

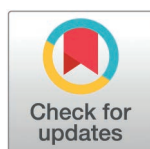
RESEARCH ARTICLE

Comparing efficiency of literature-based estimation of general biomass using Length-Weight Relationships (LWRs) between freshwater and seawater fishes

Jaehoon Yeom  Sang Don Kim *

1 School of Environment and Energy Engineering, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea

* sdkim@gist.ac.kr



OPEN ACCESS

Citation: Yeom J, Kim SD (2025) Comparing efficiency of literature-based estimation of general biomass using Length-Weight Relationships (LWRs) between freshwater and seawater fishes. PLoS One 20(6): e0321571. <https://doi.org/10.1371/journal.pone.0321571>

Editor: Abir Ishtiaq, The Islamia University of Bahawalpur Pakistan, PAKISTAN

Received: September 25, 2024

Accepted: March 7, 2025

Published: June 3, 2025

Copyright: © 2025 Yeom, Kim. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data availability statement: "All relevant data are within the paper and its [Supporting Information](#) files."

Funding: Korea Environmental Industry and Technology Institute (KEITI) through the Aquatic Ecosystem Health Technology Development Project, funded by the Korean Ministry of Environment (MOE) (No. RS-2022-KE002211).

Abstract

Fish biomass estimation is crucial for understanding aquatic ecosystem dynamics and managing fisheries resources effectively. This study evaluated the applicability of estimating fish biomass in Korean aquatic ecosystems using literature-derived General Total Length (GTL) and length–weight relationships (LWRs). We compared the estimated biomasses with measured field biomasses from literatures for freshwater and seawater fish species. Data on individual biomass, habitat, and length-weight coefficients were collected from the literature and databases for 245 fish species. Biomass was estimated using the GTL and representative LWR coefficients and then they were compared to the measured field biomasses from literature. The results showed that this estimation method was applicable only to freshwater species ($R^2 = 0.7133$), whereas the estimates for seawater species showed a poor correlation (negative R^2 values). The removal of outliers ($Q > 6$) improved the estimation accuracy for the freshwater species. This study demonstrates that literature-based biomass estimation using GTL and LWRs is appropriate for freshwater fish in Korea, but not for seawater species. These findings contribute to the generation of fundamental biomass data for ecosystem modeling and highlight the need for habitat-specific approaches for biomass estimation. Since habitat specific biomass data is deficient, future research should explore biomass estimation through species extrapolation to address data gaps in aquatic ecological studies.

Introduction

Recently, concerns have arisen regarding the aquatic ecology of freshwater and saltwater ecosystems, with respect to global biodiversity. Aquatic ecosystems, including freshwater and saltwater, act as habitats for approximately 20% of all species on Earth [1–3]. Therefore, aquatic environments have long been the primary focus of ecological studies. Consequently, research on aquatic ecosystems has been continuously conducted over extended periods across various regions, primarily focusing on ecosystem monitoring, community structure analyses, life cycle studies, and ecosystem changes [4,5]. In addition, progress in ecological

Competing interests: The authors have declared that no competing interests exist.

modeling has encouraged scientists to analyze ecological phenomena using mathematical and simulation-based systems [6–8].

Recent advancements in machine learning techniques have led scientists to explore the application of machine learning in ecological modeling [9]. Despite advancements in methodologies, limited accessibility to essential information and parameters for ecological research has resulted in data gaps in several regions. This hinders the development of ecological models based on big data and their applications in real-world environments [10–13]. Supplementation of insufficient ecological datasets and reliable processing of existing databases is crucial for advancing long-term aquatic ecosystem research. Therefore, to promote the development of long-term studies on aquatic ecosystems, it is essential to address the gaps in ecological data while ensuring the accurate and trustworthy handling of existing datasets [14].

Biomass is a fundamental parameter used in various ecological modeling studies. Therefore, accurate information of the biomass of organisms living within ecosystems is essential in many areas of ecological research [15–17]. However, the complex structure of aquatic ecosystems makes it difficult to estimate the biomass of the organisms living within different habitats. The biomass of organisms in an ecosystem is typically estimated by either inferring the biomass from compiled environmental data or converting easily measurable parameters into biomass [15,18,19]. The biomass of aquatic organisms in an aquatic ecosystem is typically estimated through a regression analysis between body length and weight, known as the length–weight relationships (LWRs) [15,20–22]. LWRs can be widely applied to aquatic invertebrates and fish, and estimations using this regression method are known to be highly accurate [15,21–24]. Therefore, when the direct measurement of mass is not possible, using LWRs to approximate the biomass of a species is a suitable approach. Therefore, estimating the weight of species that are difficult to measure directly through the proper application of length–weight relationships is a crucial methodology in ecosystem research. However, LWRs are not universally applicable to all species, and LWR coefficients can vary depending on the age, sex, maturity, food availability, parasitism and condition factor, leading to different estimates for the same species [20,25–29]. This is attributed to nutritional changes in fish caused by seasonal variations and different life stages [27,30]. In general, these differences can be evaluated using condition factors [20,31].

The studied area, Korea is surrounded by the sea on three sides and has four major rivers running through the country. As a result, both freshwater and saltwater regions coexist in Korea, with estuarine areas forming downstream of freshwater systems. Owing to these geographical characteristics, many fish species in Korean aquatic ecosystems inhabit both freshwater and seawater, alternating between the two environments. In addition, some species live exclusively in freshwater or seawater [32,33]. Fish species in Korea are distributed across various habitats, including freshwater, seawater, estuaries, and mudflats, and each species occupies a distinct environment. The habitats of these species have been identified in various preliminary aquatic ecology studies conducted in Korea. Therefore, when studying Korean aquatic ecosystems including LWRs research, it is crucial to account for these environmental factors, because Korea hosts a wide range of coexisting aquatic environments [34,35].

Although there is a lack of comprehensive information regarding the coefficients of length–weight relationships (LWRs) of fish species living in Korea, LWRs of freshwater and saltwater fish species in the Korean aquatic environment have been analyzed in several studies [36,37]. However, comparative studies of LWR between fish species known to inhabit freshwater and saltwater have not been conducted much. In addition, Korean ecologists and fishery scientists have collected quantified samples of numerous species, documented the lengths of fish species in web databases, and compiled field guides since the 1980s. Therefore, the generalized body length of fish species can be assessed using these literature sources [38,39].

This study focused on developing a methodology to estimate the generalized biomass of individual fish using length-weight relationships for fish species inhabiting Korea, thereby contributing to fields such as aquatic ecological modelling.

Method

Collecting monitored individual biomass of fish species in aquatic ecosystem in Korea including freshwater and seawater

Recorded individual biomass data for each species living in Korean aquatic ecosystems were collected from the literature on Korean aquatic ecosystems, including research articles, theses, and domestic conference papers (Fig 1). All datasets from the literature were included without any limitations on the investigation period, site, or species. However, monitoring data from the sea far from the mainland is excluded.

Biomass data from the collected literature are typically reported as the number density and mass density. To calculate the individual biomass of each species, biomass density was divided by number density using the following formula:

$$\text{Individual biomass (g)} = \frac{\text{Biomass density (g / m}^2\text{)}}{\text{number density (/m}^2\text{)}} \quad (1)$$

The individual biomass for each species was calculated separately for each dataset organized by month. Individual biomass values calculated for each species were recorded as separate data points.

Evaluating general total body length (GTL) and habitat of each species by using literature studies

General Total Body Length (GTL) was used as the representative body length for each species and was calculated based on averages determined by experts, field guides, or reliable

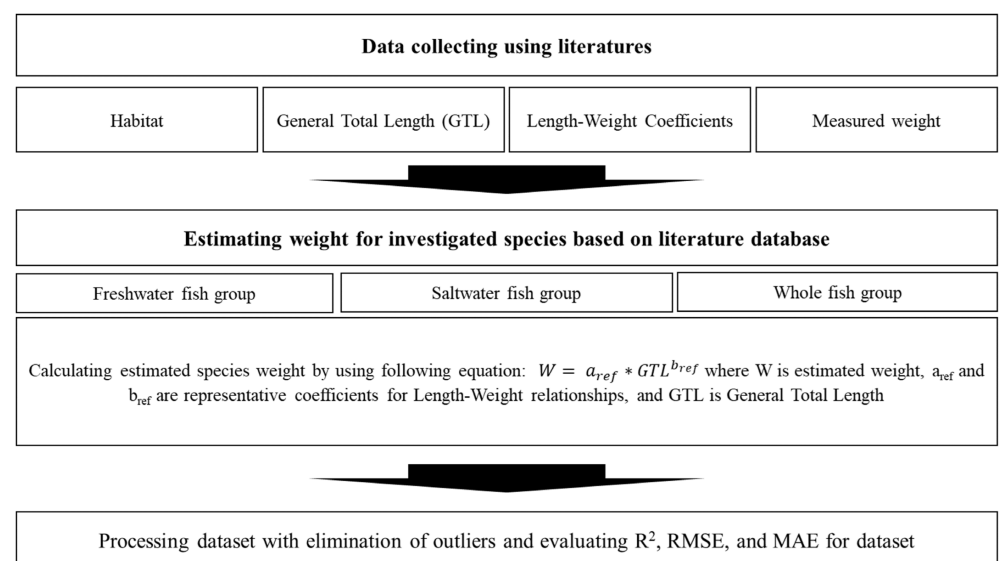


Fig 1. Scheme of methodology for whole study process.

<https://doi.org/10.1371/journal.pone.0321571.g001>

databases. This study assumed that GTL represents the average length of all individuals of a species inhabiting Korean freshwater ecosystems.

The GTL for each species was derived from literature sources, including web databases, such as FishBase, online government databases, and field guides [40]. The GTL was calculated through a series of steps.

First, meta-web databases such as FishBase were used to estimate the GTL for the target species [40]. For species found in Korea, GTL was calculated by averaging the minimum and maximum lengths recorded in the datasets specifically related to the Korean region. Datasets with small sample sizes ($n < 20$) were excluded from analysis.

Second, for species lacking data in the meta-databases, GTL was determined using online and offline field guides that provided general species information.

The habitats of the target species were assessed by reviewing the literature, including government publications, domestic field guides, and online databases. The habitats of all the fish species mentioned in the studies were categorized into four groups: freshwater, seawater, mudflats, and estuaries. Species occupying multiple habitats were also recorded. This study aimed to identify species restricted to freshwater, species restricted to seawater, and species that inhabited mixed habitats.

Selecting freshwater and seawater fish species and collecting length-weight relationship coefficients

Species that inhabit freshwater and seawater were selected from the database. The species were classified into three groups: 'Only Freshwater,' 'Only Seawater,' and a combined group labeled 'Whole Species,' which included both freshwater and oceanic species. The length-weight coefficients were collected for each group using a web database.

Because the target species were based on biomass studies conducted in Korea, we utilized length-weight coefficients from studies that focused on species inhabiting Korean waters. These constants were primarily gathered for data from Korean studies that were used in the FishBase web database. When multiple constants were available for a species from the studies conducted in Korea, a representative constant was calculated for each species. The representative coefficients a and b for the LWRs were calculated using the following formulae:

$$\text{representative } a = \text{Geometric mean of } a = \sqrt[n]{\prod_{k=1}^n a_k} \quad (2)$$

$$\text{representative } b = \text{Arithmetic mean of } b = \frac{\sum_{k=1}^n b_k}{n} \quad (3)$$

Biomass estimation and accuracy assessment

The representative weight of each species was calculated using the Length-Weight Relationship (LWR) coefficients and the General Total Length (GTL) recorded for each fish species. The estimated weight of each species was derived using the corresponding length-weight equations.

The concordance of the estimated general biomass of each species was evaluated using the coefficient of determination (R^2) for the $y = x$ regression, in which the field biomass measured dataset (x) from literatures (x) was compared with the estimated dataset (y). The equation for

the coefficient of determination is defined as $R^2 = 1 - \frac{\sum_{i=1}^n (y - \hat{y})^2}{\sum_{i=1}^n (y - \bar{y})^2}$, where y represents the

recorded values of individual biomasses from literatures for each species, $\hat{y}$ is the estimated value, and $\bar{y}$ is mean of the field measured biomasses from literatures. The coefficient of determination (R^2) quantifies the proportion of variance in the estimated dataset that matches the field biomass dataset (x) from literature for an individual biomass.

In the regression analysis, the outliers were excluded using the Q value, which was calculated based on the residuals between the recorded data from field and estimated data. The Q value is defined as $Q_i = \frac{|y_i \text{ deviation}|}{MedAD}$ where Q_i is the score used to identify outliers, y_i is the monitored biomass for each species, and MedAD is the median absolute difference between the estimated and monitored weights [41]. In this study, data points with $Q > 6$ were excluded to mitigate the influence of outliers.

Cook's distance was used to identify and eliminate influential data points in the regression analysis. Cook's distance is a measure used in regression analysis to identify influential data points. The formula for Cook's distance D_i for the i -th observation is given by

$$D_i = \frac{\sum_{j=1}^n (\hat{y}_j - \hat{y}_{j(i)})^2}{p \cdot MSE} \text{, where } \hat{y}_j \text{ is the predicted value for the } j\text{th observation with all the}$$

data points included, $\hat{y}_{j(i)}$ is the predicted value for the j th observation with the i th data point excluded, p is the number of predictors (including the intercept), and MSE is the mean-squared error of the regression model [42]. Ordinary least squares regression was applied to the dataset after the removal of these outlier points to ensure the accuracy of the analysis.

The calculated representative weight for each species was compared with the actual recorded weight from the field data to determine the method that produced the most accurate estimates. The accuracy of the estimated values was quantitatively compared across different groups using R^2 values.

Results

Results of literature-based monitoring data, habitat classification, and collection of length-weight coefficients

The compiled dataset for the measured field biomasses from literatures of fish species in Korea was obtained from various sources [43–49]. The detailed habitat composition of the compiled dataset is presented in Table 1. Species in the dataset were classified into four categories: species inhabiting only seawater and mudflats; species inhabiting only freshwater; species inhabiting both freshwater and seawater or estuaries; and species with no recorded habitats (Table 1). In this study, species groups living exclusively in seawater and mudflats, species living only in freshwater, and a combined group of all species from both categories were used to evaluate the accuracy of the proposed methodology.

The number of investigated records was 197 for fish species inhabiting seawater and mudflats and 160 for species inhabiting only freshwater (Table 1). However, when considering the availability of length data and length-weight coefficients for each species, the number of usable data points was reduced to 114 for seawater and mudflat species and 131 for freshwater

Table 1. Number of Fish Species in Korea by Habitat Based on Literature Review.

Habitat name	Only seawater fish and mud	Only fresh-water fish	Mixed area	No record	Total
Number of species	51	55	48	6	160
Number of data point in field data	197	160	167	11	535

<https://doi.org/10.1371/journal.pone.0321571.t001>

species (Table 2 in [S1 Appendix](#)). The number of data points for each species within each dataset is provided in the Supporting information.

The distributions of individual biomass for 131 freshwater, 114 seawater, and 245 species from the database are presented as box plots ([Fig 2](#)). Since biomass values range widely, from less than 1 g for small species to over 1000 g for larger species such as carp, applying a simple box plot would result in skewness due to scale differences. Therefore, a logarithmic scale was applied to the boxplots ([Fig 2](#)). To detect outliers within each dataset, the 1.5 times the interquartile range (IQR) method was used to detect outliers in each dataset. No outliers were found in any of the datasets when this criterion was used ([Fig 2](#)). The range of field measured weights from literature for each dataset is 0.1299 g/ind. to 2024.4 g/ind. for freshwater species, and 0.0471 g/ind. to 143.3 g/ind. for seawater species and 0.0471 g/ind. to 2024.4 g/ind. in the combined species groups ([Fig 2](#)).

The length-weight coefficients for each species were collected from FishBase [40], with only data from studies conducted in Korea on accurately identified species. The length-weight coefficients collected for each species are presented in tabular form (Table 1 in [S1 Appendix](#)).

The General Total Length (GTL) representing each species were recorded based on data from FishBase, online databases, and field guides. For each species, length was recorded as either the average of the minimum and maximum lengths reported from these sources or the explicitly stated average length.

Data processing and analysis of concordance between estimated and measured field weight from literature

The results presented in [Fig 3](#) illustrate the comparison between the estimated and measured field biomasses from literatures for freshwater and seawater species with each data point analyzed ([Fig 3](#)). The estimated weights, calculated using the GTL and representative constants

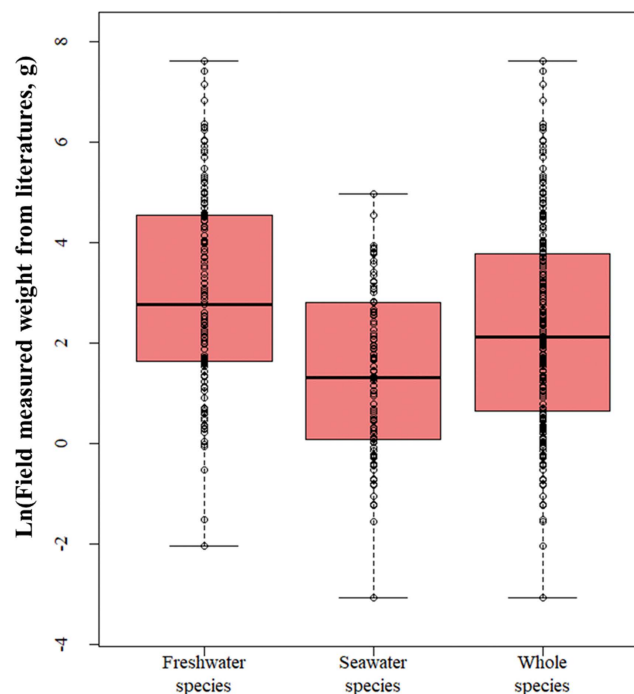


Fig 2. Box plot for Ln- recorded field weight of grouped species.

<https://doi.org/10.1371/journal.pone.0321571.g002>

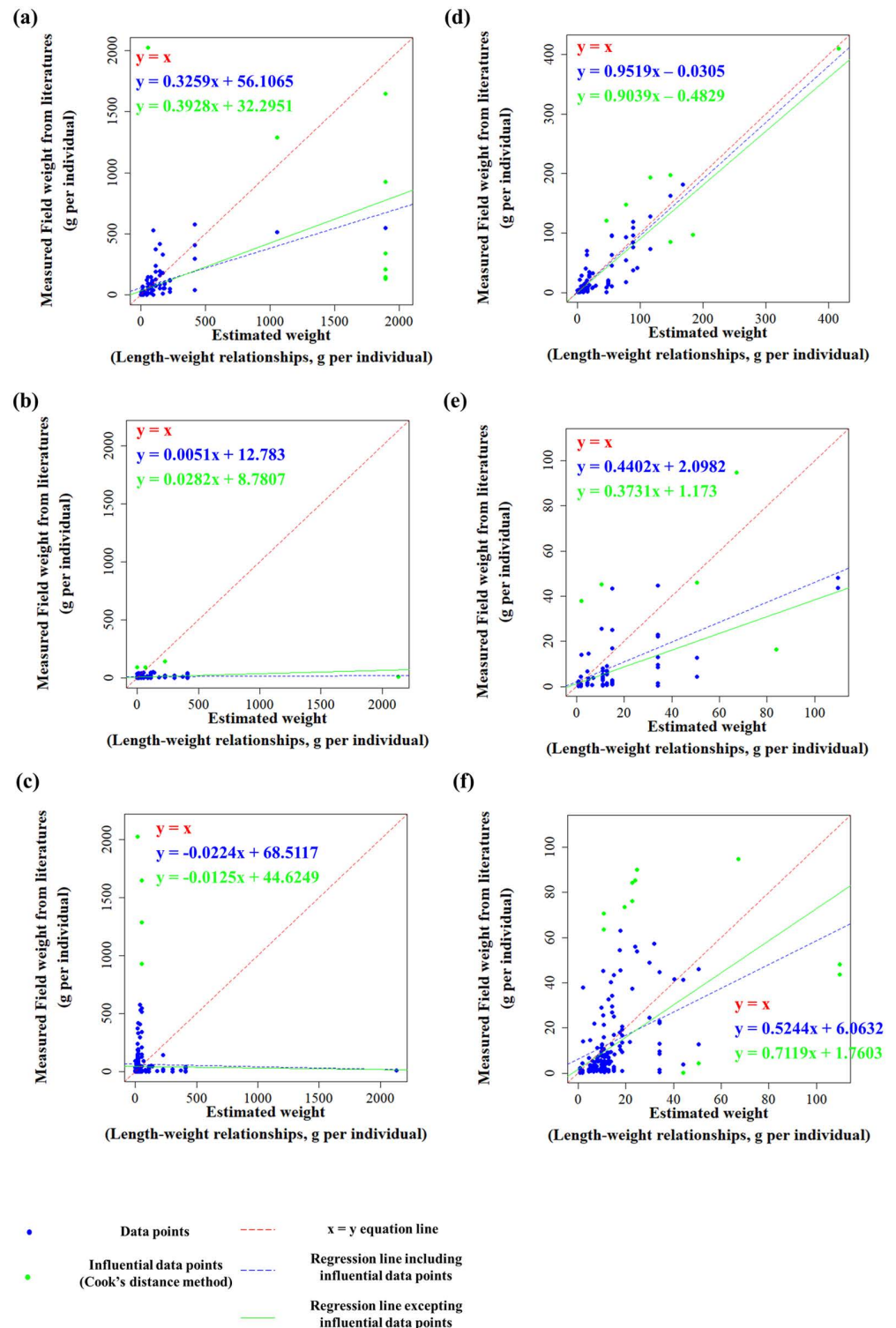


Fig 3. Graph plotting for dataset group (freshwater, seawater, whole species) with outlier processing. (a) Targeted only freshwater species with no exception on outlier. (b) Targeted only seawater species with no exception on outlier. (c) Targeted whole species with no exception on outlier. (d) Targeted only freshwater species without outlier with $Q > 6$. (e) Targeted only seawater species without outlier with $Q > 6$. (f). Targeted whole species without outlier with $Q > 6$.

<https://doi.org/10.1371/journal.pone.0321571.g003>

of LWRs from the literature, are plotted on the x-axis, whereas the field weights reported in the literature are plotted on the y-axis. A Q value was calculated for each data point and data with $Q > 6$ were removed from the dataset for analysis. Consequently, each group was divided into two datasets: the original dataset and a modified dataset, excluding data points with $Q > 6$. The analysis was conducted across six datasets categorized by freshwater species, seawater species, and all species combined for the original and $Q > 6$ eliminated datasets.

In each graph, Cook's distance was used to identify the influential data points for trend analysis, and separate trend lines were presented for cases in which these data points were included or excluded. The blue line represents the first-order linear regression line calculated using the entire dataset, including influential data points (Fig 3). The green line represents the first-order linear regression line calculated after removing the influential data points identified by the Cook's distance. In addition, the red dashed line represents the $y = x$ line, which serves as a reference line where the estimated weight equals the measured field biomasses from literature, indicating an ideal alignment between the two values (Fig 3).

In panel (a), for freshwater species without removing the $Q > 6$ values, the slope of the trend line before removing the influential points was 0.3259; after removing them, the slope slightly increased to 0.3928 (Fig 3a). In panel (b), for seawater species without removing $Q > 6$, the slope of the trend line before removing the influential points was 0.0051; after removal, the slope increased slightly to 0.0282 (Fig 3b). In panel (c), for the whole dataset of freshwater and seawater species without removing $Q > 6$ values, the slope of the trend line was -0.0224 before removing the influential points, and after removal, it slightly increased to -0.0125 (Fig 3c).

In contrast, in Panel (d), for freshwater species with $Q > 6$ values removed, the slope of the trend line was 0.9519 before removing the influential points and decreased slightly to 0.9039 after removal (Fig 3d). In panel (e), for seawater species with $Q > 6$ values removed, the slope decreased from 0.4402 to 0.3731 after removing influential points (Fig 3e). In panel (f), for the entire dataset of freshwater and seawater species with $Q > 6$ values removed, the slope of the trend line increased from 0.5244 to 0.7119 after removing influential points (Fig 3e). Overall, the change in slope after removing the influential points varied depending on the dataset.

Analysis of the coefficient of determination (R^2) for each dataset comparing estimated and measured field weight from literatures

In Fig 4, the correlation between the estimated and measured field biomasses from literatures for the six dataset groups, as shown in the chart, was evaluated by calculating the coefficient of determination (R^2) through statistical processing for Group 1, 2, 3, and 4. For Group 1, R^2 represents the value derived by applying a first-order linear regression to the data points (Fig 4). Group 2 presents the R^2 values for the data points with respect to the $y = x$ equation. Group 3 represents the R^2 values after removing influential points, as identified by Cook's distance, and applying first-order linear regression to the remaining data points. Finally, Group 4 shows the R^2 value for the data points with respect to the $y = x$ equation after removing the influential points based on Cook's distance (Fig 4).

For freshwater species, R^2 for the regression of the dataset was 0.2583, and after removing influential points, it increased to 0.4558 (Fig 4). The R^2 for the $y = x$ equation was -0.8989, improving to -0.6644 after removing influential points, indicating an increase in R^2 after the removal of influential points (Fig 4). For the dataset of freshwater species with $Q > 6$ removed, the regression R^2 was 0.8140, decreasing to 0.7292 after removing the influential points, whereas the R^2 for the $y = x$ equation was 0.8109, decreasing to 0.7133 after removing the influential points (Fig 4).

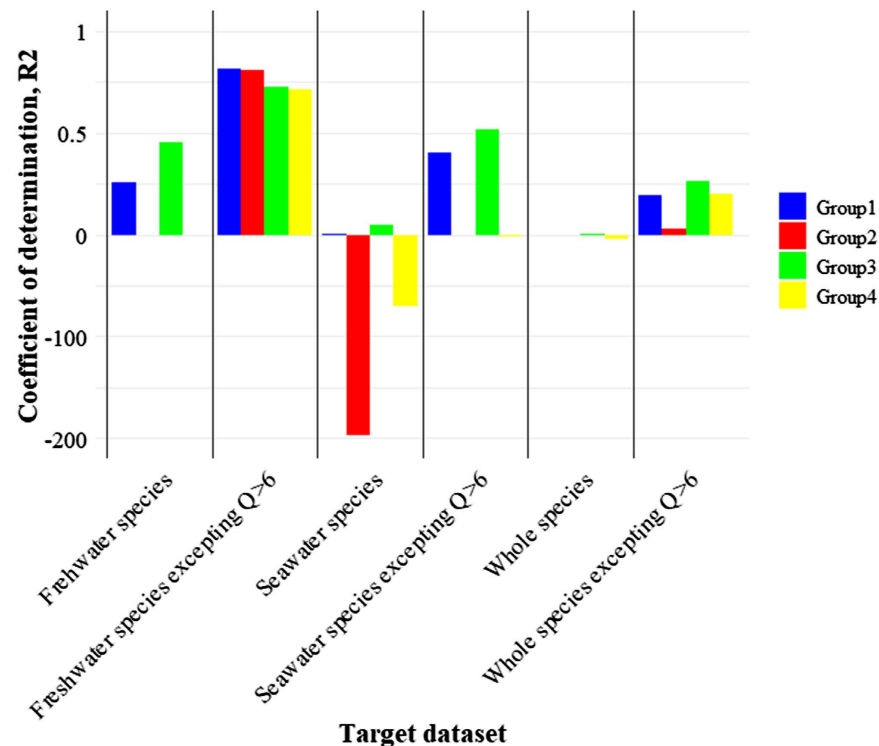


Fig 4. Chart for coefficient of determination (R^2) of each dataset.

<https://doi.org/10.1371/journal.pone.0321571.g004>

For seawater species, the regression R^2 of the dataset was 0.0047, increasing to 0.0474 after removing influential points, while the R^2 for the $y = x$ equation was -197.24, improving to -70.682 after removing influential points (Fig 4). For the dataset of seawater species with $Q > 6$ values removed, the regression R^2 was 0.4049, increasing to 0.5161 after removing influential points, and the R^2 for the $y = x$ equation was -0.4748, decreasing to -1.514 after removing influential points (Fig 4).

For the entire species group, the regression R^2 of the dataset was 0.00005, increasing to 0.0008 after removing the influential points, whereas the R^2 for the $y = x$ equation was -0.98952, further decreasing to -4.959 after removing the influential points (Fig 4). For the entire species dataset with $Q > 6$ values removed, the regression R^2 was 0.1914 and increased to 0.2606 after removing the influential points, whereas the R^2 for the $y = x$ equation was 0.02896 and increased to 0.1994 after removing the influential points (Fig 4).

In summary, the dataset with the highest regression R^2 was the group of freshwater species with $Q > 6$ values removed, and the group with the highest R^2 for the $y = x$ equation was the freshwater species group with $Q > 6$ values removed. In contrast, the seawater species and whole species groups exhibited lower R^2 and R^2 values in the $y = x$ equation. Additionally, the RMSE and MAE values for evaluating the efficiency of the regression model and $y = x$ model for each dataset group are presented in the Supplementary Materials (Table 1 in S1 Appendix).

Discussions

The aim of this study was to assess the applicability of biomass estimates for fish species in Korea by utilizing the GTL derived from literature and representative LWR coefficients. This

study compared the concordance between estimated and actual measured field biomasses from literature for both freshwater and seawater species.

As shown in Fig 3, the effect of outlier removal varies depending on the dataset for each fish group. In some panels, such as (b), (c), and (d), the slope and intercept of the trend line do not change significantly after removing influential data points, whereas in other panels, such as (a), (e), and (f), the slope and intercept change considerably (Fig 3). This suggests that specific data points have a substantial effect on the overall analysis results for certain species. In datasets where the measured field biomasses from literatures or estimated biomasses values exhibited extreme outliers, the regression slope was heavily influenced, and the removal of influential points proved effective [41,42,50].

Comparing the slope and interceptive values across the panels in Fig 3, Panels (a), (b), and (c), which include $Q > 6$ data points, show slopes that deviate significantly from the ideal trend line, $y = x$ (Fig 3a, b, and c). Because Q is calculated based on the absolute value of $y - x$, datasets containing outliers corresponding to high Q values exhibit large errors when the relationship $y = x$ [41]. Consequently, as observed in Panels (d), (e), and (f), removing data points with $Q > 6$ improves the alignment of the regression line with $y = x$.

Although panel (e) for seawater species still does not represent an ideal trend line after removing $Q > 6$ data points, panels (d) and (f) for freshwater and whole species show an improved alignment with the $y = x$ line (Fig 3e, d, and f). In panel (d), the slope of the regression line is 0.9519, which is close to the $y = x$ line, with a small intercept (Fig 3d). However, after removing the influential points to improve data reliability, the slope decreased slightly to 0.9039, moving further from the ideal trend (Fig 3d). In panel (f), the slope before removing the influential points is 0.5244, significantly deviating from 1; however, after their removal, the slope improves to 0.7119, moving closer to 1 (Fig 3f). This increase in the slope towards $y = x$ after removing $Q > 6$ points indicated that the regression line became more aligned with the ideal trend.

The increase in the slope after removing $Q > 6$ values suggests that the removed data points likely had much higher estimated values than the measured field biomasses from literature. These removed points were often small individuals, such as juveniles, whose measured field biomasses from literatures were much smaller than the average, leading to their classification as outliers [20,51,52]. Therefore, the presence of juvenile fish appears to be a significant factor contributing to inaccuracies in general biomass estimation.

In Fig 3, the datasets that display the trend line closest to the $y = x$ line, representing the agreement between the estimated and measured field biomasses from literatures, correspond to the datasets for freshwater species and whole species with $Q > 6$ values removed, as shown in panels (d) and (f) (Fig 3d and f). In contrast, the regression graph for seawater species followed a trend that was difficult to explain using the $y = x$ line, indicating that the estimated biomasses derived from the literature-based GTL and representative LWR coefficients did not match the actual measured field biomasses from literatures for coastal fish species in Korea (Fig 3b and e). Therefore, the biomass estimation method based on the literature-derived GTL and representative LWR coefficients demonstrates a more suitable correlation for freshwater species than for seawater species in Korea [20,34].

In Fig 4, the relationship between the predicted and measured field biomasses from literatures in this study was evaluated using the R^2 value with respect to the $y = x$ line, as this represents the true correspondence between the two values, and indicates the validity of the biomass estimation method. Thus, the R^2 values assessed in Groups 2 and 4 reflect the degree of agreement between the estimated and measured field biomasses from literatures in each dataset, whereas the R^2 values in Groups 1 and 3 represent the consistency of trends within the dataset itself. Among the six datasets analyzed, the dataset for freshwater species with $Q > 6$

values removed was the only one in which the biomass estimation methodology was appropriately applied (Fig 4). In this dataset, the R^2 value for the $y = x$ equation was 0.8109, and after removing the influential points, it decreased slightly to 0.7133; however, the estimates generated by the length-weight relationship still showed a meaningful fit with the biomass recorded in the literature. In comparison, the R^2 value for the $y = x$ equation in the dataset for freshwater species without Q removal was -0.899, and after removing influential points, it improved to -0.664, showing a significant correlation with the $y = x$ line after removing Q outliers (Fig 4).

The reason for this improvement is that, as previously mentioned, the measured field biomasses from literatures for freshwater species in this study included a significant number of juveniles [20]. By removing the outliers corresponding to Q values greater than six, the alignment between the predicted biomass calculated from the general body length and the biomass recorded in the literature improved [41]. Notably, 32 data points (approximately 25% of the total) were excluded because $Q > 6$ (Fig 4). Furthermore, the R^2 values for the datasets with simple linear regression were generally higher than those for the $y = x$ equation across all the datasets, indicating a general correlation between the measured field biomasses from literatures and estimated biomasses. However, the R^2 values for the $y = x$ equation, which represent perfect agreement between the estimated and measured field biomasses from literatures, were very low for all datasets except for that for freshwater species with $Q > 6$ removed (Fig 4). The seawater species dataset produced negative R^2 values, indicating a lower level of agreement than expected from the simple average biomass estimates. The applicability of length-weight coefficients appears to vary across different environments for fish, as demonstrated in several studies [20,34,53].

The study about length-weight relationship of *Mystus tengara* in pond and river habitat in India showed r^2 of length-weight relationship for *Mystus tengara* living in river was higher than *Mystus tengara* living in pond. However, Pearson's correlation between habitat and condition factor shows no significant correlation in same study [54]. Similarly, a study on the length-weight relationship of *Tenualosa ilisha* in Bangladesh rivers (freshwater) and the Bay of Bengal (seawater) revealed no significant differences in b values between samples obtained from riverine and seawater habitats [55]. However, the size range and weight range of the target species were found to be higher in the Bay of Bengal (seawater) compared to the Bangladesh rivers (freshwater) [55]. However, this is a study applied to sampling results of a single species. What can be inferred from this is that the length-weight constants a and b for a single species do not vary significantly according to habitat, but individuals inhabiting seawater may generally exhibit larger sizes [55].

This can explain why the applicability of the standard biomass estimation methodology for freshwater species in this study appears to be higher than that for seawater species. Although the applicability of length-weight constants may not differ significantly between habitats, seawater samples are generally likely to be larger and have a wider size distribution compared to river samples [54,55]. As a result, the General Total Length (GTL) used in this study may be inadequate in representing the wide range of sizes for each species in seawater, even if it was recorded based on rivers or seawater standards for the same species. Therefore, while the applicability of length-weight constants may not differ greatly between rivers and seawater, it appears that the applicability of representative lengths (GTL) for each species differs between seawater and freshwater environments

Therefore, the biomass estimates derived from the GTL-based methodology in this study, which used coefficients for LWRs obtained from the literature on fish species in Korea, were applicable only to freshwater species.

Based on the results of this study, when considering the relationship and coefficient of determination between the estimated and measured field biomasses from literature, it is

evident that applying literature-based biomass estimation for fish species in Korea is appropriate only when the species has been confirmed to inhabit freshwater environments.

Conclusion

The purpose and originality of this study was to estimate the biomass of fish species inhabiting freshwater and seawater in Korea using literature-based General Total Length (GTL) and Length-Weight Relationships (LWRs), and to evaluate the accuracy of the estimates by comparing them with measured field biomasses from literature. The applicability of the general biomass estimates calculated using LWRs and GTL from the literature was compared across different fish groups. When comparing freshwater species, seawater species, and a combined group of both, only the freshwater species group showed a close correlation between the estimated and measured field biomasses from literatures, with R^2 values of 0.8109 before removing influential points and 0.7133 after their removal, demonstrating significant explanatory power. In contrast, the applicability of the seawater species group was minimal with negative R^2 values. Considering the coefficient of determination, the biomass estimates derived from the GTL and LWRs were found to be appropriate only for species that inhabit freshwater in Korea. Based on these results, when performing biomass estimates using data from the literature instead of measured field biomasses from literatures, it is essential to confirm the habitat of the species and limit the estimates to freshwater species to obtain accurate results. This study is expected to contribute to the generation of fundamental data on biomass that can be used in future studies on ecosystem indicators and ecological modeling. Furthermore, this study highlights the need for future research to expand beyond the existing species data and explore biomass estimation through extrapolation among species.

Supporting information

S1 Appendix. Table dataset for supporting information.
(DOCX)

S2 Appendix. Dataset for measured and estimated dry weight.
(XLSX)

Author contributions

Conceptualization: Jaehoon Yeom.

Data curation: Jaehoon Yeom.

Funding acquisition: Sang Don Kim.

Methodology: Jaehoon Yeom.

Supervision: Sang Don Kim.

Writing – original draft: Jaehoon Yeom.

Writing – review & editing: Sang Don Kim.

References

1. Appeltans W, Ah Yong ST, Anderson G, Angel MV, Artois T, Bailly N, et al. The magnitude of global marine species diversity. *Current Biology*. 2012;22(23):2189–202. <https://doi.org/10.1016/j.cub.2012.09.036>
2. Carrete Vega G, Wiens JJ. Why are there so few fish in the sea?. *Proc Biol Sci*. 2012;279(1737):2323–9. <https://doi.org/10.1098/rspb.2012.0075> PMID: 22319126

3. Helfrich L, Neves R, Parkhurst J. Sustaining America's aquatic biodiversity. What is aquatic biodiversity; why is it important?. 2005;420. www.ipbes.net/
4. Collen B, Whitton F, Dyer EE, Baillie JEM, Cumberlidge N, Darwall WRT, et al. Global patterns of freshwater species diversity, threat and endemism. *Glob Ecol Biogeogr*. 2014;23(1):40–51. <https://doi.org/10.1111/geb.12096> PMID: 26430385
5. Huang L, Meng J-N, Xu F, Zhou Y, He G, Wang K, et al. A holistic view of aquatic ecosystems: integrating health and integrity, network, stability, and regime shift assessments. *Int J Sediment Res*. 2023. <https://doi.org/10.1016/j.ijsrc.2023.12.006>
6. Evans MR. Modelling ecological systems in a changing world. *Philos Trans R Soc Lond B Biol Sci*. 2012;367(1586):181–90. <https://doi.org/10.1098/rstb.2011.0172> PMID: 22144381
7. Coll M, Steenbeek J, Pennino MG, Buszowski J, Kaschner K, Lotze HK, et al. Advancing global ecological modeling capabilities to simulate future trajectories of change in marine ecosystems. *Front Mar Sci*. 2020;7. <https://doi.org/10.3389/fmars.2020.567877>
8. Larocque GR. Ecological modelling in the 21st century: examining potential research directions and challenges. *Procedia Environ Sci*. 2012;13:331–9. <https://doi.org/10.1016/j.proenv.2012.01.032>
9. O'Gorman EJ. Machine learning ecological networks. *Science*. 2022;377(6609):918–9. <https://doi.org/10.1126/science.add7563> PMID: 36007050
10. Salguero-Gómez R, Jackson J, Gascoigne SJL. Four key challenges in the open-data revolution. *J Anim Ecol*. 2021;90(9):2000–4. <https://doi.org/10.1111/1365-2656.13567> PMID: 34525215
11. Farley SS, Dawson A, Goring SJ, Williams JW. Situating ecology as a big-data science: current advances, challenges, and solutions. *BioScience*. 2018;68(8):563–76. <https://doi.org/10.1093/biosci/biy068>
12. Hampton SE, Strasser CA, Tewksbury JJ, Gram WK, Budden AE, Batcheller AL, et al. Big data and the future of ecology. *Frontiers in Ecol & Environ*. 2013;11(3):156–62. <https://doi.org/10.1890/120103>
13. Culina A, Adriaensen F, Bailey LD, Burgess MD, Charmantier A, Cole EF, et al. Connecting the data landscape of long-term ecological studies: The SPI-Birds data hub. *J Anim Ecol*. 2021;90(9):2147–60. <https://doi.org/10.1111/1365-2656.13388> PMID: 33205462
14. Augustine SP, Bailey-Marren I, Charton KT, Kiel NG, Peyton MS. Improper data practices erode the quality of global ecological databases and impede the progress of ecological research. *Glob Chang Biol*. 2024;30(1):e17116. <https://doi.org/10.1111/gcb.17116> PMID: 38273575
15. Li D, Hao Y, Duan Y. Nonintrusive methods for biomass estimation in aquaculture with emphasis on fish: a review. *Rev Aqua*. 2019;12(3):1390–411. <https://doi.org/10.1111/raq.12388>
16. Galaiduk R, Radford BT, Harvey ES. Utilizing individual fish biomass and relative abundance models to map environmental niche associations of adult and juvenile targeted fishes. *Sci Rep*. 2018;8(1):9457. <https://doi.org/10.1038/s41598-018-27774-7> PMID: 29930311
17. Moscoso JE, Bianchi D, Stewart AL. Controls and characteristics of biomass quantization in size-structured planktonic ecosystem models. *Ecological Modelling*. 2022;468:109907. <https://doi.org/10.1016/j.ecolmodel.2022.109907>
18. Llopis-Belenguer C, Blasco-Costa I, Balbuena JA. Evaluation of three methods for biomass estimation in small invertebrates, using three large disparate parasite species as model organisms. *Sci Rep*. 2018;8(1):3897. <https://doi.org/10.1038/s41598-018-22304-x> PMID: 29497143
19. Peña H, Macaulay GJ, Ona E, Vatnehol S, Holmin AJ. Estimating individual fish school biomass using digital omnidirectional sonars, applied to mackerel and herring. *ICES J Marine Sci*. 2021;78(3):940–51. <https://doi.org/10.1093/icesjms/fsaa237>
20. Froese R. Cube law, condition factor and weight-length relationships: history, meta-analysis and recommendations. *J Appl Ichthyol*. 2006;22(4):241–53. <https://doi.org/10.1111/j.1439-0426.2006.00805.x>
21. González-Sansón G, Aguilar-Betancourt C, Kosonoy-Aceves D, Lucano-Ramírez G, Ruiz-Ramírez S, Flores-Ortega JR, et al. Weight-length relationships for 38 fish species of Barra de Navidad coastal lagoon, Jalisco, Mexico. *J Appl Ichthyol*. 2013;30(2):428–30. <https://doi.org/10.1111/jai.12327>
22. Benke AC, Huryn AD, Smock LA, Wallace JB. Length-mass relationships for freshwater macroinvertebrates in north america with particular reference to the Southeastern United States. *J North Am Benthol Soc*. 1999;18(3):308–43. <https://doi.org/10.2307/1468447>
23. Smock LA. Relationships between body size and biomass of aquatic insects. *Freshwater Biology*. 1980;10(4):375–83. <https://doi.org/10.1111/j.1365-2427.1980.tb01211.x>
24. Freitas TMS, Dutra GM, Salvador GN. Length-weight relationships of 18 fish species from Paraíba do Sul basin, Minas Gerais, Brazil. *J Appl Ichthyol*. 2017;33:652–54. <https://doi.org/10.1111/jai.13346>

25. Mocq J, Dekanová V, Boukal DS. Length-mass allometries of the larvae of aquatic dipterans: differences between taxa, morphological traits, and methods. *J Insect Sci.* 2024;24(1):10. <https://doi.org/10.1093/jisesa/ieae012> PMID: 38367014
26. Mehanna SF, Farouk AE. Length-weight relationship of 60 fish species from the Eastern Mediterranean Sea, Egypt (GFCM-GSA 26). *Front Mar Sci.* 2021;8. <https://doi.org/10.3389/fmars.2021.625422>
27. Muhammad N, Abir I. Proximate composition of *Mystus bleekeri* in relation to body size and condition factor from Nala Daik, Sialkot, Pakistan. *Afr J Biotechnol.* 2011;10(52):10765–73. <https://doi.org/10.5897/ajb10.2339>
28. Salam A, Naeem M, Kauser S. Weight Length and Condition Factor Relationship of a Fresh Water Wild *Puntius chola* from Islamabad, Pakistan. *Pakistan J of Biol Sci.* 2005;8(8):1112–4. <https://doi.org/10.3923/pjbs.2005.1112.1114>
29. Naeem M, Salam A, Ashraf M, Khalid M, Ishtiaq A. External morphometric study of hatchery reared mahseer (*Tor putitora*) in relation to body size and condition factor. *Afr J Biotechnol.* 2011;10:7071–7. <https://doi.org/10.5897/AJB11.715>
30. Ashraf M, Zafar A, Naeem M. Comparative studies on the seasonal variations in the nutritional values of three carnivorous fish species. *Int J Agric Biol.* 2011;13:1560–8530.
31. Ngodhe SO. Assessment of Length-Weight Relationship and Condition Factor of Nile Tilapia (*Oreochromis niloticus*) in Cage and Open Waters in Winam Gulf of L. Victoria, Kenya. *IJESNR.* 2019;22(3). <https://doi.org/10.19080/ijesnr.2019.22.556088>
32. Lee E, Lee J, Lee C, Choi K. Connectivity and effectiveness of marine protected areas on the west coast of Korea within the Yellow Sea large marine ecosystem. *Ocean Polar Res.* 2022;44:249–60. <https://doi.org/10.4217/OPR.2022019>
33. Atique U, An K-G. Potential risky exotic fish species, their ecological impacts and potential reasons for invasion in Korean aquatic ecosystems. *J Ecol Environ.* 2022;46. <https://doi.org/10.5141/jee.22.008>
34. Chandran R, K Singh R, Singh A, Ganesan K, Thipramalai Thangappan AK, K Lal K, et al. Evaluating the influence of environmental variables on the length-weight relationship and prediction modelling in flathead grey mullet, *Mugil cephalus* Linnaeus, 1758. *Peer J.* 2023;11:e14884. <https://doi.org/10.7717/peerj.14884> PMID: 36860765
35. Kwon Y-S, Shin M-S, Yoon H-N. Relationships between fish communities and environmental variables in islands, South Korea. *Proc National Instit Ecol Republic of Korea.* 2022;3(2):84–96. <https://doi.org/10.22920/PNIE.2022.3.2.84>
36. Yoon H-S. Length-weight relationships for 11 fishes caught by fish pots in the coastal water off Baekya Island, Korea. *Korean J Environ Biol.* 2013;31(2):154–7. <https://doi.org/10.11626/kjeb.2013.31.2.154>
37. Jo H, Baek K-W, Jeong K-S, Gim J-A, Kim JY, Joo G-J, et al. Length-weight relationships of three freshwater fishes from Hoiya River in South Korea. *J Appl Ichthyol.* 2018;34(4):1075–7. <https://doi.org/10.1111/jai.13717>
38. Yoon J-D, Kim J-H, Park S-H, Jang M-H. The distribution and diversity of freshwater fishes in Korean peninsula. *Korean J Ecol Environ.* 2018;51(1):71–85. <https://doi.org/10.11614/ksl.2018.51.1.071>
39. Kim J, Kwun H, Ji H-S, Park J, Myoung S, Song Y. A Guide Book to Marine Fishes in Korea. Ministry of Oceans and Fisheries; 2020.
40. Froese R, Pauly D. FISHBASE. fishbase.org. 2010.
41. Moen MO, Griffin KJ, Kalantar AH. Simple regression and outlier detection using the median method. *Anal Chim Acta.* 1993;277(2):477–87. [https://doi.org/10.1016/0003-2670\(93\)80459-x](https://doi.org/10.1016/0003-2670(93)80459-x)
42. Cook RD. Detection of Influential Observation in Linear Regression. *Technometrics.* 1977;19(1):15–8. <https://doi.org/10.1080/00401706.1977.10489493>
43. Nam KS, Sung-Hoi HUH. Changes in species composition of fishes in the nakdong river estuary. *Korean J Fish Aqua Sci.* 2003;36(2):129–35. <https://doi.org/10.5657/KFAS.2003.36.2.129>
44. Choi WS, Han JS, Choi JK, Lee HG. Fish fauna and characteristics of *Carassius auratus* population in the Boryeong Reservoir. *Korean J Environ Biol.* 2020;38(4):667–77. <https://doi.org/10.11626/kjeb.2020.38.4.667>
45. Kim J-S, Choi J-Y, Lee Y-J, Gwak W-S. Species Composition of Fishes in Eelgrass Bed of Minyang in Tongyeong, Korea. *The Sea:J Korean Soc Oceanograp.* 2013;18(4):227–33. <https://doi.org/10.7850/JKSO.2013.18.4.227>
46. Sun-Wan HW, Hak-Bin HW, Hyung-Soo N. Seasonal variation in species composition of fish collected by a bag net in the geum river estuary, Korea. *Korean J Fisheries Aquatic Sci.* 2005;38(1):39–54. <https://doi.org/10.5657/KFAS.2005.38.1.039>

47. Hong Y-K, Kim K-H, Song M-Y, Lee W-O. Ichthyofauna and Structure of the Fish Community in Lake Goesan, Korea. *Korean J Environ Ecol*. 2015;29(4):552–63. <https://doi.org/10.13047/kjee.2015.29.4.552>
48. Hong Y-K, Kim K-H, Kim K-M, Lim G-H, Song M-Y, Lee W-O. Characteristics of fish fauna and community structure in Wangpicheon. *Korean J Environ Ecol*. 2016;30(5):874–87. <https://doi.org/10.13047/kjee.2016.30.5.874>
49. Lee TW, Hwang HB, Hwang SW. Change in fish species composition in the Saemangeum reservoir after the construction of dike in 2006-2007. *J Korean Soc Oceanogr*. 2007;12(3):191–9. <https://doi.org/10.7850/jkso.2007.12.3.191>
50. Cousineau D, Chartier S. Outliers detection and treatment: a review. *Int j psychol res*. 2010;3(1):58–67. <https://doi.org/10.21500/20112084.844>
51. Xu L, Mazur M, Chen X, Chen Y. Improving the robustness of fisheries stock assessment models to outliers in input data. *Fisheries Res*. 2020;230:105641. <https://doi.org/10.1016/j.fishres.2020.105641>
52. Barletta M, Barletta-Bergan A, Saint-Paul U, Hubold G. Seasonal changes in density, biomass, and diversity of estuarine fishes in tidal mangrove creeks of the lower Caeté Estuary (northern Brazilian coast, east Amazon). *Mar Ecol Prog Ser*. 2003;256:217–28. <https://doi.org/10.3354/meps256217>
53. Rypel AL, Richter TJ. Empirical Percentile Standard Weight Equation for the Blacktail Redhorse. *North Am J Fish Manage*. 2008;28(6):1843–6. <https://doi.org/10.1577/m07-193.1>
54. Jana A, Sit G, Das P, Chanda A, Sahu S. Seasonal length-weight relationships and condition factors of mystus tengara (Hamilton, 1822) in Two Habitats. *ase*. 2022;0(0):0–0. <https://doi.org/10.26650/ase202221159748>
55. Faruque MH, Das R. Size-specific variations in the length-weight relationship and relative condition factor of Hilsa shad (*Tenualosa ilisha*) across its habitats in Bangladesh. *Heliyon*. 2024;10(13):e33586. <https://doi.org/10.1016/j.heliyon.2024.e33586> PMID: 39050429