left[\frac{failed\ value_i}{objective_j}\right] - 1 \quad (5)$$

If a value exceeds the objective,

$$excursion_i = \left[\frac{objective_j}{failed\ value_i}\right] - 1 \quad (6)$$

$$nse = \left[\frac{\sum_{i=1}^{n} excursion_j}{total\ number\ of\ test}\right] - 1 \quad (7)$$

Due to the inability to determine the precise value, left-censoring poses a challenge in identifying failure objectives. Therefore, left-censored value that exceeded the objective was estimated using the Bayesian-Tobit model and replaced with zero, one-half of the detection limit, and the limit of detection (LOD). For each parameter, Table 2 shows the quantity and percentage of left-censored.

After substituting and estimating the left-censored, the CCME-WQI was calculated for each site. Then, there are five categories for the water quality of the samples: excellent (95-100), good (80-94), fair (65-79), marginal (45-64) or poor (0-44). In addition to utilizing the estimation results from these methodologies, CCME-WQI measurements were also obtained, excluding parameters with left-censored values that exceeded the objective, for comparative analysis.

*Handling Left-Censored*

The most conventional approach in environmental chemistry to deal with non-detects is to substitute some fraction of the detection limit. This method is more commonly referred to as 'fabrication'. In water chemistry, the most typical fraction used is one-half (P. C. Bürkner, 2017). Left-censored values are replaced with zero for comparisons concerning underestimated results, and with the detection limit for overestimated results. The CCME-WQI method is then carried out to calculate the water quality index value by comparing the substitution values with the standard.

The Bayesian-Tobit model was run using RStudio version 4.5.1 with brms package version 2.22.0. The brms (Bayesian Regression Models using 'Stan') package is the implementation of Bayesian generalized (non-)linear multivariate multilevel models using the 'Stan' program that is essential for conducting comprehensive Bayesian inference (P. C. Bürkner, 2017).

The Bayesian-Tobit model is a censored regression model that is estimated in a Bayesian framework. Prior distributions are assigned to the parameters, and inference is based on the posterior distributions obtained using Bayes' theorem (Nurfajriah et al., 2024). The use of a Bayesian Tobit model is for analyzing data with values falling below or above a detection limit, by treating these observations as censored and incorporating prior knowledge into the parameter estimation via Markov Chain Monte Carlo methods (Dagne & Huang, 2022).

To operate the Bayesian Tobit model in RStudio, first mark left-censored data with the cens code on the script.

```
data_raw <- read_excel (location of the file that contains data)
data_all <- data_raw %>% mutate
    (cens = ifelse (str_detect (Data, "<"), "left", "none"),
    y = as.numeric (str_replace (Data, "<", "")),
    LOD = ifelse (cens == "left", y, NA))
```

The priors were determined using a weakly informative prior with partially pooling hierarchical structure from sites and time as a varying effect. We used lognormal and gamma distributions.

```
Median_LOD <- median (data_all$LOD, na.rm = TRUE)
prior_lognorm <- c(prior(normal(log(median_LOD), 2),
    class = "Intercept"),
    prior (student_t (3, 0, 1), class = "sd"),
    prior (student_t (3, 0, 1), class = "sigma"))
prior_gamma <- c(prior(normal(log(median_LOD), 1),
    class = "Intercept"),
    prior (student_t (3, 0, 1), class = "sd"),
    prior (gamma (2, 0.5), class = "shape"))
```

**Table 1.** Location Site Information

| Sites | Location | Latitude | Longitude |
|---|---|---|---|
| A | Coal-fired power plant, Cinangka-Cipeteuy river basin | 5°52'23"S | 106°02'26"E |
| B | Coal-fired power plant, Cipeteuy river basin | 5°52'59"S | 106°02'00"E |
| C | Coal-fired power plant, Cipeteuy river basin | 5°52'51"S | 106°01'14"E |
| D | Bulk liquid storage terminal, Cipeteuy river basin | 5°54'36"S | 106°00'14"E |
| E | Bulk liquid storage terminal, Cipeteuy river basin | 5°54'39"S | 106°00'01"E |
| F | Bulk liquid storage terminal, Cipeteuy river basin | 5°54'38"S | 106°00'04"E |

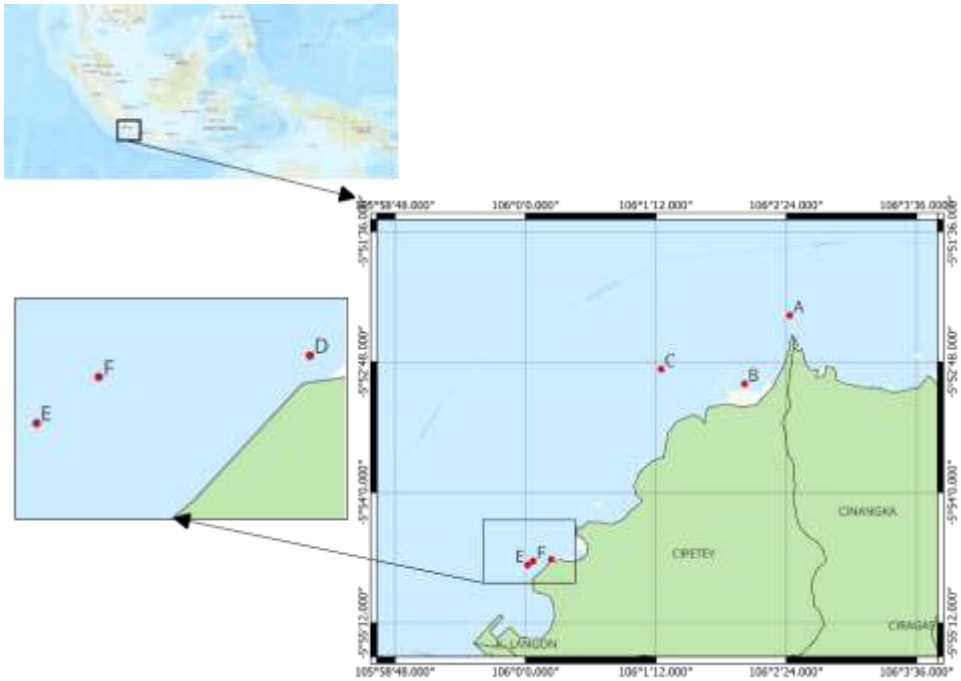



**Figure 2.** Map of the study area

**Table 2.** Dataset description

| Parameter | Standard[a] (unit) | Percentage of Left-Censored | Median of Limit Detection | Range of Limit Detection | Total of Observed exceed standard | Total of Left-Censored exceed standard |
|---|---|---|---|---|---|---|
| Transparency | >5 (m) | 0 | - | - | 37 | 0 |
| Turbidity | 5 (NTU) | 0 | - | - | 1 | 0 |
| Suspended Solid | 20 (mg/L) | 12.50 | 3 | 1 - 3 | 9 | 0 |
| Temperature | 28-30 (°C) | 0 | - | - | 4 | 0 |
| pH | 7 - 8.5 | 0 | - | - | 0 | 0 |
| Salinity | 33-34 (‰) | 0 | - | - | 24 | 0 |
| DO | >5 (mg/L) | 0 | - | - | 7 | 0 |
| BOD | 20 (mg/L) | 18.75 | 2 | 2 | 0 | 0 |
| Total Ammonia | 0.3 (mg/L) | 50 | 0.016 | 0.01 - 0.035 | 0 | 0 |
| Orthophosphate | 0.015 (mg/L) | 83.33 | 0.006 | 0.003 - 0.03 | 3 | 5 |
| Nitrates | 0.06 (mg/L) | 4.17 | 0.005 | 0.005 | 0 | 0 |
| Cyanide | 0.5 (mg/L) | 75 | 0.003 | 0.002 - 0.0995 | 0 | 0 |
| Sulfide | 0.01 (mg/L) | 77.08 | 0.0022 | 0.00022 - 0.01 | 0 | 0 |
| Total Phenol | 0.002 (mg/L) | 100 | 0.0008 | 0.0002 - 0.002 | 0 | 0 |
| PAH | 0.003 (mg/L) | 0 | 0.0002 | 0.000006 - 0.0.001 | 0 | 0 |
| PCB | 0.01 (µg/L) | 95.83 | 0.001 | 0.00001 - 0.01 | 0 | 0 |
| MBAS | 1 (mg/L) | 79.17 | 0.01 | 0.008 - 0.029 | 0 | 0 |
| Oil and grease | 1 (mg/L) | 47.92 | 0.3 | 0.3 - 1 | 1 | 0 |
| TBT | 0.01 (µg/L) | 100 | 0.004 | 0.00001 - 0.01 | 0 | 0 |
| Hg | 0.001 (mg/L) | 83.33 | 0.0005 | 0.00005 - 0.001 | 0 | 0 |
| As | 0.012 (mg/L) | 58.33 | 0.0064 | 0.002 - 0.0064 | 0 | 0 |
| Cd | 0.001 (mg/L) | 100 | 0.0005 | 0.0001 - 0.001 | 0 | 0 |
| Pb | 0.008 (mg/L) | 62.5 | 0.0033 | 0.0002 - 0.008 | 0 | 0 |
| Cu | 0.008 (mg/L) | 81.25 | 0.002 | 0.0003 - 0.007 | 0 | 0 |
| Zn | 0.05 (mg/L) | 62.5 | 0.0048 | 0.0003 - 0.005 | 1 | 0 |
| Ni | 0.05 (mg/L) | 68.75 | 0.0059 | 0.0008 - 0.009 | 0 | 0 |
| Total Coliform | 1000 (/100 mL) | 25 | 1.8 | 1.8 - 2 | 0 | 0 |

[a]Source: (Penyelenggaraan Perlindungan Dan Pengeolaan Lingkungan Hidup, 2021)

The brms package's default_prior function was also executed for comparison.

```
Default_prior_lognorm <- default_prior (cens ~ 1 + Semester + (1 | Site) + (1 | Year), family = lognormal(),
    data = data_all)
Default_prior_gamma <- default_prior(cens ~ 1 + Semester + (1 | Site) + (1 | Year), family = Gamma(),
    data = data_all)
```

To fit the Bayesian Tobit model, we use iter 4000 with warm up 1000 and 4 chains of MCMC. Default adapt delta is 0.99 and max treedepth 15. However, if there is divergent transition, we use higher adapt delta until there is no divergent transition. The script of fit the model used,

```
brm_lognorm <- brm (bf (y | cens(cens) ~ 1 + Semester + (1 | Site) + (1 | Year)),
    data = data_all,
    family = lognormal(),
    prior = prior_lognorm,
    iter = 4000, warmup = 1000,
    chains = 4, cores = parallel::detectCores(),
    save_pars = save_pars (all = TRUE),
    control = list (adapt_delta = 0.99,
    max_treedepth = 15), seed = 123)
brm_gamma <- brm (bf (y | cens(cens) ~ 1 + Semester + (1 | Site) + (1 | Year)),
    data = data_all,
    family = Gamma (link = "log"),
    prior = prior_gamma,
    iter = 4000, warmup = 1000,
    chains = 4, cores = parallel::detectCores(),
    save_pars = save_pars(all = TRUE),
    control = list (adapt_delta = 0.99,
    max_treedepth = 15), seed = 123)
```

After running fit Bayesian Tobit model, the rhat and neff tests were conducted to determine the convergence of the model. Moreover, we do posterior predictive check to see from pp_check graph whether the model fits the distribution of the data and then we conduct LOO (Leave-One-Out cross-validation) test to determine lognormal or gamma distribution more fit with the data.

```
Summary(brm_lognorm)
summary(brm_gamma)
plot(brm_lognorm)
plot(brm_gamma)
pp_check(brm_lognorm)
pp_check(brm_gamma)
loo_lognorm <- loo(brm_lognorm)
loo_gamma <- loo(brm_gamma)
loo_compare (loo_lognorm, loo_gamma)
```

The estimation of predictive median with lower and upper of 95 % credible intervals were applied in the model.

```
y_lognorm <- posterior_predict(brm_lognorm)
y_medlognorm <- apply (y_lognorm, 2, median)
y_cilognorm <- apply (y_lognorm, 2, quantile, probs = c (0.025, 0.975))
y_gamma <- posterior_predict(brm_gamma)
y_medgamma <- apply (y_gamma, 2, median)
y_cigamma <- apply (y_gamma, 2, quantile, probs = c (0.025, 0.975))
imput_lognorm <- data_all %>% mutate (y_imput_median = y_medlognorm,
    y_CI_lower = y_cilognorm [1,],
    y_CI_upper = y_cilognorm [2,])
print(imput_lognorm)
imput_gamma <- data_all %>% mutate (y_imput_median = y_medgamma,
    y_CI_lower = y_cigamma [1,],
    y_CI_upper = y_cigamma [2,])
print(imput_gamma)
```

After obtaining the posterior predictions, it is also necessary to know the percentage of values below and above the detection limit to assess the consistency of the modeling results.

```
id_cens <- which (data_all$cens == "left")
perc_below_lognorm <- sapply(seq_along(id_cens), function(j) {mean (y_lognorm [, id_cens[j]] <= data_all$LOD[id_cens[j]], na.rm = TRUE)})
perc_lognorm <- data_table %>% filter (cens == "left") %>% mutate (perc_below_LOD = perc_below_lognorm) %>% summarise (mean_perc_below_lognorm = mean (perc_below_lognorm, na.rm = TRUE))
print(perc_lognorm)
perc_below_gamma <- sapply(seq_along(id_cens), function(j) {mean (y_gamma [, id_cens[j]] <= data_all$LOD [id_cens[j]], na.rm = TRUE)})
perc_gamma <- data_table %>% filter (cens == "left") %>% mutate (perc_below_LOD = perc_below_gamma) %>% summarise ( mean_perc_below_gamma = mean (perc_below_gamma, na.rm = TRUE))
print(perc_gamma)
```

After estimating values from left-censored data exceeding the standard, CCME-WQI was calculated and the water quality category for the Sunda Strait was determined for the period 2021 to 2024. As comparison, CCME-WQI also computed without left-censored data.

## Result and Discussion

*Bayesian-Tobit Modeling and Model Evaluation*

In this study, CCME-WQI was determined using twenty-seven parameters listed in the Indonesian government's marine water quality standards. Table 2 presents a description of the parameter data on the study site. Of the twenty-seven parameters, nine parameters have observed data (the data value that above the

detection limit) exceeding the standard, while one parameter, orthophosphate, have left-censored data exceeding the standard. This parameter will then be estimated using the Bayesian-Tobit model.

The Bayesian-Tobit model is a statistical modelling approach that combines prior knowledge about parameters with censored regression data (Tobit model) to generate a posterior distribution with credible intervals after viewing the data (P. Bürkner, 2018; Dagne & Huang, 2022; Huynh et al., 2014). Running the Bayesian-Tobit model depends on determining the prior. In Bayesian inference, a prior is used to quantify a researcher's conviction about particular hypotheses (such as parameter values) prior to observing any evidence. Typically depicted as a probability distribution across various levels of belief (Ravenzwaaij et al., 2018). The prior used in this study was run on two types of distributions, namely lognormal and gamma. The lognormal and gamma distributions were chosen for this model because the data on marine water chemical concentration were positive and right-skewed, making these two distributions the most suitable.

Prior plays an important role in generating posterior draws that represent estimates of the data distribution (Gelman et al., 2017; Safford et al., 2022). Recent study revealed that for meta-analyses with random effects, it is recommended to use Bayesian methods by employing a weakly informative prior distribution for the heterogeneity parameter, especially when there are only a few studies (Röver et al., 2021). Weakly informative priors provide a prior that is broad enough but not excessively restricted, allow it to be used without particular knowledge. It generates more stable and realistic estimates than maximum likelihood estimation or the non-informative prior. To improve the robustness of parameter estimation and reduce its sensitivity to extreme patterns and data noise, partial pooling was also applied (Gelman et al., 2017). Therefore, in this study, the Bayesian-Tobit model uses partial pooling with semester data as a fixed effect (b) plus year and site data as varying effects (sd). Semester is used as a model coefficient to reflect the difference between the first and second semesters, while year and site are used as varying effects to reflect the effects of year and site on the model.

The prior family lognormal uses normal(log(median_LOD), 2) at the "intercept" with mean = log(median_LOD) and standard deviation = 2, reflecting prior knowledge that the median left-censored (on the log scale) is approximately −5. Prior student_t (3, 0, 1) is used for "sd" (prior for variation between sites and years) and "sigma" (prior for residual variation) with degrees of freedom (df) = 3, mean = 0, and scale = 1. The use of df = 3 is done so that the distribution has heavier tails than normal and makes the prior more "robust" against outliers, while mean = 0 because the data values are close to zero and scale = 1 so that the prior allows for variation of up to 3 on the log scale (Gelman et al., 2017; McElreath, 2020). The default prior in the brms package gives the same "intercept", "sd", and "sigma" prior, the family student_t(3, 0, 2.5). The default prior also suggests df = 3, which is heavy-tailed for "robust" results. The mean is the same at 0, but the scale is larger with a wider variation up to 18.75 on the log scale, thus suggesting a weakly informative prior.

The prior family gamma uses normal(log(median_LOD), 1) at the "intercept" with mean = log(median_LOD) and standard deviation = 1, reflecting prior knowledge that the median left-censored (on the log scale) is approximately −5. The prior student_t (3, 0, 1) is still used for "sd" and "shape" (prior for residual variation in gamma) uses gamma (2, 0.5) with shape = 2 and rate = 0.5 to ensure positive and moderate variance = 8. The default prior in the brms package produces the same "intercept" and "sd" priors as the default lognormal prior family, the Student_t (3, 0, 2.5), but with a different "shape" prior that gives the default gamma (0.01, 0.01) prior, which has quite a large variance = 100. The use of normal-gamma and hierarchically weakly informative priors as default regression priors is also consistent with ecological research which recommended this prior (Lemoine, 2019).

The Bayesian-Tobit model was run using the probabilistic programming language 'Stan', with sampling from the posterior distribution using the Markov Chain Monte Carlo (MCMC) method (No-U-Turn Sampler (NUTS) version). The MCMC iteration was performed 4000 times, with 1000 warm-ups and 4 chains. The control parameter adapt_delta, with a target acceptance rate of 0.99 or higher, was used to ensure more careful and accurate sampling, thereby preventing divergent transitions (P. C. Bürkner, 2017). The results of the summary fit of the Bayesian-Tobit model on orthophosphate parameter to the lognormal distribution with weakly informative prior ($\text{lognormal}_{WIP}$), the lognormal distribution with default prior ($\text{lognormal}_{DP}$), the gamma distribution with weakly informative prior ($\text{gamma}_{WIP}$) and the gamma distribution with default prior ($\text{gamma}_{DP}$) can be seen in

Table ***3*** and Fig 3.

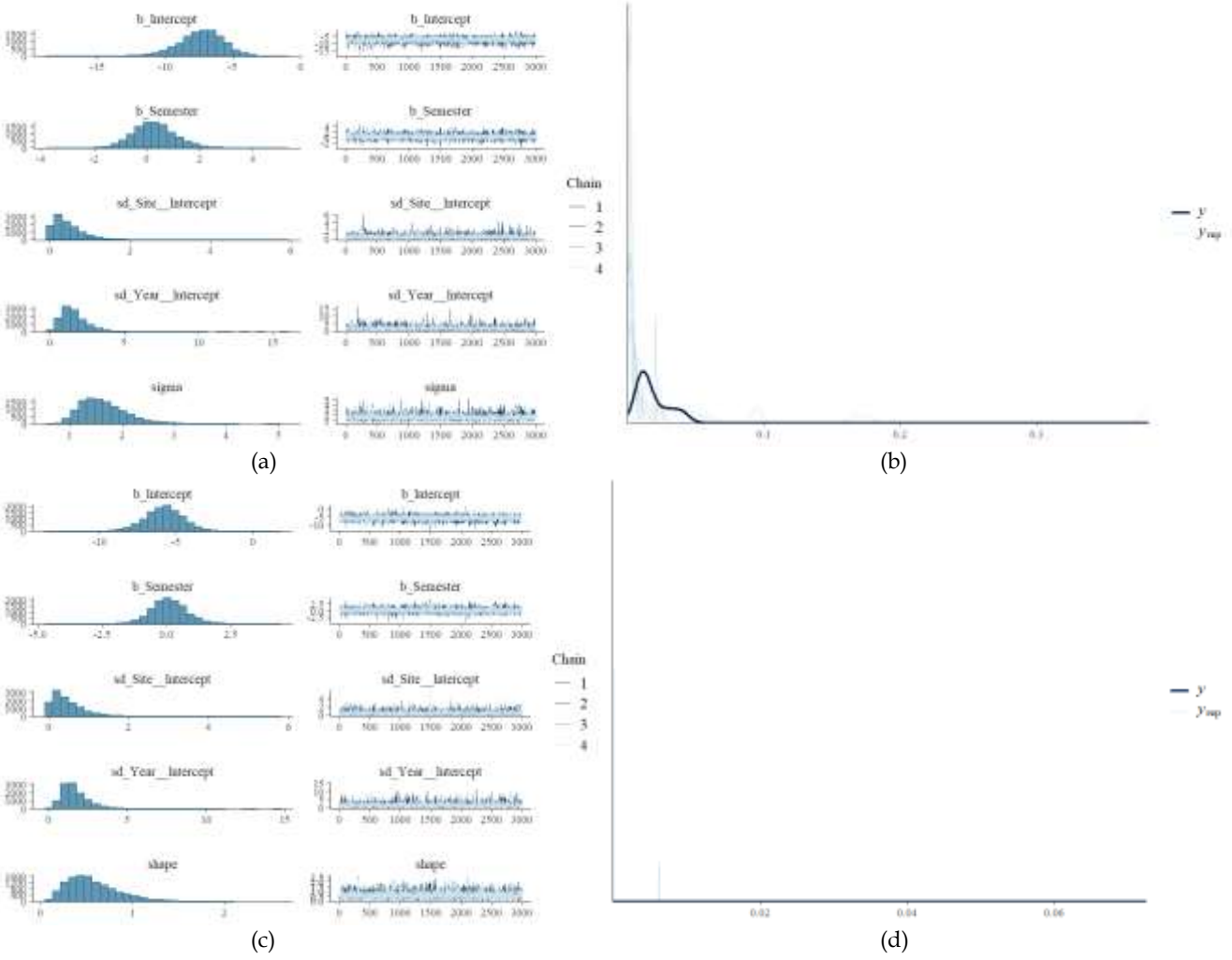
Before interpreting the results, it is necessary to check whether the model can interpret the data distribution properly, namely by seeing whether the model has converged by looking at the trace-plot graph or through the Rhat and ESS (Effective Sample Size) values. In the trace-plot graph, convergence can be seen whether the chains are mixing well separately but also

overlapping one another (Gelman et al., 2013). In the trace-plot graphs in Fig 3 (a), (c), (e), and (g), it can be seen that chains 1 - 4 are well mixed and overlaying one another. This indicates that MCMC has converged well. In addition to the trace-plot graphs, the convergence of MCMC to the model can also be seen numerically through the Rhat and ESS values. In

Table **3**, all Rhat values = 1.00, which means that the model converges well. The Bulk_ESS and Tail_ESS values show the number of effective sample sizes in the mean and credible intervals distributions, respectively. The largest Bulk_ESS and Tail_ESS are $gamma_{WIP}$, which indicates the best model convergence. However, it should be noted that (e) and (g), which are the $lognormal_{DP}$ and $gamma_{DP}$ trace-plot graphs, respectively, require a higher adapt_delta compared to $lognormal_{WIP}$ and $gamma_{WIP}$, because $lognormal_{DP}$ resulting 2 divergent transitions and $gamma_{DP}$ resulting 54 divergent transitions. So, it can be concluded that the prior on $lognormal_{WIP}$ and $gamma_{WIP}$ is more suitable for the data distribution. Therefore, after investigating both graphical and numerical indicators of convergence, we choose $lognormal_{WIP}$ and $gamma_{WIP}$ as the models with the best convergence.

**Table 3.** Convergence check of MCMC sampling and LOO test of distribution family on orthophosphate

| | Rhat | Bulk_ESS | Tail_ESS | Elpd_diff | SE_diff | Mean percentage below LOD | Divergent transitions | Adapt_delta |
|---|---|---|---|---|---|---|---|---|
| $lognormal_{WIP}$ | 1.00 | 5866 | 5681 | -2.4 | 1.3 | 0.855 | 0 | 0.99 |
| $gamma_{WIP}$ | 1.00 | 7590 | 6863 | 0 | 0 | 0.855 | 0 | 0.99 |
| $lognormal_{DP}$ | 1.00 | 3741 | 4504 | -2.1 | 0.7 | 0.871 | 2 | 0.999 |
| $gamma_{DP}$ | 1.00 | 4470 | 5787 | 0 | 0 | 0.871 | 54 | 0.9999 |



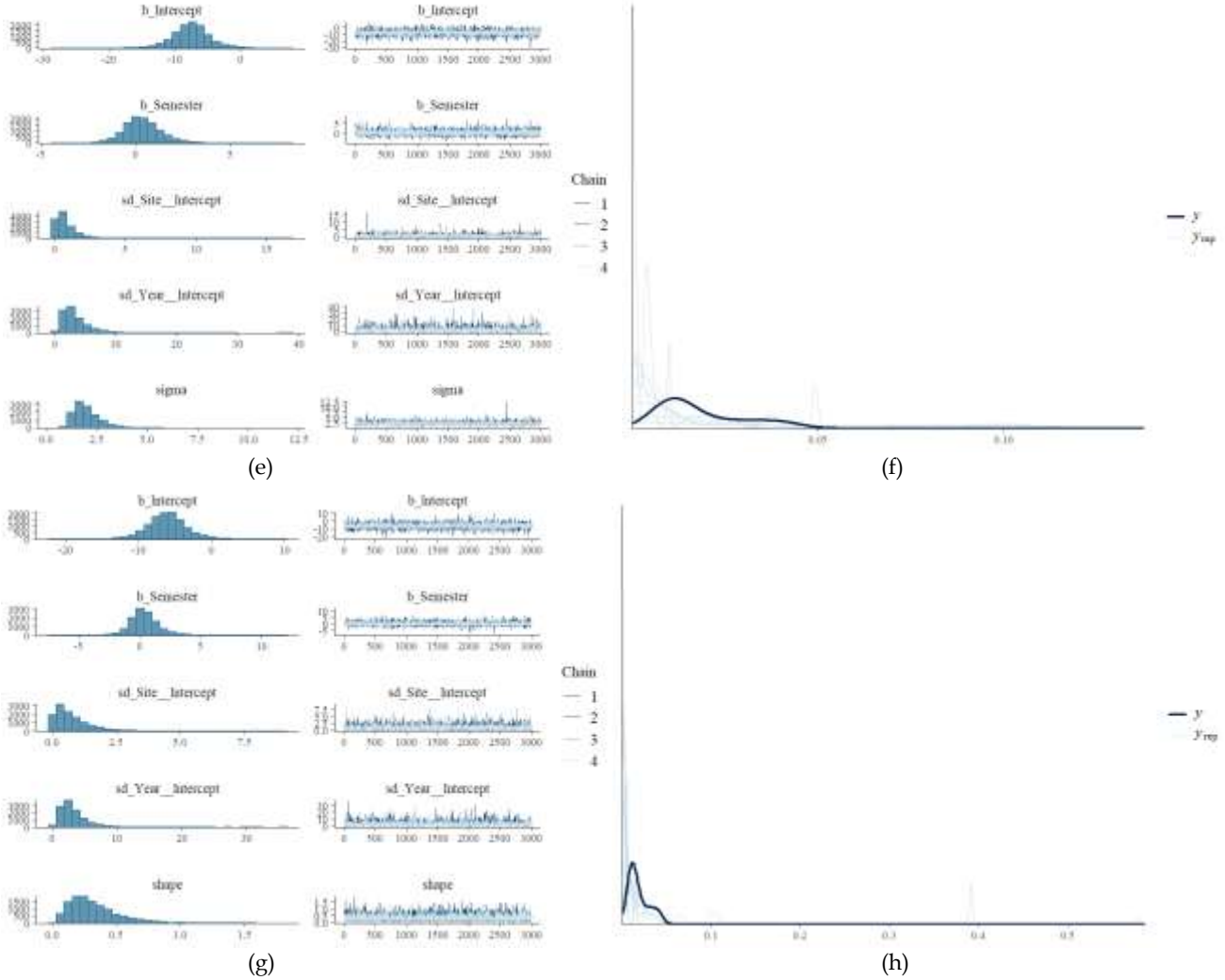



Fig 3. Summary of posterior predict distribution, (a) traceplot $lognormal_{WIP}$, (b) pp_check $lognormal_{WIP}$, (c) traceplot $gamma_{WIP}$, (d) pp_check $gamma_{WIP}$, (e) traceplot $lognormal_{DP}$, (f) pp_check $lognormal_{DP}$, (g) traceplot $gamma_{DP}$, and (h) pp_check $gamma_{DP}$

After determining the model that converges well, the next step is to compare which distribution best fits the data using LOO_compare.

Table ***3*** shows that elpd_diff and se_diff between the $lognormal_{WIP}$ and $gamma_{WIP}$ distributions have differences of -2.4 and 1.3, respectively, with $gamma_{WIP}$ having a higher value than $lognormal_{WIP}$. Therefore, it can be concluded that $gamma_{WIP}$ can predict the model better, and the estimated value is taken from the posterior predict $gamma_{WIP}$ results. Fig 3 (a), (c), (e), and (g) show the posterior distribution graphs. The graphs show that imputation of the posterior distribution results from MCMC produces b_intercept values of -10 to -5 (on a log scale) for $lognormal_{WIP}$ and $gamma_{WIP}$, while $lognormal_{DP}$ and $gamma_{DP}$ produce a wider distribution of -10 to 0. This is relevant because $lognormal_{DP}$ and gamma use a wider prior. The distribution of b_semester across all graphs shows results around 0, which means that the semester effect is not significant. The distribution of b_semester across all graphs shows results around 0, which means that the semester effect is not significant, and varying effects on site and year show a right-skewed distribution with results using a wider default prior. In the posterior predictive check (pp_check) graphs, Fig 3 (b), (d), (f), and (h) show that $gamma_{WIP}$ has an yrep simulation graph that closely follows the observed data y graph, so it can be concluded that the $gamma_{WIP}$ model can produce reasonable estimates.

*Estimation Result and Impact on CCME-WQI*

The posterior predict can be seen in

Table ***4***, where left-censored exceeding the standard are found in orthophosphate data site B year 2021 semester 2, C year 2021 semester 2, D year 2024 semester 2, E year 2024 semester 2, and F year 2024 semester 2. The posterior predicts results in

Table ***4*** show a lower median value compared to left-censored substitution with half of limit detection, indicating that the Bayesian-Tobit model can correct bias when using the substitution method. Additionally, the advantage of using the Bayesian-Tobit model to handle left-censored data is that credible intervals are used to form the basis for analyzing the resulting estimation distribution.

**Table 4.** Posterior predict distribution

| Site | Year | Semester | Left-censored | Median of posterior predict | Lower CI 95% | Upper CI 95% |
|---|---|---|---|---|---|---|
| B | 2021 | 2 | 0.018 | 0.00187 | 2.64e-07 | 0.030 |
| C | 2021 | 2 | 0.018 | 0.00189 | 2.75e-07 | 0.032 |
| D | 2024 | 2 | 0.030 | 0.00034 | 2.92e-08 | 0.013 |
| E | 2024 | 2 | 0.030 | 0.00028 | 2.55e-08 | 0.010 |
| F | 2024 | 2 | 0.030 | 0.00031 | 2.88e-08 | 0.012 |

**Table 5.** CCME-WQI

| | | Site | CCME-WQI | CCME-WQI Lower | CCME-WQI Upper |
|---|---|---|---|---|---|
| Substitution methods | Substitute with zero | A | 90.48 | | |
| | | B | 85.91 | | |
| | | C | 88.36 | | |
| | | D | 90.39 | | |
| | | E | 86.11 | | |
| | | F | 88.51 | | |
| | Substitute with half LOD | A | 90.48 | | |
| | | B | 85.91 | | |
| | | C | 88.36 | | |
| | | D | 90.39 | | |
| | | E | 86.11 | | |
| | | F | 88.51 | | |
| | Substitute with LOD | A | 90.48 | | |
| | | B | 83.85 | | |
| | | C | 86.28 | | |
| | | D | 90.22 | | |
| | | E | 85.97 | | |
| | | F | 88.38 | | |
| Bayesian-Tobit model | | A | 90.48 | 90.48 | 90.48 |
| | | B | 85.91 | 85.91 | 83.81 |
| | | C | 88.36 | 88.36 | 86.24 |
| | | D | 90.39 | 90.39 | 90.39 |
| | | E | 86.11 | 86.11 | 86.11 |
| | | F | 88.51 | 88.51 | 88.51 |
| Exclude Parameter with LOD | | A | 90.12 | | |
| | | B | 85.37 | | |
| | | C | 87.92 | | |
| | | D | 92.14 | | |
| | | E | 87.80 | | |
| | | F | 90.22 | | |

The posterior distribution was then used in the CCME-WQI calculation and determined the seawater quality category for each site in the 2021-2024 period at the Sunda Strait. The results in

Table ***5*** showed that the CCME-WQI increased when compared to the substitution of left-censored with LOD, such as at site B from 83.85 to 85.91. This change indicates that Bayesian-Tobit can correct underestimation. Meanwhile, imputation of upper credible intervals 95% also affects the CCME-WQI, such as at site B from 85.91 to 83.81, which means that the model can correct overestimated CCME-WQI, as we could see at Figure 4. Simple substitution techniques have been shown in numerous studies to create systematic bias in the interpretation of environmental data. These techniques frequently result in an overestimation or underestimate of the concentrations of pollutants (Albert et al., 2024; George et al., 2021). The Bayesian-Tobit addresses the uncertainty of censored data and provides a posterior distribution that can be used to predict environmental risk. These findings highlight the importance of using a Bayesian approach

when analyzing environmental data with a high left-censored.

For comparison, this study also calculated the CCME-WQI by excluding parameters that had left-censored values exceeding the standard. It can be seen in Figure 4 that the CCME-WQI resulting from not including these parameters in the calculation can cause the CCME-WQI value change to a higher value, so it can be concluded that the CCME-WQI result is greatly influenced by the number and deviation of the parameters used in the calculation. In general, the CCME-WQI results for the Sunda Strait are in the "good" category, with the highest index value at site D and the lowest index value at site B. This indicates that the types of activities around the location can affect differences in water quality index.

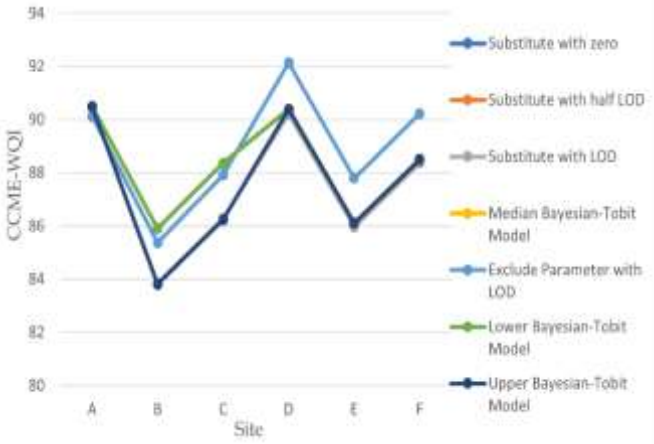



**Figure 4.** CCME-WQI

## Conclusion

Calculating water quality indices can be difficult due to factors such as left-censored data, which cannot be used directly in calculations. In fact, water quality indices can be useful in environmental management policy analysis. As significant left-censored values are present in marine parameter data, a method for handling left-censored data is required for the calculation of the CCME-WQI. The Bayesian-Tobit model can estimate values below the detection limit and is a more reliable method than conservative substitution, which can produce underestimated or overestimated values. As a convergence check and posterior predict result, a $\text{gamma}_{\text{WIP}}$ model can produce reasonable estimate, so the quality of seawater in the Sunda Strait using the CCMI-WQI method is in the "good" category (83.8-92.1) for the period 2021 to 2024. By knowing the marine condition through the water quality index, it can be a reference for the local government in making coastal zone management policies, and the Bayesian-Tobit model approach can be used in determining CCME-WQI and in other scientific applications.

**Acknowledgment**

Thank you to all parties who have helped in this research so that this article can be published.

**Author Contributions**

Conceptualization, I.A., S.S., and N.F.; methodology, I.A., S.S., and N.F; data and software, I.A.; visualization, I.A.; validation, S.S., and N.F.; analysis, I.A., S.S., and N.F.; writing, I.A., S.S., and N.F.; supervision, S.S., and N.F. All authors have read and agreed to the published version of the manuscript.

**Funding**

This research received no external funding.

**Conflicts of Interest**

There are no conflicts to declare.

## References

Al-Qadami, E. H. H., Razi, M. A. M., Shah, S. M. H., Mokhtar, A., Mahamud, M., Jamal, M. H., & Ismail, Z. A. (2025). Marine Water and Sediment Quality Assessment of Mainland Kedah and Langkawi Island Shorelines: A Case Study. *IOP Conference Series: Earth and Environmental Science*, *1453*(012058), 1–11. https://doi.org/10.1088/1755-1315/1453/1/012058

Albert, D. L., Hardy, W. N., & Kemper, A. R. (2024). Effect of data censoring on thoracic injury risk curves. *Traffic Injury Prevention*, *25*(sup1), S33–S42. https://doi.org/10.1080/15389588.2024.2405643

Bürkner, P. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, *10*(July), 395–411. https://journal.r-project.org/articles/RJ-2018-017/RJ-2018-017.pdf

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Canadian Council of Ministers of the Environment. (2017). CCME Water Quality Index user's manual 2017 Update. *Canadian Water Quality Guidelines for the Protection of Aquatic Life*, 1–5. https://ccme.ca/en/res/wqimanualen.pdf

Chasanah, N., Armono, H. D., & Radianta, W. (2020). Pemodelan Penjalaran Tsunami Akibat Erupsi Gunung Anak Krakatau Beserta Skenario Dike, Studi Kasus Teluk Jakarta. *Jurnal Teknik ITS*, *9*(1), 23–30. https://doi.org/10.12962/j23373539.v9i1.50609

Dagne, G. A., & Huang, Y. (2022). Bayesian Semiparametric Mixture Tobit Models with Left-Censoring, Skewness and Covariate Measurement Errors. *BMC Cancer*, *23*(1), 1–7.

https://doi.org/10.1002/sim.5799.Bayesian

Elgendy, A. R., El, A., El, M. S., Sawy, M. A. El, Alprol, A. E., & Zaghloul, G. Y. (2024). A comparative study of the risk assessment and heavy metal contamination of coastal sediments in the Red sea , Egypt , between the cities of El - Quseir and Safaga. *Geochemical Transactions*, *25*(3), 1–23. https://doi.org/10.1186/s12932-024-00086-8

Fahlevi, M. R., Bayhaqi, A., Sugianto, D. N., Fadli, M., Wang, H., Susanto, R. D., & Wouthuyzen, S. (2022). Karakteristik Massa Air di Selat Sunda dan Perairan Lepasnya. *Buletin Oseanografi Marina*, *11*(3), 231–247. https://doi.org/10.14710/buloma.v11i2.41323

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (Third). Chapman and Hall/CRC. https://doi.org/10.1201/b16018

Gelman, A., Simpson, D., & Betancourt, M. (2017). The Prior Can Often Only Be Understood in the Context of the Likelihood. *Entropy*, *19*(555), 1–13. https://doi.org/10.3390/e19100555

George, B. J., Gains-germain, L., Broms, K., Black, K., Hays, M. D., Thomas, K. W., & Simmons, J. E. (2021). Censoring trace-level environmental data: statistical analysis considerations to limit bias. *Environmental Science & Technology*, *55*(6), 3786–3795. https://doi.org/10.1021/acs.est.0c02256.Censoring

Giovanni, N.-G. S. D. and I. S. (2025). *Average Total Precipitation from Days*. https://giovanni.gsfc.nasa.gov/giovanni/

Haeruddin, Widowati, I., Rahman, A., Rumanti, M., & Iryanthony, S. B. (2021). Bioconcentration of lead (Pb) and cadmium (cd) in green-lipped mussels (perna viridis) in the coastal waters of semarang bay, indonesia. *AACL Bioflux*, *14*(3), 1581–1595. https://bioflux.com.ro/docs/2021.1581-1595.pdf

Huynh, T., Ramachandran, G., Banerjee, S., Monteiro, J., Stenzel, M., Sandler, D. P., Engel, L. S., Kwok, R. K., Blair, A., & Stewart, P. A. (2014). Comparison of Methods for Analyzing Left- Censored Occupational Exposure Data. *Ann. Occup. Hyg.*, *58*(9), 1126–1142. https://doi.org/10.1093/annhyg/meu067

Iqbal, M., Denhi, A. D. A., Kristianto, & Prayoga, A. (2023). Morphological Analysis of Anak Krakatau Volcano after 22 December 2018 Eruption using Differential Interferometry Synthetic Aperture Radar (DInSAR). *Journal of Geoscience, Engineering, Environment, and Technology*, *8*(2), 90–98. https://doi.org/10.25299/jgeet.2023.8.2.11651

Kartini, N., Boer, M., & Affandi, R. (2017). Pola Rekrutmen, Mortalitas, dan Laju Eksploitasi Ikan Lemuru (Amblygaster sirm, Walbaum 1792) di Perairan Selat Sunda. *Biospecies*, *10*(1), 11–16. https://doi.org/10.22437/biospecies.v10i1.3483

Kementerian Lingkungan Hidup. (2025). *Sistem Informasi Pelaporan Elektronik Lingkungan Hidup*. https://simpel.menlhk.go.id/2023/pemda/evaluasi

Lemoine, N. P. (2019). Moving beyond noninformative priors: why and how to choose weakly informative priors in Bayesian analyses. *Oikos*, *128*, 912–928. https://doi.org/10.1111/oik.05985

Lestari, D. A., Anzani, L., Zamil, A. S., Prasetyo, A., Simbolon, E. F., & Apriansyah, M. R. (2020). Pengaruh Gunung Laut Anak Krakatau Terhadap Pertumbuhan Rumput Laut di Selat Sunda. *Jurnal Kemaritiman : Indonesian Journal of Maritime*, *1*(2), 75–88. https://doi.org/10.17509/ijom.v1i2.25590

Li, S., Wei, Z., Susanto, R. D., Zhu, Y., Setiawan, A., Xu, T., Fan, B., Agustiadi, T., Trenggono, M., & Fang, G. (2018). Observations of intraseasonal variability in the Sunda Strait throughflow. *Journal of Oceanography*, *74*(5), 541–547. https://doi.org/10.1007/s10872-018-0476-y

Liu, S., Shi, J., Wang, J., Dai, Y., Li, H., Li, J., Liu, X., Chen, X., Wang, Z., & Zhang, P. (2021). Interactions Between Microplastics and Heavy Metals in Aquatic Environments: A Review. *Frontiers in Microbiology*, *12*(April), 1–14. https://doi.org/10.3389/fmicb.2021.652520

Ma, Z., Li, H., Ye, Z., Wen, J., Hu, Y., & Liu, Y. (2020). Application of modified water quality index (WQI) in the assessment of coastal water quality in main aquaculture areas of Dalian, China. *Marine Pollution Bulletin*, *157*(May), 1–8. https://doi.org/10.1016/j.marpolbul.2020.111285

McElreath, R. (2020). *Statistical Rethinking - A Bayesian Course with Examples in R and Stan* (Second). CRC Press. https://doi.org/10.1201/9780429029608

Munandar, E., Susanto, A., Nurdin, H. S., Syafrie, H., Hamzah, A., Khalifa, M. A., Budiaji, W., Febrio, E. P., & Dewi, I. Y. (2022). Dynamics of Coral Reefs Condition Before and After Sunda Strait Tsunami in Badul Island. *Jurnal Perikanan Dan Kelautan*, *12*, 221–229. http://dx.doi.org/10.33512/jpk.v12i2.17423

Nugroho, T. M., Hartoko, A., & Suryanti. (2024). Analysis Of Mangrove Damage Before and After Tsunami in Sangiang Island Based on NDVI. *Journal of Marquares*, *11*(1), 65–71. https://doi.org/10.14710/marj.v11i1.30111

Nurfajriah, U., Ijonu, S., Nyoman, I. G., Jaya, M., & Arisanti, R. (2024). Spatially Varying Regression Coefficient Model For Predicting Stunting

Hotspots In Indonesia. *Journal of Research in Science Education*, *10*(10), 7748–7755. https://doi.org/10.29303/jppipa.v10i10.8270

Penyelenggaraan Perlindungan Dan Pengeolaan Lingkungan Hidup, Pub. L. No. 22, VIII Kementrian Sekretarian Negara Republik Indonesia, 483 (2021). http://www.jdih.setjen.kemendagri.go.id/

Ramazanova, E., Bahetnur, Y., Yessenbayeva, K., Lee, S. H., & Lee, W. (2022). Spatiotemporal evaluation of water quality and risk assessment of heavy metals in the northern Caspian Sea bounded by Kazakhstan. *Marine Pollution Bulletin*, *181*(113879), 1–10. https://doi.org/10.1016/j.marpolbul.2022.113879

Ravenzwaaij, D. Van, Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov Chain Monte – Carlo sampling. *Psychon Bull Rev*, *25*, 143–154. https://doi.org/10.3758/s13423-016-1015-8

Ristanto, D., Ambariyanto, A., & Yulianto, B. (2021). Water Quality Assessment Based on National Sanitations Foundation Water Quality Index during Rainy Season in Sibelis and Kemiri Estuaries Tegal City. *IOP Conference Series: Earth and Environmental Science*, *750*(012013), 1–10. https://doi.org/10.1088/1755-1315/750/1/012013

Röver, C., Schmidli, H., Schmid, C. H., Weber, S., & Friede, T. (2021). On weakly informative prior distributions for the heterogeneity parameter in Bayesian random-effects meta- analysis. *Research Synthesis Methods*, *12*, 448–474. https://doi.org/10.1002/jrsm.1475

Safford, H., Zuniga-montanez, R. E., Kim, M., Wu, X., Wei, L., Sharpnack, J., Shapiro, K., & Bischel, H. N. (2022). Wastewater-Based Epidemiology for COVID-19: Handling qPCR Nondetects and Comparing Spatially Granular Wastewater and Clinical Data Trends. *ACS EST Water*, *2*, 2114–2124. https://doi.org/10.1021/acsestwater.2c00053

Sobaruddin, D. P., Armawi, A., & Martono, E. (2017). Model Traffic Separation Scheme (TSS) Di Alur Laut Kepulauan Indonesia (AlKI) I Di Selat Sunda Dalam Mewujudkan Ketahanan Wilayah. *Jurnal Ketahanan Nasional*, *23*(1), 104. https://doi.org/10.22146/jkn.22070

Subekti, S., Amiin, M. K., Ardiyanti, H. B., Yudarana, M. A., Achmadi, I., & Akbar, R. E. K. (2021). Molecular epidemiology of helminth diseases of the humpback grouper , Cromileptes altivelis , as a pattern for mapping fish diseases in the. *Veterinary World*, *14*(5), 1324–1329. https://www.doi.org/10.14202/vetworld.2021.1324-1329

Susanto, A., Nurdin, H. S., Khalifa, M. A., Munandar, E., Syafrie, H., Alansar, T., Sulistyono, B., & Raihan, A. (2023). Sunda Strait Coastal Management with Mangrove Planting as an Effort for Mitigation of Disaster and Climate Change (Blue Carbon). *Journal of Maritime Empowerment*, *5*(2), 48–55. https://doi.org/10.31629/jme.v5i2.5711

Tolkou, A. K., Toubanaki, D. K., & Kyzas, G. Z. (2023). Detection of Arsenic, Chromium, Cadmium, Lead, and Mercury in Fish: Effects on the Sustainable and Healthy Development of Aquatic Life and Human Consumers. *Sustainability (Switzerland)*, *15*(23), 1–17. https://doi.org/10.3390/su152316242

Uddin, M. G., Nash, S., & Olbert, A. I. (2021). A review of water quality index models and their use for assessing surface water quality. *Ecological Indicators*, *122*(107218), 1–21. https://doi.org/10.1016/j.ecolind.2020.107218

Wood, M. D., Beresford, N. A., & Copplestone, D. (2011). Limit of detection values in data analysis: Do they matter? *Radioprotection*, *46*(6 SUPPL.), 85–90. https://doi.org/10.1051/radiopro/20116728s

Yonvitner, Y., LIoret, J., Boer, M., Kurnia, R., Akmal, S. G., Yuliana, E., Yani, D. E., Gómez, S., & Setijorini, L. E. (2020). Vulnerability of marine resources to small-scale fishing in a tropical area: The example of Sunda Strait in Indonesia. *Fisheries Management and Ecology*, *27*(5), 472–480. https://doi.org/10.1111/fme.12428