applied sciences

MDPI

Article

A Weighted Bayesian Kernel Machine Regression Approach for Predicting the Growth of Indoor-Cultured Abalone

Seung-Won Seo [1,†], Gyumin Choi [2,†], Ho-Jin Jung [1], Mi-Jin Choi [3,4], Young-Dae Oh [3], Hyun-Seok Jang [3], Han-Kyu Lim [3,4,*] and Seongil Jo [2,*]

[1] Silicogen Inc., Yongin 16954, Republic of Korea; swseo@insilicogen.com (S.-W.S.); hjjung@insilicogen.com (H.-J.J.)
[2] Department of Statistics and Data Science, Inha University, Incheon 22212, Republic of Korea; a1b1c2d3e5@naver.com
[3] Smart Aqua Farm Convergence Research Center, Mokpo National University, Muan 58554, Republic of Korea; coacoawls@gmail.com (M.-J.C.); oh9108@mnu.ac.kr (Y.-D.O.); houston7777@naver.com (H.-S.J.)
[4] Department of Biomedicine, Health & Life Convergence Science, BK21 Four, Mokpo National University, Muan 58554, Republic of Korea
* Correspondence: limhk@mnu.ac.kr (H.-K.L.); bstatsjo@gmail.com (S.J.)
† These authors contributed equally to this work.

Abstract: The cultivation of abalone, a species with high economic value, faces significant challenges due to its slow growth rate and sensitivity to environmental conditions, resulting in prolonged cultivation periods and increased mortality risks. To address these challenges, we propose a novel probabilistic machine learning approach based on a Bayesian framework to predict abalone growth by modeling key environmental factors, including water temperature, pH, salinity, nutrient supply, and dissolved oxygen levels. The proposed method employs a weighted Bayesian kernel machine regression model, integrating Gaussian processes with a spike-and-slab prior to identify influential variables. This approach accommodates heteroscedasticity, capturing varying levels of variance across observations, and models complex, non-linear relationships between environmental factors and abalone growth. Our analysis reveals that time, dissolved oxygen, salinity, and nutrient supply are the most critical factors influencing growth, while water temperature and pH play relatively minor roles under controlled indoor farming conditions. Interaction analysis highlights the non-linear dependencies among factors, such as the combined effects of salinity and nutrient supply. The proposed model not only improves prediction accuracy compared to baseline methods, but also provides actionable insights into the environmental dynamics that optimize abalone growth. These findings underscore the potential of advanced machine learning techniques in enhancing aquaculture practices and offer a robust framework for managing complex, multi-variable systems in sustainable farming.

Keywords: abalone growth; heteroscedasticity; indoor abalone farming; weighted Bayesian kernel machine regression

check for updates

Academic Editors: Vlasoula Bekiari and Franco Mutinelli

Received: 31 October 2024
Revised: 13 December 2024
Accepted: 7 January 2025
Published: 13 January 2025

Citation: Seo, S.-W.; Choi, G.; Jung, H.-J.; Choi, M.-J.; Oh, Y.-D.; Jang, H.-S.; Lim, H.-K.; Jo, S. A Weighted Bayesian Kernel Machine Regression Approach for Predicting the Growth of Indoor-Cultured Abalone. *Appl. Sci.* **2025**, *15*, 708. https://doi.org/10.3390/app15020708

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Abalone, a highly valued marine resource, holds significant economic importance, particularly in Korea, where its market value reached KRW 347.5 billion in 2016 (Korea National Statistical Office, 2018 [1]). This makes it the most valuable cultured shellfish in the country. However, abalone cultivation faces considerable challenges that threaten its sustainability and profitability. The species' inherently slow growth rate necessitates prolonged cultivation periods, leaving it vulnerable to environmental stresses. These

challenges are compounded by additional factors, including viral infections (Won et al., 2013 [2]) and the adverse effects of high-density farming practices (Kim et al., 2013 [3]). Furthermore, fluctuations in water temperature present a critical concern, often resulting in mass mortality events. Such temperature shifts increase oxygen demand, leading to hypoxic conditions that severely stress the abalone (Vosloo et al., 2013 [4]).

The cultivation of abalone is not only vital for its economic contributions, but also for meeting the growing demand for high-value seafood products globally. Addressing the multifaceted challenges associated with its farming requires innovative strategies and tools. Specifically, there is an urgent need for advanced predictive models capable of accurately assessing the effects of environmental factors on abalone growth. By developing such models, we can mitigate risks, improve survival rates, and optimize growth conditions, contributing to more sustainable aquaculture practices.

In this paper, we address these challenges by introducing a probabilistic machine learning-based approach using a Bayesian framework to assess the effects of key environmental factors, including water temperature, pH, salinity, nutrient supply, and dissolved oxygen levels. The proposed model is based on a Bayesian kernel machine regression model (Bobb et al., 2015 [5]; Bobb et al., 2018 [6]) with a Gaussian process prior (Williams and Rasmussen, 2006 [7]).

This model employs a spike-and-slab prior on the inverse of the length-scale parameters to determine the influence of growth-related factors. By leveraging Gaussian processes, the model captures complex, non-linear relationships among the factors. Furthermore, it addresses heteroscedasticity by modeling the error variance, allowing for varying levels of variance across observations. The spike-and-slab prior also enables automatic variable selection, identifying the key environmental factors with the most significant impact on growth. This approach provides not only accurate growth predictions, but also deeper insights into the relative importance of these factors, contributing to a better understanding of abalone growth dynamics.

This paper makes the following key contributions:

- We propose a novel weighted Bayesian kernel machine regression (WBKMR) model that integrates Gaussian processes with a spike-and-slab prior, enabling both accurate predictions and variable selection.
- The model explicitly accounts for heteroscedasticity, allowing for varying levels of error variance across observations, which is critical for analyzing aquaculture data.
- Using Posterior Inclusion Probability (PIP) analysis, the model identifies the most critical environmental factors, such as dissolved oxygen, nutrient supply, and salinity, and quantifies their relative influence on abalone growth.
- Comparative experiments demonstrate that the proposed model achieves superior predictive accuracy compared to baseline methods.

The remainder of this paper is organized as follows. Section 2 reviews the related work on abalone growth prediction and machine learning models. Section 3 introduces the definitions and properties of Gaussian processes and kernel machine regression, which are foundational to our proposed approach. Section 4 describes the growth data of northern and disk abalones, along with the Bayesian kernel machine regression model used for analysis. The results of the analysis are presented in Section 5, followed by a detailed discussion in Section 6. Finally, Section 7 summarizes the findings and outlines potential directions for future research.

2. Related Work

Several methods have been proposed in various studies for predicting abalone growth. For example, Helidoniotis (2011) [8] employed regression-based models to predict abalone

growth. This study focused specifically on Haliotis rubra, examining how different environmental factors, such as water temperature, feeding regimes, and stocking density, influence growth rates and overall productivity in aquaculture. By utilizing these regression models, the research aimed to enhance our understanding of optimal growth conditions and to provide actionable insights for improving productivity.

Jabeen and Ahamed (2016) [9] introduced artificial neural networks to predict the age of abalone, which is closely related to growth. The study utilized the Levenberg–Marquardt algorithm for its model, which is known for its fast and stable convergence, making it suitable for medium-sized predictive modeling problems. Misman et al. (2019) [10] discussed the use of regression-based neural networks to predict the age of abalone from physical characteristics, a key factor in understanding and predicting growth rates.

More recently, Khiem et al. (2023) [11] proposed an ensemble method for predicting the growth of abalone reared in land-based aquaculture. This study uses an ensemble of machine learning algorithms, including random forest, gradient boosting, support vector machine, and neural networks, to predict the growth of indoor-cultured abalone. The study highlights the importance of environmental control for maximizing growth and evaluates the effects of various environmental factors like water temperature and flow speed on abalone growth. Although these proposed methods are known to predict abalone growth well, they unfortunately do not identify the importance of the factors influencing growth.

3. Background

This section briefly discusses Gaussian processes (GPs) and kernel machine regression (KMR), which are crucial for building probabilistic predictive models based on Bayesian statistics.

3.1. Gaussian Processes

A Gaussian process (GP) is a powerful statistical tool that defines a distribution over functions, allowing for flexible modeling of unknown, complex functions with continuous domains (see, e.g., Williams and Rasmussen, 2006 [7]). Let $\mathcal{T}$ be an arbitrary index set with a continuous domain, such as space or time, and let $f = (f(t), t \in \mathcal{T})$ be a stochastic process defined over $\mathcal{T}$. Then, f is called a second-order GP if the marginal distribution of any finite-dimensional subset of f is a multivariate Gaussian distribution. That is, for every n and $t_1, \ldots, t_n \in \mathcal{T}$,

$$f \equiv (f(t_1), \ldots, f(t_n))^\top \sim \mathcal{N}_n(\mu, K),$$

where $\mu = (\mu(t_1), \ldots, \mu(t_n))^\top$ with $\mu(t_i) = E[f(t_i)]$ is the mean vector, and $K = (K_{ij})_{i,j=1}^n$ with $K_{ij} \equiv K(t_i, t_j) = E[(f(t_i) - \mu(t_i))(f(t_j) - \mu(t_j))]$ is the covariance matrix. A GP is usually denoted as

$$f \sim \mathcal{GP}(\mu, K).$$

In Bayesian modeling, GPs serve as a prior distribution over function spaces, which is especially useful for non-parametric functions when the form of the function is unknown or too complex to model parametrically (see, e.g., Gelman et al., 2014 [12]; Murphy, 2022 [13]). To illustrate this, suppose that we obtain a dataset $\mathcal{D} = \{(y_i, x_i),\ i = 1, \ldots, n\}$, where y_i is a response variable and $x_i = (x_{i1}, \ldots, x_{ip})^\top$ is a p-dimensional vector of predictors that non-parametrically explains y_i. Then, by assuming additive Gaussian noise, we model the relationship between the response and predictors as follows:

$$\begin{aligned} y_i &= f(x_i) + \epsilon_i,\ \epsilon_i \sim \mathcal{N}(0, \sigma^2), \\ f \equiv (f_1, \ldots, f_n) &\sim \mathcal{GP}(\mu, K),\ f_i = f(x_i), \end{aligned} \tag{1}$$

where the mean vector (or mean function) μ is typically taken to be zero, and the covariance matrix (or covariance function) is defined by commonly used the squared exponential kernel,

$$K_{ij} = \tau_f^2 \exp\left[-\frac{1}{2}\|x_i - x_j\|^2 / l^2\right].$$

Here, the parameter l is called a length-scale hyperparameter and τ_f^2 denotes the marginal variance, and these parameters control the shape of the Gaussian process (Hu and Dey, 2023 [14]).

One of the key advantages of GPs is their ability to provide a full posterior predictive distribution when making predictions. This means that for any new predictors, GPs can provide both the expected value (mean) of the prediction and a measure of the uncertainty associated with that prediction. For example, let $x_* = (x_{*1}, \ldots, x_{*p})^\top$ be a new set of predictors that were not utilized in the model specified by (1). The predictions can be computed in the following closed form:

$$f_* \mid x_*, y, x \sim \mathcal{N}(\mu_*, \Sigma_*),$$

where the mean function and covariance function are given by

$$\begin{aligned} \mu_* &= K_{x_*,x}[K + \sigma^2 I]^{-1} y, \\ \Sigma_* &= K_{x_*,x_*} - K_{x_*,x}[K + \sigma^2 I]^{-1} K_{x,x_*}. \end{aligned}$$

3.2. Kernel Machine Regressions

KMR is a non-parametric regression approach designed to capture complex, non-linear, and non-additive relationships between predictors and responses, while adjusting for potential confounding factors (Bobb et al., 2018 [6]). Unlike traditional parametric regression, which assumes a predefined functional form for the relationship between predictors and the response, KMR allows the model to be more flexible, adapting to the underlying structure of the data without such assumptions. This flexibility is crucial when the relationship between variables is highly non-linear and difficult to model explicitly with standard methods.

Suppose $f : \mathbb{R}^p \to \mathbb{R}$ is a function belonging to a function space $\mathcal{H}_K$, which is equipped with a positive semidefinite reproducing kernel $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$. There are two approaches to characterize f. The first approach is to use a basis-function representation, where $f(x) = \sum_{m=1}^{M} \phi_m(x)\eta_m$ for some set of basis function $\{\phi_m\}_{m=1}^{M}$ and coefficients $\{\eta_m\}_{m=1}^{M}$. The second approach directly utilizes the positive-definite kernel function $K(\cdot,\cdot)$, where $f(x) = \sum_{i=1}^{n} K(x_i, x)\alpha_i$ for a set of coefficients $\{\alpha_i\}_{i=1}^{n}$. Mercer's theorem (Mercer, 1909 [15]) provides theoretical foundation, ensuring that a kernel function $K(\cdot,\cdot)$ representing f implicitly defines a unique function that is spanned by a specific set of orthogonal basis functions used in the basis-function representation of f. For specific examples, see Liu et al. (2007) [16], Bobb et al. (2015) [5] and references therein.

4. Materials and Methods

4.1. Data Description

In this paper, we consider a dataset collected from the study by Oh et al. (2018) [17]. The dataset targets the growth data of 9600 northern abalone (Haliotis discus hannai) and disk abalone (H. discus discus) cultured at the Abalone Research Institute (Department of Fisheries Seed) of the Jeonnam Marine and Fisheries Science Institute in Wando, Jeollanam-do. The abalones were spawned through artificial fertilization into four breeds (northern × northern cross, northern × disk cross, disk × disk cross, disk × northern

cross), and then cultured in land-based tanks for 250 days. From these, individuals with a shell length of over 30 mm were selected, totaling 9600 across the breeds, and transferred to sea cages in Wando County, where their growth was monitored over 967 days. Each cage, measuring 2.4 m in length, 1.2 m in width, and 2 m in height, housed 800 abalones, utilizing a total of 12 cages. The abalones' feed consisted of cultured wakame (Undaria pinnatifida) and kelp (Laminaria japonica), including wakame stems, supplied every two weeks at 40 kg.

To evaluate the impact of environmental factors on abalone growth, water temperature (water.temp), salinity (salt), dissolved oxygen (DO), nutrient supply (NS), and pH were measured every two weeks over a 967-day cultivation period, resulting in 61 sets of environmental data. These variables are listed in Table 1. To minimize the impact of rearing management on growth, no other work such as selection or net cleaning was performed, except for one net change on the 600th day.

Table 1. List of potential variables used in predictions.

Variable	Description
Water temperature (°C)	Measured at two-week intervals for 950 days using marine environmental survey equipment (YSI 5908, Xylem, Yellow Springs, OH, USA).
DO (mg/L)	Dissolved oxygen levels in the water.
pH	Acidity or alkalinity of the water.
Salinity (psu)	Salinity levels in the water.
Nutrient supply (kg)	Supply of 40 kg of seaweed (*Undaria pinnatifida*), kelp (*Laminaria japonica*), and seaweed stems cultivated in nearby areas every two weeks.

Summary statistics for these factors are presented in Table 2, and the patterns of water temperature and DO is plotted in Figure 1. In the table, Q2 is the median value, and Q1 and Q3 represent the 25th percentile and the 75th percentile, respectively. From the figure, we can see that water temperature showed a clear seasonal pattern, while dissolved oxygen levels generally decreased in summer and increased in winter, without showing a consistent overall pattern. In contrast, salinity and pH did not show significant changes during the same period. Salinity was maintained at 32–34 psu regardless of seasonal or rainy season effects, and pH also maintained a value between 7 and 9.

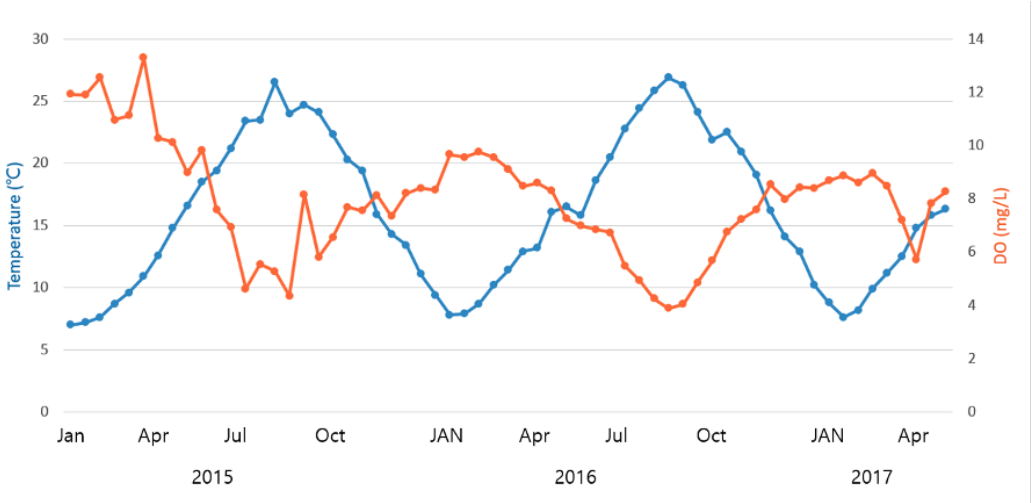


Figure 1. Temperature and DO values over time.

Table 2. Summary statistics for environmental factors.

Variables	Mean	Sd	Min	Q1	Q2	Q3	Max
Water Temperature	16.94	5.840	8.700	12.900	15.900	22.800	26.500
DO	7.801	2.126	4.050	5.820	7.990	8.480	13.300
pH	8.385	0.522	7.210	8.440	8.560	8.760	8.940
Salinity	33.080	0.673	31.820	32.540	33.040	33.520	34.690
NS	45.840	8.124	40.000	40.000	40.000	50.000	60.000

The growth of abalones was assessed every 50 days by randomly sampling 90 individuals per breed and measuring shell length, width, height, and weight. Growth data were collected 16 times over the 967-day period, and we focus specifically on weight, as it is the most important factor for understanding growth based on economic value. The growth and environmental data were integrated by matching measurement dates, with the closest environmental data used when exact matches were unavailable.

Prior to conducting the analysis, we performed additional data preprocessing. This involved removing missing values for abalone weight and eliminating outliers from the abalone weight measurements at each time point. Outliers were identified using the criterion of Q1 $\pm$ 1.5 IQR, where IQR is the interquartile range. This resulted in the removal of 109 data points. Figure 2 shows scatter plots of the dataset before and after outlier removal. The left panel of the figure displays the scatter plot before outlier removal, with the red points indicating the identified outliers, while the right panel shows the scatter plot after outlier removal.

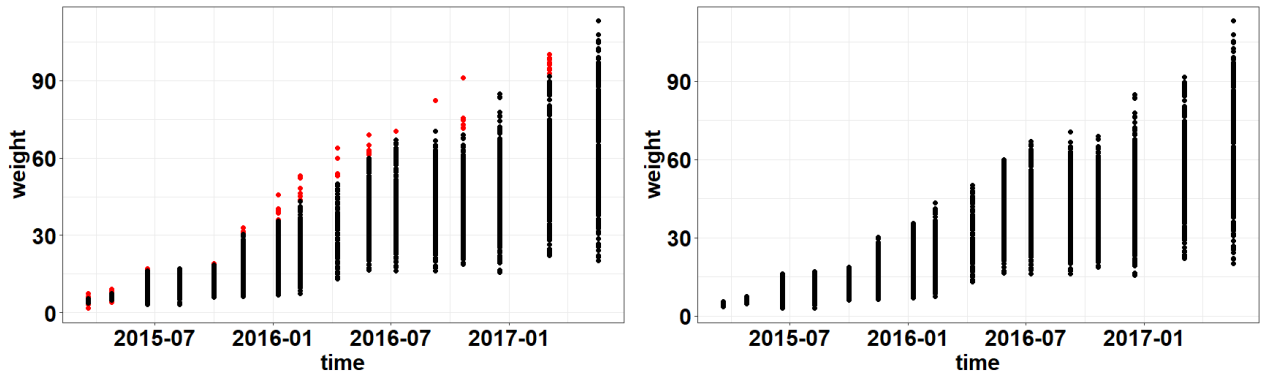


Figure 2. Scatter plots of weight over time: The (**left**) panel displays the plot before outlier removal, while the (**right**) panel shows the scatter plot after outlier removal.

Figure 3 shows the correlation matrix plot between various environmental factors and the abalone weight. From the figure, it is evident that weight has a strong correlation with factors. Additionally, there are significant relationships among the environmental factors themselves, indicating a complex structure. These complex interdependencies among predictors suggest that a traditional linear model may not be sufficient to capture the interactions influencing abalone growth.

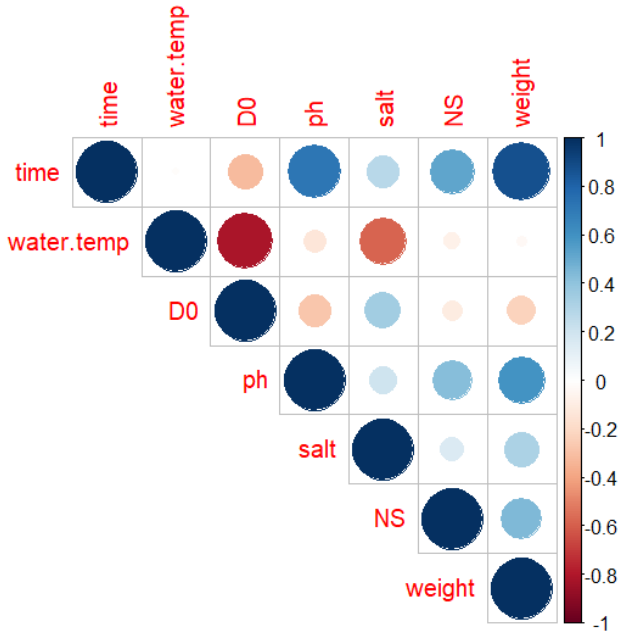


Figure 3. Correlation plot of weight and growth-related environmental factors.

4.2. Weighted Bayesian Kernel Machine Regression

We consider a semiparametric Bayesian model based on the Bayesian kernel machine regression (BKMR), which was proposed in Bobb et al. (2015) [5], with the Gaussian process (GP) prior (see, e.g., Williams and Rasmussen, 2006 [7]). as a novel probabilistic approach for predicting the growth of indoor-cultured abalone. To describe the model in detail, let $\bar{y}_t = n_t^{-1} \sum_{j=1}^{n_t} \log y_{tj}, t = 1, \ldots, T$ represent the average log-transformed growth measurements of indoor-cultured abalones at time t. Let $\boldsymbol{z}_t = (z_{t1}, \ldots, z_{tp})^\top$ be a covariate vector, including dummy variables representing species, and let $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tM})^\top$ be a M-dimensional vector of features influencing prediction of the growth of the indoor-cultured abalone at time t, such as water temperature, dissolved oxygen, power of hydrogen, and water salinity. We introduce a weighted BKMR model, which includes both linear and nonlinear effects of the predictors, as

$$\bar{y}_t = \boldsymbol{z}_t^\top \boldsymbol{\beta} + f(\boldsymbol{x}_t) + \epsilon_t, \;\; \epsilon_t \overset{indep.}{\sim} \mathcal{N}(0, \sigma^2 / n_t), \tag{2}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a p-dimensional coefficient vector and $f(\cdot) : \mathbb{R}^M \to \mathbb{R}$ is an unknown and non-linear function modeled through a GP prior as

$$\boldsymbol{f} = (f_1, \ldots, f_T)^\top \sim \mathcal{GP}(0, \sigma^2 K), \; f_t = f(\boldsymbol{x}_t). \tag{3}$$

Here, $K = (K_{il})_{i,l=1}^T$ is an $T \times T$ positive semi-definite covariance matrix, and $K_{il} : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}$ is a kernel function that governs the smoothness of the realizations derived from the GP and determines the extent of shrinkage towards the mean (Gelman et al., 2014 [12]). In the paper, we use the squared exponential kernel for the kernel function K_{il}, which is defined as

$$K_{il} \equiv K(\boldsymbol{x}_i, \boldsymbol{x}_l) = \lambda_f \exp\left[- \sum_{m=1}^{M} \gamma_m (x_{im} - x_{lm})^2 \right],$$

where $\gamma_m \geq 0, \; m = 1, \ldots, M$, are inverse length-scale parameters, and λ_f is a positive real-valued scaling parameter. The parameter λ_f determines the overall magnitude of the covariance matrix. It effectively controls the amplitude of the realizations from the GP, allowing the model to adapt to the scale of the response variable. The inverse length-scale

parameters, on the other hand, determine the degree of influence of the m-th environmental factor. A larger γ_m value indicates a stronger sensitivity to changes in $\boldsymbol{x}_m$, which results in a smoother function along that dimension. Conversely, smaller values of γ_m allow for greater flexibility and variability in the GP realizations with respect to the corresponding predictor.

For the linear fixed effects, we assign the widely used noninformative flat prior

$$\pi(\boldsymbol{\beta}) \propto 1,$$

We assume gamma priors for the variance parameters λ_f and σ^2 as

$$\pi(\lambda_f) = \mathcal{G}amma(a_\lambda, b_\lambda) \quad \text{and} \quad \pi(\sigma^{-2}) = \mathcal{G}amma(a_\sigma, b_\sigma),$$

where a_λ and a_σ denote the shape parameters, and b_λ and b_σ are the rate parameters. We set $a_\lambda = 100$, $b_\lambda = 1$, $a_\sigma = 0.001$, and $b_\sigma = 0.001$.

In the context of abalone growth prediction, identifying the importance of growth-related farming environment factors is of paramount importance. To achieve this, we utilize the following spike-and-slab prior for the factor weights γ_m, which enables variable selection by distinguishing between relevant and irrelevant factors. The prior is defined as follows:

$$\pi(\gamma_m \mid \omega, a_\gamma, b_\gamma) = \omega\delta_0(\gamma_m) + (1-\omega)g(\gamma_m \mid a_\gamma, b_\gamma),\ m = 1, \ldots, M, \tag{4}$$

where $\delta_0(\cdot)$ denotes a Dirac measure with point mass at 0 and ω represents the probability that $\gamma_m = 0$. The mixing parameter ω is assigned the Jeffreys' prior distribution (see, e.g., Gelman et al., 2014 [12]; Murphy, 2022 [13]; Berger et al., 2024 [18]) over the interval $(0, 1)$, ensuring a non-informative prior that allows the model to adaptively determine which variables are important. The slab component $g(\cdot \mid a_\gamma, b_\gamma)$ is modeled as an inverse uniform distribution on $\mathbb{R}^+$, defined as

$$g(\gamma_m \mid a_\gamma, b_\gamma) \propto \frac{1}{b_\gamma - a_\gamma} I(a_\gamma < \gamma_m^{-1} < b_\gamma), \tag{5}$$

where $I(\cdot)$ is an indicator function. The hyperparameters a_γ and b_γ are set such that the mean of γ_m is 5, and the variance is 5, implying that the model assumes the importance of most variables to be moderate on average, while allowing for a sufficiently wide range of variability. This configuration strikes a balance between emphasizing variables with significant contributions and excluding irrelevant ones, enabling the model to adapt flexibly to the data.

In this prior, the mixing probability ω plays a critical role by balancing the spike-and-slab components. When ω is high, the variable is likely irrelevant, and its weight is set to zero. Conversely, a low ω indicates that the variable contributes significantly to the prediction model, and its weight is sampled from the slab distribution. By incorporating the spike-and-slab prior, the model simultaneously achieves both variable selection and parameter estimation in a unified framework. This ensures that only the most influential environmental factors, such as dissolved oxygen, salinity, and nutrient supply, are included in the final predictive model, improving both interpretability and predictive accuracy.

5. Results

In this section, we present the results of analyzing the growth data of indoor-cultured abalone, as introduced in Section 4.1. We begin by applying the proposed weighted Bayesian Kernel Machine Regression (WBKMR) to the data and discussing the findings. Then, we assess the effectiveness of the model by comparing its predictive performance

with that of the well-known BKMR model. The BKMR model assumes equal variance of errors. In all experiments, we draw 10,000 samples from the Markov chain Monte Carlo (MCMC) algorithm (Appendix A), discarding the first 5000 as burn-in, and use the remaining 5000 as posterior samples for inference. The experiments are conducted on an iMac Pro (Apple Inc., Cupertino, CA, USA) with 128 GB of 2666 MHz DDR4 memory and a 3 GHz 10-Core Intel Xeon W CPU.

Before presenting the results of the data analysis, we first examine the convergence of the MCMC algorithm using trace plots for selected parameters, as shown in Figure 4. The trace plots indicate that the MCMC chain has reached convergence, as evidenced by the stable fluctuations around a central value with no clear trends. This suggests good mixing and effective exploration of the parameter space.

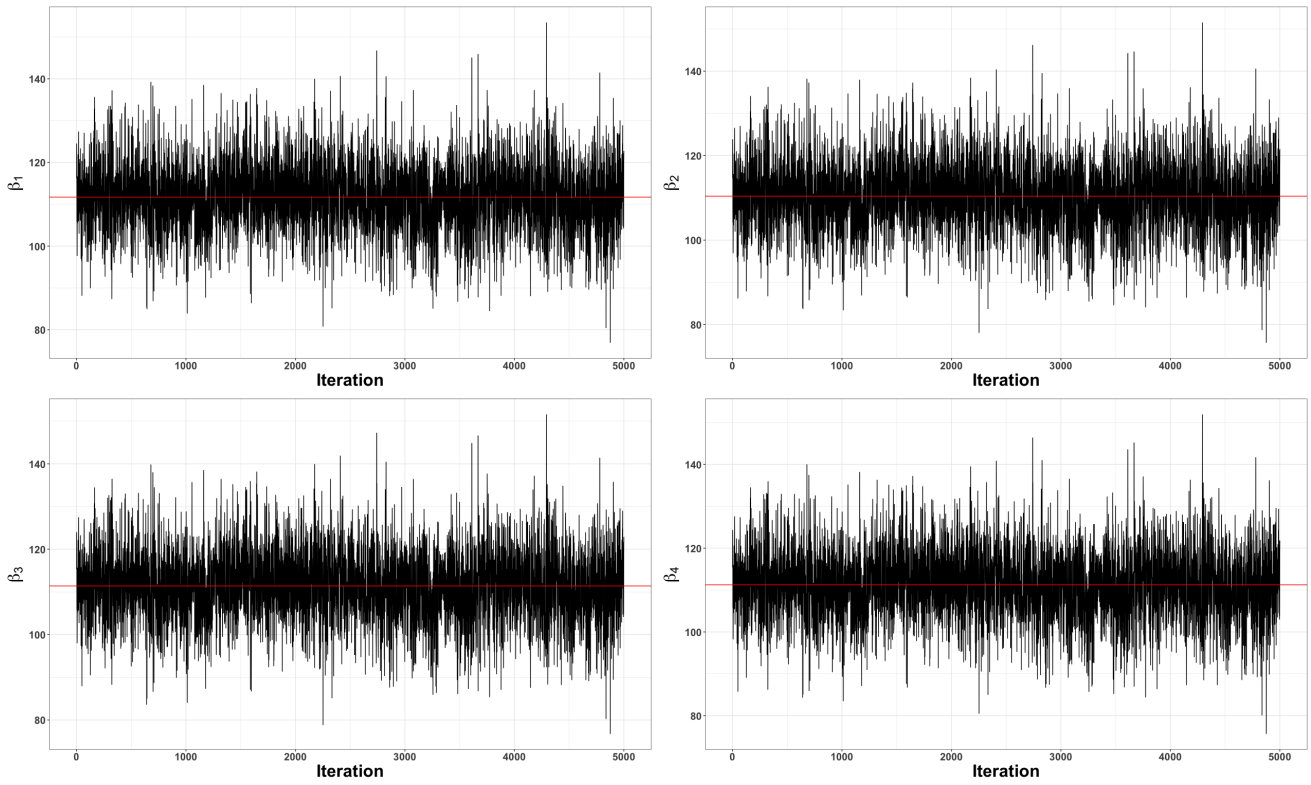


Figure 4. Trace plots of the linear effects $\boldsymbol{\beta}$ for checking the convergence of the MCMC algorithm.

Below, we present the results of the data analysis. We start by examining the importance of key environmental factors in predicting abalone growth. Then, we provide the estimated effects of these factors on abalone growth. Figure 5 provides the Posterior Inclusion Probability (PIP) for each key environmental factor, including time, DO, salinity, NS, pH, and water temperature. The PIP is derived from the spike-and-slab prior, which enables the identification of influential variables by quantifying the probability of each factor being included in the model. As shown in the figure, variables such as time, DO, salt, and NS exhibit high PIP values, underscoring their significant roles in predicting abalone growth. Specifically, the high PIP for time reflects the natural progression of growth over the study period, while the elevated PIP for DO, salt, and NS suggests their influence on the environmental conditions that support abalone growth. In contrast, the low PIP value for water temperature suggests that its influence on growth is minimal within the observed environmental conditions. The intermediate PIP for pH suggests a moderate, yet less pronounced, role compared to the other variables.

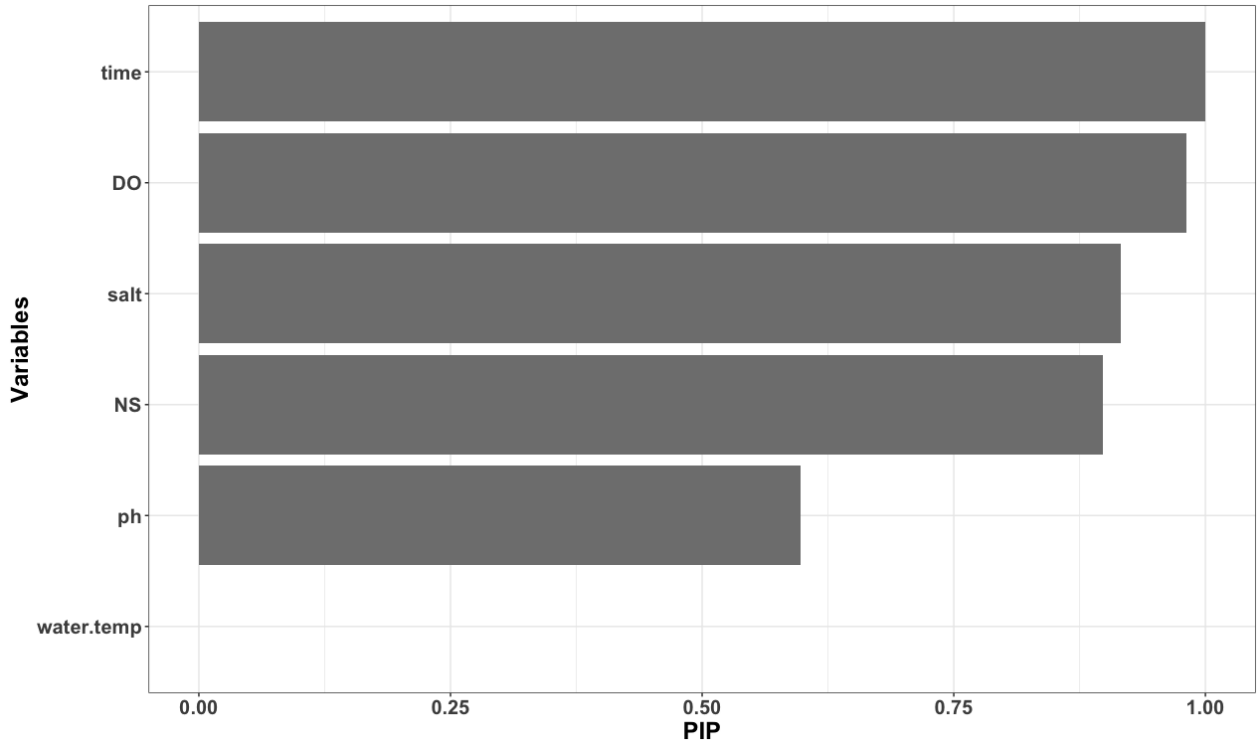


Figure 5. Barplot of Posterior Inclusion Probability (PIP) representing the Importance of key environmental factors.

Figure 6 illustrates the estimated effects of key environmental factors selected based on PIP on abalone growth. The blue line in each plot represents the relationship between each factor and the growth function $f(x)$, while the shaded regions indicate the 95% credible intervals, reflecting the uncertainty in the estimates. The effect of DO shows a negative trend, suggesting that higher DO levels may slightly inhibit growth, though the effect remains uncertain. The influence of pH appears minimal, as evidenced by a nearly flat relationship, indicating a neutral effect on growth. Salt concentration shows a positive effect on growth, whereas nutrient supply has a non-linear relationship, with growth increasing as supply rises but eventually leveling off. Time exhibits a strong positive influence on growth, reflecting the natural progression of abalone growth over the study period.

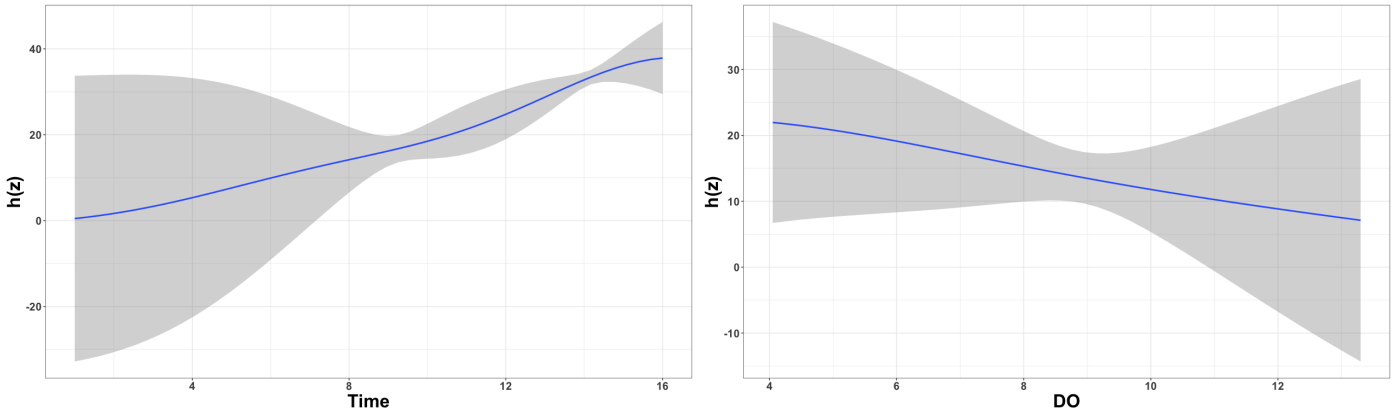


Figure 6. *Cont.*

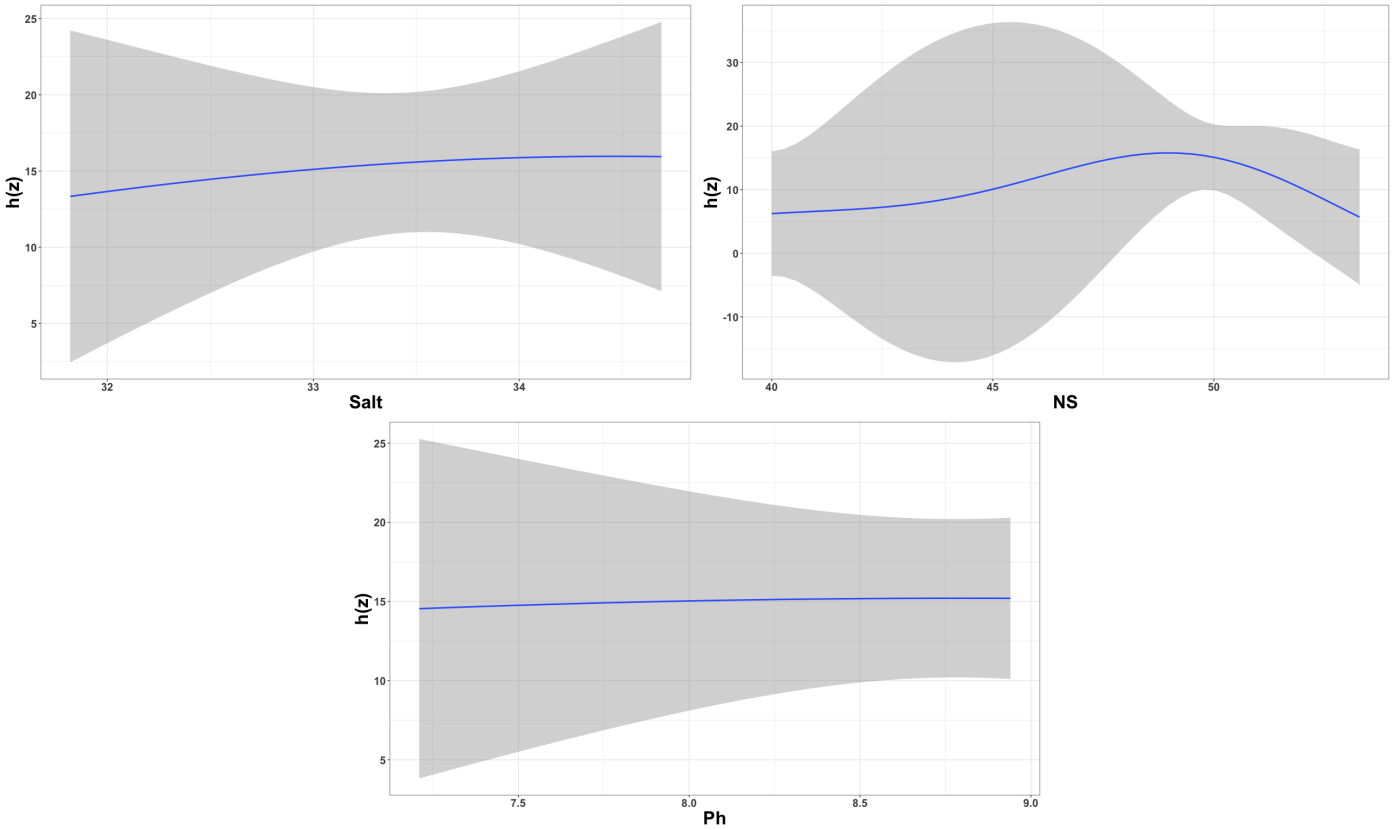


Figure 6. Estimated effects of factors selected based on Posterior Inclusion Probability (PIP) on abalone growth, showing their independent contributions and associated uncertainties (shaded regions).

Next, we present the estimated joint effects of pairs of environmental factors (DO, supply, salt) on abalone growth in Figure 7. In the figure, the first plot on the left shows the combined effect of DO and supply, with red areas indicating higher estimated growth values and blue areas indicating lower growth. This suggests that higher supply levels paired with lower DO values are associated with increased growth. The middle plot displays the interaction between DO and salt, where the highest growth estimates occur at lower DO levels and higher salt concentrations, as indicated by the red gradient. Lastly, the third plot (right) represents the combined effect of salt concentration and supply, with higher growth observed when both salt and supply levels are high. The color gradients in all three plots represent the estimated growth function $f(x)$, with red denoting a higher growth response and blue denoting a lower response.

Finally, we present the results of the prediction performance comparison. To accomplish this, we split the dataset into three training and testing sets. The first training set includes data from 20 March 2015 to 22 October 2016, with the test set consisting of data from 17 December 2016. The second training set includes data up to 17 December 2016, with the test set being 4 March 2017. Lastly, the final training set includes data up to 4 March 2017, while the test set consisted of data from 20 May 2017. Figure 8 shows the first training and test sets.

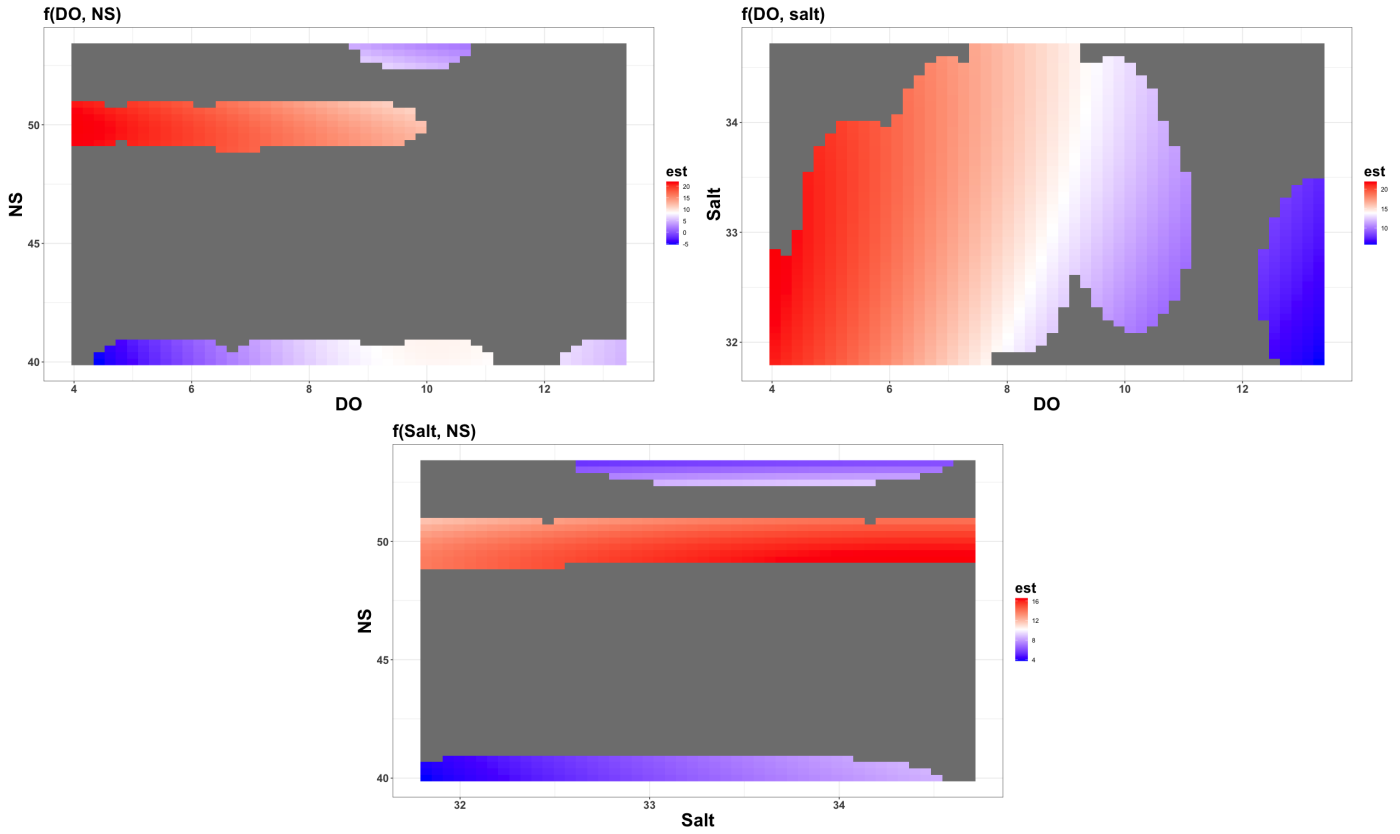


Figure 7. Estimated joint effects of dissolved oxygen (DO), nutrient supply (NS), and salinity (salt) on abalone growth, highlighting their combined influence on growth patterns.

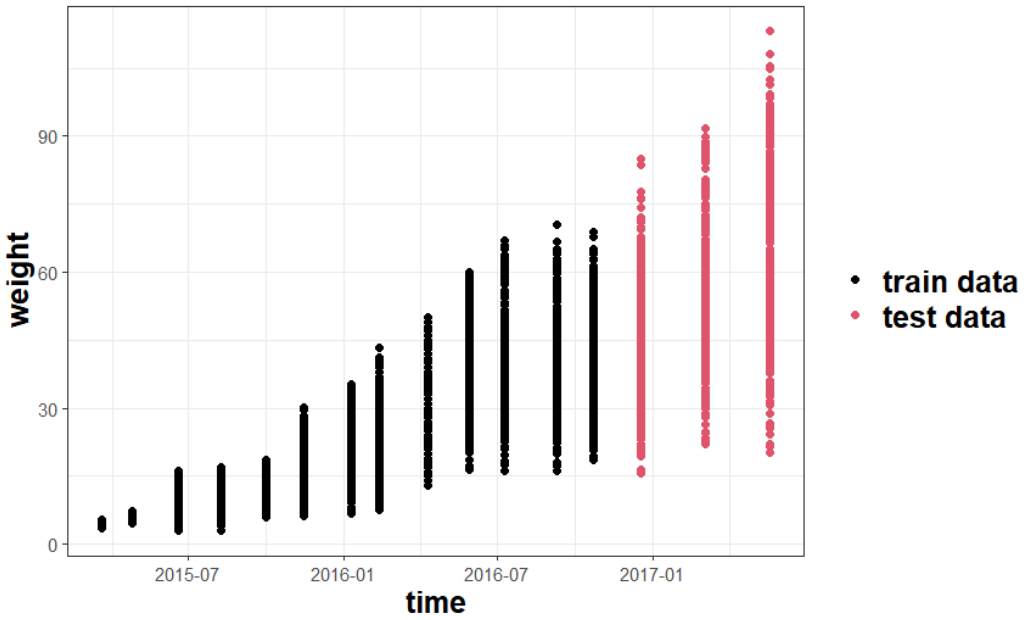


Figure 8. Training and testing sets.

Table 3 presents the root mean square errors (RMSE) for both the training and test sets to compare the predictive performance of two models: WBKMR and the BKMR with equal variance. The table includes results from three different dataset splits (Set 1, Set 2, and Set 3). For the training sets, WBKMR consistently shows lower RMSE values compared to BKMR, indicating that WBKMR provides better fitting to the training data. In the test sets, WBKMR also generally outperforms BKMR, with lower RMSE values

in Sets 2 and 3. However, in Set 1, BKMR shows a slightly lower RMSE than WBKMR, suggesting that WBKMR's performance might vary slightly based on the training data split. Overall, WBKMR demonstrates more reliable predictive performance across the majority of the datasets.

Table 3. Root mean square errors for comparison.

	Training Set		Test Set	
	WBKMR	**BKMR**	**WBKMR**	**BKMR**
Set 1	0.130	0.263	2.938	2.546
Set 2	0.124	0.259	0.949	1.482
Set 3	0.121	0.247	0.569	0.731

6. Discussions

The PIP analysis highlights the critical environmental factors influencing abalone growth, such as time, DO, salinity, and nutrient supply. These factors demonstrate their significant roles in determining growth patterns, as evidenced by their high PIP values. Specifically, the high PIP for time reflects the biological progression of growth over the cultivation period, while the elevated PIP for DO, salinity, and nutrient supply underscores their importance in maintaining optimal environmental conditions. In contrast, the low PIP for water temperature indicates that its influence is minimal under the stable, controlled indoor farming conditions evaluated in this study. The intermediate PIP for pH suggests a secondary role compared to other factors, potentially contributing to maintaining a balanced environment.

The estimated effects further reveal nuanced relationships between environmental factors and abalone growth. For instance, the negative trend observed for DO suggests that higher levels might slightly inhibit growth, potentially due to stress caused by excessive oxygen concentrations. The minimal influence of pH and water temperature aligns with the hypothesis that these variables remain relatively stable in indoor aquaculture systems. Salt concentration and nutrient supply, however, exhibit strong positive or non-linear effects, highlighting their pivotal roles in optimizing growth.

The interaction effects between pairs of environmental factors, as shown in Figure 7, emphasize the complexity of the relationships governing abalone growth. For example, the combined influence of salinity and nutrient supply indicates that simultaneous optimization of these factors could lead to significant improvements in growth. Similarly, the interaction between DO and salinity further underscores the importance of carefully balancing environmental variables to minimize stress while maximizing growth.

Finally, the predictive performance comparison demonstrates the robustness of the WBKMR model. By outperforming the BKMR model in most scenarios, WBKMR proves its capability in handling heteroscedasticity and capturing complex, non-linear relationships. The observed variation in performance across different data splits suggests that the model's effectiveness is influenced by the specific characteristics of the training data. These findings indicate the potential of WBKMR as a reliable tool for both predicting growth rates and identifying critical environmental factors in aquaculture settings.

7. Conclusions

In this paper, we introduced a probabilistic machine learning approach using a Bayesian kernel machine regression model to assess the growth of indoor-cultured abalone, focusing on key environmental factors such as water temperature, dissolved oxygen, pH, salinity, and nutrient supply. The model employs a spike-and-slab prior to determine the influence of growth-related factors, effectively selecting significant predictors. By incor-

porating heteroscedasticity and leveraging Gaussian processes, the proposed weighted Bayesian kernel machine regression (WBKMR) model with unequal variance provides a novel framework for capturing complex, non-linear relationships among the predictors.

The Posterior Inclusion Probability (PIP) analysis highlighted that factors such as time, dissolved oxygen, and nutrient supply are critical in influencing abalone growth, while interaction analysis highlighted the combined effects of environmental variables, such as DO and nutrient supply. These findings offer actionable insights into the optimal management of environmental conditions for maximizing abalone growth in controlled aquaculture settings.

The comparative analysis demonstrated that WBKMR consistently outperformed the baseline BKMR model, achieving lower root mean square errors (RMSE) across multiple datasets. This indicates the proposed model's superior predictive accuracy and its ability to provide meaningful interpretations of environmental dynamics affecting growth.

While the WBKMR model represents a significant advancement, this study is not without limitations. The model is currently constrained to specific environmental factors and indoor farming conditions. Future work could extend the approach to include additional variables, such as interactions with other species or broader aquaculture practices. Incorporating real-time data collection and expanding the temporal scale could further enhance the model's predictive power and practical applicability, offering valuable tools for sustainable abalone farming.

Author Contributions: Conceptualization: H.-J.J., H.-S.J., H.-K.L. and S.J.; Data curation: S.-W.S., M.-J.C. and Y.-D.O.; Exploratory Data Analysis: G.C.; Analysis: G.C. and S.J.; Writing: S.J. and S.-W.S. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by Korea Institute of Marine Science & Technology Promotion(KIMST) funded by the Ministry of Oceans and Fisheries (RS-2022-KS221673, Big data-based aquaculture productivity improvement technology). Seongil Jo was also supported by INHA UNIVERSITY research grant and the National Research Foundation of Korea (NRF) grant funded by the Korea government (RS-2023-00209229).

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: Authors Seung-Won Seo and Ho-Jin Jung were employed by the company Silicogen Inc. The remaining authors declare that the re-search was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BKMR	Bayesian Kernel Machine Regression
GP	Gaussian Process
MCMC	Markov chain Monte Carlo
KMR	Kernel Machine Regression
PIP	Posterior Inclusion Probability
WBKMR	Weighted Bayesian Kernel Machine Regression

Appendix A. Markov Chain Monte Carlo Algorithm

In this appendix, we provide a Gibbs sampling algorithm for calculating the posterior distributions of the parameters of interest, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \lambda_f, \boldsymbol{\gamma}, \omega)$. The algorithm is modified based on the Markov chain Monte Carlo (MCMC) sampler proposed in Bobb et al. (2015) [5]

and Bobb et al. (2018) [6]. To give the detailed description of the algorithm, we first re-express the spike-and-slab prior in (4) hierarchically using binary latent variables as

$$\begin{aligned} \gamma_m \mid r_m, a_\gamma, b_\gamma &\sim r_m \delta_0(\gamma_m) + (1 - r_m) g(\gamma_m \mid a_\gamma, b_\gamma), \\ r_m \mid \omega &\sim \mathcal{B}ernoulli(\omega), \\ \omega &\sim \mathcal{B}eta(a_\omega, b_\omega),\ a_\omega = b_\omega = 1/2. \end{aligned} \tag{A1}$$

Then, after integrating over ω and $\boldsymbol{f}$, the joint posterior distribution is proportional to

$$\begin{aligned} \pi(\boldsymbol{\theta} \mid \bar{y}_1, \ldots, \bar{y}_T) \propto\ & \mathcal{N}(\bar{\boldsymbol{y}} \mid \boldsymbol{Z}\boldsymbol{\beta}, \sigma^2 \boldsymbol{\Sigma}) \left\{ \prod_{m=1}^{M} [r_m \delta_0(r_m) + (1 - r_m) g(\gamma_m \mid a_\gamma, b_\gamma)] \right\} \\ & \times \Gamma(\sum_m r_m + a_\omega) \Gamma(M - \sum_m r_m + b_\omega) \\ & \times \mathcal{G}amma(\sigma^{-2} \mid a_\sigma, b_\sigma) \mathcal{G}amma(\lambda_f \mid a_\lambda, b_\lambda), \end{aligned}$$

where $\bar{\boldsymbol{y}} = (\bar{y}_1, \ldots, \bar{y}_T)^\top$, $\boldsymbol{Z} = (\boldsymbol{z}_1^\top, \ldots, \boldsymbol{z}_T^\top)^\top$ is the design matrix for linear effects, and $\boldsymbol{\Sigma} = \text{diag}(1/n_1, \ldots, 1/n_T) + \lambda_f K$. The algorithm proceeds proceeds iteratively by sampling from the full conditional posterior distributions of each parameter. Below, we outline the main steps of the Gibbs sampling procedure:

1. Set initial values for all parameters in $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \lambda_f, \boldsymbol{\gamma}, \omega)$.
2. For each iteration, sample from the following full conditional posterior distributions:
   (a) Step 1: Sample the coefficients for linear fixed effects, $\boldsymbol{\beta}$, from the full conditional distribution
   $$\boldsymbol{\beta} \mid \sigma^2, \lambda_f, \boldsymbol{r}, \bar{\boldsymbol{y}} \sim \mathcal{N}(\boldsymbol{\beta} \mid \boldsymbol{\Sigma}_n \boldsymbol{Z}^\top \boldsymbol{\Sigma}^{-1} \bar{\boldsymbol{y}}, \sigma^2 \boldsymbol{\Sigma}_n),$$
   where $\boldsymbol{\Sigma}_n = (\boldsymbol{Z}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{Z})^{-1}$.
   (b) Step 2: Sample the variance, σ^2, from the full conditional distribution given by
   $$\sigma^{-2} \mid \boldsymbol{\beta}, \lambda_f, \boldsymbol{r}, \bar{\boldsymbol{y}} \sim \mathcal{G}amma(\sigma^{-2} \mid a_\sigma + n/2, b_\sigma + SSE/2),$$
   where $SSE = (\bar{\boldsymbol{y}} - \boldsymbol{Z}\boldsymbol{\beta})^\top \boldsymbol{\Sigma}_n^{-1} (\bar{\boldsymbol{y}} - \boldsymbol{Z}\boldsymbol{\beta})$.
   (c) Step 3: Sample λ_f using a Metropolis-Hastings method because the full conditional distribution does not the closed form and is proportional to
   $$\pi(\lambda_f \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{r}, \bar{\boldsymbol{y}}) \propto |\boldsymbol{\Sigma}|^{-1/2} \exp\Big\{ -SSE/(2\sigma^2) \Big\} \lambda_f^{a_\lambda - 1} \exp\{-b_\lambda/\sigma^2\}.$$
   Specifically, we generate a candidate sample from a gamma distribution, where the mean is set to the value of λ_f from the previous iteration, and the variance is adjusted to achieve an appropriate acceptance rate.
   (d) Step 4: Sample $(\boldsymbol{\gamma}, \boldsymbol{r})$ using an adaptive Metropolis-Hastings algorithm jointly from the following distribution
   $$\begin{aligned} \pi(\boldsymbol{\gamma}, \boldsymbol{r} \mid \boldsymbol{\beta}, \sigma^2, \lambda_f, \bar{\boldsymbol{y}}) \propto\ & \Gamma(\sum_m r_m + a_\omega) \Gamma(M - \sum_m r_m + b_\omega) \\ & \left\{ \prod_{m=1}^{M} [r_m \delta_0(r_m) + (1 - r_m) g(\gamma_m \mid a_\gamma, b_\gamma)] \right\}. \end{aligned}$$
   For more details, see Bobb et al. (2015) [5].

References

1. Korea National Statistical Office. KOSIS Statistical DB. 2018. Available online: http://kosis.kr/index/index.do (accessed on 11 March 2018).
2. Won, K.; Kim, B.; Jin, Y.; Park, Y.; Son, M.; Cho, M.; Park, M.; Park, M. Infestation of the Abalone, Haliotis discus hannai, by the Polydora under Intensive Culture Conditions in Korea. *J. Fish Pathol.* **2013**, *26*, 139–148. [CrossRef]
3. Kim, B.; Park, M.; Son, M.; Kim, T.; Myeong, J.; Cho, J. A Study on the Optimum Stocking Density of the Juvenile Abalone, Hailotis discus hannai Net Cage Culture or Indoor Tank Culture. *Korean J. Malacol.* **2013**, *29*, 189–195. [CrossRef]
4. Vosloo, D.; van Rensburg, L.; Vosloo, A. Oxidative stress in abalone: The role of temperature, oxygen and L-proline supplementation. *Aquaculture* **2013**, *416–417*, 265–271. [CrossRef]
5. Bobb, J.F.; Valeri, L.; Claus Henn, B.; Christiani, D.C.; Wright, R.O.; Mazumdar, M. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics* **2015**, *16*, 493–508. [CrossRef] [PubMed]
6. Bobb, J.F.; Henn, B.C.; Valeri, L.; Coull, B.A. Statistical software for analyzing the health effects of multiple concurrent exposures via Bayesian kernel machine regression. *Environ. Health* **2018**, *17*, 67. [CrossRef] [PubMed]
7. Williams, C.K.; Rasmussen, C.E. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006; Volume 2.
8. Helidoniotis, F. Growth of Abalone (*Haliotis rubra*) with Implications for Its Productivity. Ph.D. Thesis, University of Tasmania, Hobart, TAS, Australia, 2011.
9. Jabeen, K.; Ahamed, K.I. Abalone Age Prediction using Artificial Neural Network. *IOSR J. Comput. Eng.* **2016**, *18*, 34–38. [CrossRef]
10. Misman, M.F.; Samah, A.A.; Aziz, N.A.A.; Majid, H.A.; Shah, Z.A.; Hashim, H.; Harun, M.F. Prediction of Abalone Age Using Regression-Based Neural Network. In Proceedings of the 2019 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS), Ipoh, Malaysia, 19 September 2019; pp. 23–28. [CrossRef]
11. Khiem, N.M.; Takahashi, Y.; Masumura, T.; Kotake, G.; Yasuma, H.; Kimura, N. A machine learning ensemble approach for predicting growth of abalone reared in land-based aquaculture in Hokkaido, Japan. *Aquac. Eng.* **2023**, *103*, 102372. [CrossRef]
12. Gelman, A.; Carlin, J.B.; Stern, H.S.; Dunson, D.B.; Vehtari, A.; Rubin, D.B. *Bayesian Data Analysis*, 3rd ed.; CRC Press: Boca Raton, FL, USA, 2014.
13. Murphy, K.P. *Probabilistic Machine Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2022.
14. Hu, Z.; Dey, D.K. Generalized variable selection algorithms for Gaussian process models by LASSO-Like penalty. *J. Comput. Graph. Stat.* **2023**, *33*, 477–486. [CrossRef]
15. Mercer, J. Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. R. Soc. A* **1909**, *209*, 415–446.
16. Liu, D.; Lin, X.; Ghosh, D. Semiparametric regression of multidimensional genetic pathway data: Least squares kernel machines and linear mixed models. *Biometrics* **2007**, *63*, 1079–1088. [CrossRef]
17. Oh, Y.; Sun, S.; Lee, K.; Lim, H. Growth and survival of purebred and hybrid according to intraspecific hybridization between Haliotis discus hannai and H. discus discus. *Korean J. Malacol.* **2018**, *34*, 31–41. [CrossRef]
18. Berger, J.O.; Bernardo, J.M.; Sun, D. *Objective Bayesian Inference*; World Scientific: Singapore, 2024.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.