

Evaluation of Smoothing Functions in Generalized Additive Models Applied to the Relative Abundance of Green Turtle (*Chelonia mydas*)

Letícia Verônica Santos

leticia.veronica@outlook.com

Federal University of Rio Grande (FURG)

Thayana Gião

Federal University of Santa Catarina (UFSC)

Raquel da Fontoura Nicolette

Federal University of Rio Grande (FURG)

Juliano César Marangoni

Federal University of Rio Grande (FURG)

Gustavo Martinez-Souza

Federal University of Rio Grande (FURG)

Research Article

Keywords: Distribution families, Model selection, Overdispersion, Sighting data

Posted Date: August 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10594651/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Evaluation of Smoothing Functions in Generalized Additive Models Applied to the Relative Abundance of Green Turtle (*Chelonia mydas*)

Letícia Verônica dos Santos ^{1,2,3*}, Thayana Gião ^{2,3,4†}, Raquel da Fontoura Nicolette ^{1,5†}, Juliano César Marangoni ^{1†} and Gustavo Martinez-Souza^{1,2,4†}

¹ Federal University of Rio Grande (FURG), Institute of Mathematics, Statistics and Physics (IMEF), Graduate Program in Ambientometrics (PPGAMB), Av. Itália, km 8, Carreiros, Rio Grande, 96203-900, RS, Brazil.

² Laboratório de Biometria & Conservação (BIOCO), Av. Itália, km 8, Carreiros, Rio Grande, 96203-900, RS, Brazil.

³ Instituto Caminho Marinho (ICMar), Rua A8, SN, Quadra012B Lote 004, Praia de Itapirubá, Laguna, 88790-000, SC, Brazil.

⁴ Federal University of Santa Catarina (UFSC), Rua Desembargador Vitor Lima, 222, Trindade, Florianópolis, 88040-400, SC, Brazil.

⁵ Environmental Statistics Laboratory (LEA), Av. Itália, km 8, Carreiros, Rio Grande, 96203-900, RS, Brazil.

*Corresponding author(s). E-mail(s):

leticia.veronica@outlook.com;

Contributing authors: thayana.giao@gmail.com

raquelnicolette@furg.br; julianomarangoni@furg.br;

gustavomarsouza@gmail.com;

†These authors contributed equally to this work.

Abstract

This study evaluates how response variable distribution families and smoothing techniques influence Generalized Additive Model (GAM) performance, using juvenile *Chelonia mydas* abundance data from southern Brazil. We systematically compared Normal, Log-normal, Poisson, and Negative Binomial distributions combined with Thin Plate, Cubic, and Penalized Regression Splines. The Negative Binomial distribution outperformed other frameworks across all information criteria and residual diagnostics, effectively mitigating overdispersion. Among smoothers, Thin Plate Regression Splines demonstrated superior efficiency in capturing complex spatial and seasonal patterns, yielding the lowest Generalized Cross Validation (GCV) and Root Mean Square Error (RMSE) scores. Ultimately, the correct specification of the response

distribution family proves just as critical as smoother selection for ensuring robust ecological inferences and accurately representing species abundance and site fidelity patterns.

Keywords: Distribution families, Model selection, Overdispersion, Sighting data.

1 Introduction

Generalized Additive Models (GAMs) represent a robust non-parametric framework for capturing nonlinear and heterogeneous relationships in complex environmental systems by replacing traditional linear predictors with a sum of smooth functions. A critical component of GAM specification is the selection of the error distribution family, which dictates the likelihood function and defines the fundamental mean-variance relationship. Because biological processes shape the underlying variance structure of ecological data, an inadequate distribution choice compromises the likelihood framework, resulting in biased inferences, underestimation of standard errors, and inflated Type I error rates under conditions such as overdispersion.

To circumvent these inferential errors, comparative model testing must evaluate both absolute goodness-of-fit and predictive generalization capacity. This is achieved through a multi-criteria approach incorporating Information Criteria (AIC, BIC), Generalized Cross-Validation (GCV), Root Mean Square Error (RMSE), explained deviance, and rigorous residual diagnostics.

When modeling ecological count data, such as monitoring records of juvenile *Chelonia mydas*, the classical assumptions of normality, independence, and homoscedasticity are routinely violated by zero-inflation, strong asymmetry, and overdispersion. Under these scenarios, the equidispersion constraint of the Poisson distribution is statistically untenable. The Negative Binomial distribution provides a flexible alternative by incorporating an explicit dispersion parameter that allows the variance to scale quadratically with the mean.

Within the penalized likelihood estimation framework of GAMs, the choice of the response distribution directly conditions the behavior, estimation, and optimization of the smoothing functions. The selection of the smoothing penalization degree—via automated methods like REML or GCV—occurs in direct interaction with the estimated dispersion parameter. Penalized smoothers demonstrate superior numerical stability under high residual noise, controlling artificial oscillations induced by aggregated data, while local smoothing methods excel in isolating unobserved variance stemming from complex spatiotemporal environmental gradients.

Consequently, evaluating smoother performance cannot be decoupled from error distribution analysis. Robust modeling requires joint verification using predictive cross-validation and information criteria to optimize mathematical fit while preserving the ecological interpretability of the parameters.

2 Background of Smoothers

In Generalized Additive Models (GAMs), the relationship between predictor variables and the response variable is modeled via non-parametric smoothing functions, structurally defined by the basis type argument (*bs*). The basis functions represent the mathematical foundations—typically regression splines—that construct the smoother, dictating both the flexibility and geometric configuration of the response curve. Selecting an appropriate basis function is critical to prevent overfitting and ensure the model captures genuine biological trends rather than stochastic noise, a process typically guided by information-theoretic criteria such as the Akaike Information Criterion (AIC).

From a structural perspective, smoothers provide a flexible, additive, and interpretable framework where the individual contribution of each covariate is summed (Hastie and Tibshirani, 1986). Curve fitting under classical non-parametric approaches initially seeks an unconstrained function $f(x_i)$, that minimizes the Residual Sum of Squares (RSS):

$$RSS = \sum (y_i - f(x_i))^2 \quad (1)$$

Where y_i represents the observed values of the dependent variable, $f(x_i)$ denotes the model predictions, $y_i - f(x_i)$ represents the residual, or error, for each data point; that is, the difference between the observed value and the value predicted by the model, $(y_i - f(x_i))^2$ squaring the error ensures that positive and negative deviations do not cancel each other out and imposes a greater penalty on larger errors and $\sum (y_i - f(x_i))^2$ it is the sum of all squared errors across the dataset, reflecting the model's total prediction error.

Squaring the error prevents directional cancellation and disproportionately penalizes larger discrepancies. However, minimizing RSS without structural constraints allows $RSS = 0$ through exact interpolation, leading to severe overfitting where the curve fits the residual noise rather than the underlying deterministic trend. Consequently, the objective is to estimate a function $f(x_i)$ that minimizes RSS while maintaining structural smoothness to ensure predictive generalization.

This approach prioritizes the balance between model flexibility and biological interpretability, using smoothing constraints to mitigate the risk of overfitting and ensure that nonparametric functions capture the real trend of ecological data rather than stochastic noise.

2.1 Smoothing Splines and Penalization Mechanics

Splines are piecewise polynomial functions joined at specific join-points termed knots, constructed to ensure mathematical continuity of both the function and its derivatives across adjacent segments. By dividing the predictor domain into discrete intervals and fitting low-degree polynomials within each segment, splines prevent abrupt discontinuities and provide a natural, stable fit to empirical observations. Within the GAM framework, smoothing splines balance data fidelity and structural complexity through a penalized regularization mechanism.

Mathematically, the smoothing spline function $f(x)$ is estimated by minimizing a penalized objective function that integrates the goodness-of-fit error with a roughness penalty:

$$\sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \int [f''(x)]^2 dx \quad (2)$$

Where y_i are the observed values, $f(x_i)$ is the value estimated by the curve, $f''(x)$ is the second derivative of the function (measuring its curvature) and λ is the smoothing parameter, governs the trade-off between baseline data fit and curve smoothness.

When $\lambda = 0$: The roughness penalty is nullified, resulting in an unconstrained exact interpolation that passes through all observed data points, maximizing variance and inducing severe overfitting.

When $\lambda \rightarrow \infty$: The penalty on the second derivative becomes infinitely restrictive, suppressing all curvature and forcing $f''(x)$, which reduces the smoother to a standard linear regression framework.

Within the scope of this study, emphasis is given to penalized splines, whose fitting process is governed by the minimization of an objective function balancing fidelity to the data and smoothness of the curve. The control of this balance is exerted by the penalization parameter (λ): while null values of λ lead to overfitting through exact interpolation, values tending toward infinity result in model linearization. The optimized selection of this parameter, essential for the robustness of the inferences presented here, was conducted using automated methods such as Generalized Cross-Validation (GCV) or Restricted Maximum Likelihood (REML), ensuring that model complexity is consistent with the actual biological trend observed in *Chelonia mydas* abundance data.

2.1.1 Thin Plate Regression Splines (TPRS)

Thin Plate Splines (TPS) constitute a robust non-parametric framework for multidimensional smoothing and multidimensional surface interpolation . Within the R statistical environment, TPS represents the default smoother implementation in the *mgcv* package via the `bs = "tp"` argument . The conceptual foundation of the thin plate spline derives from a physical analogy: a thin, elastomeric metal plate forced to pass through a set of fixed data points will naturally deflect to minimize its total bending energy. Mathematically, the estimated smoothing function balances data fidelity with this physical constraint, utilizing radial basis functions to ensure smooth geometric transitions across multivariate and geospatial coordinate spaces.

In a conventional generalized additive framework, the thin plate smoother resolves a penalized optimization problem by minimizing an energy functional associated with the total curvature of the multidimensional surface. In a two-dimensional spatial context, where the objective is to estimate a smooth surface $f(x_i, z_i)$, this optimization problem is formulated by minimizing the following objective function:

$$\min_f \sum_{i=1}^n (y_i - f(x_i, z_i))^2 + \lambda \int \int [(\frac{\partial^2 f}{\partial x^2})^2 + 2(\frac{\partial^2 f}{\partial x \partial z})^2 + (\frac{\partial^2 f}{\partial z^2})^2] dx dz + \varepsilon_i \quad (3)$$

Where the first term quantifies the data fitting error, the second term measures the total curvature of the surface, λ governs the trade-off between model fit and smoothness, $f(x_i, z_i)^2$ is the smoothing function to be estimated, y_i represents the observed values of the response variable and $\int \int \square$ is the penalization term that integrates the variation of the second derivative of f , representing the bending energy of the surface.

In the context of this article, it is noteworthy that TPS solves an optimization problem minimizing the energy functional associated with surface curvature, ensuring a fit that is simultaneously faithful to the data and geometrically smooth. This methodological choice is justified by its isotropic and robust nature, particularly effective in capturing complex patterns and attenuating noise in multivariate data, such as records of *Chelonia mydas* sightings as a function of environmental and temporal variables.

2.1.2 Cubic Regression Splines (CRS)

Another relevant smoothing basis for ecological time-series modeling, which uses third-degree piecewise polynomials. Each distinct segment of the predictor domain is fitted with a separate cubic polynomial, where structural continuity across adjacent intervals is strictly enforced not only for the function values themselves, but also for their first and second

derivatives at the knots. This specific differentiability characteristic minimizes abrupt geometric transitions, making cubic splines exceptionally effective at capturing underlying smooth deterministic patterns in ecological data streams without inducing spurious oscillations or artificial boundary artifacts.

Mathematically, the piecewise cubic spline function $S(x)$ is constructed as:

$$S(x) = \begin{cases} P_1(x) & \text{sex} \in [t_1, t_2] \\ P_2(x) & \text{sex} \in [t_2, t_3] \\ \dots & \dots \\ P_k(x) & \text{sex} \in [t_k, t_{k+1}] \end{cases} \quad (4)$$

Where each $P_i(x)$ is a cubic polynomial, ensuring mathematical continuity up to the second derivative at the knots $t_2, t_3, \dots, t_k$.

The formulation of cubic splines is uniquely advantageous due to its concurrent optimization of structural interpretability, computational efficiency, and precise smoothness regulation, particularly when coupled with penalized regularization techniques. Because the resulting function and its first- and second-order derivatives maintain strict mathematical continuity across all defined intervals, the estimated curve exhibits stable gradients and local curvature. This differentiability constraint effectively mitigates the severe, spurious oscillations commonly induced by unconstrained high-degree global polynomials, thereby ensuring geometrically stable transitions across adjacent data segments.

Within the scope of this study, emphasis is given to the distinction between the conventional smoother (“cr”) and its shrinkage variant (“cs”). While the former focuses strictly on smooth transitions between segments, the shrinkage version incorporates an additional penalization that allows the smooth term to be reduced to zero if its contribution is statistically irrelevant. This characteristic acts as an automatic variable selection strategy, promoting parsimony and ensuring that only effectively informative predictors compose the modeling of *Chelonia mydas* abundance, resulting in a more interpretable and robust final model.

2.1.3 Penalized Splines

P-splines, proposed by Eilers and Marx, represent a smoothing technique integrating the flexibility of B-spline basis functions with regularity control through finite-difference penalization.

The primary divergence of P-splines from traditional B-spline frameworks lies in the explicit integration of a regularization penalty applied directly to the finite differences of adjacent basis coefficients. This difference penalty acts as a numerical constraint against excessive model variability, effectively restricting the curvature complexity of the fitted function and adaptively enforcing structural smoothness. Consequently, the resulting smoothing estimator achieves a concurrent optimization of empirical data fidelity and regular geometric curvature.

Mathematically, the general formulation of a penalized B-spline can be representing as:

$$f(x) = \sum_{j=1}^K \beta_j B_j(x) \quad (5)$$

Where $B_j(x)$ represents the B-spline basis function predetermined by the knot distribution, β_j denotes the regression coefficients estimated from the empirical data and K is the total number of B-spline basis functions, which scales as a function of both the knot density and the spline degree.

Within the scope of this article, the main advantage of P-splines lies in the application of penalization over differences of order d among coefficients, restricting excessive variability and adaptively imposing smoothness. This formulation offers superior numerical stability and computational efficiency, being particularly advantageous for modeling nonlinear trends in large ecological datasets. By balancing residual error and curvature penalization, P-splines ensure that *Chelonia mydas* abundance estimates preserve biological interpretability while avoiding artifacts generated by inadequate knot distributions.

2.1.4 Adaptive Splines

Unlike global penalization approaches, adaptive smoothing allows model flexibility to vary dynamically across the predictor domain. This technique is particularly effective in scenarios of high structural heterogeneity, where data exhibit abrupt trend changes or variable density.

This localized adaptability is functionally conditioned by both data density and the spatial or temporal heterogeneity of the underlying ecological process. In coordinate regions characterized by sparse data or abrupt threshold transitions, the model adaptively relaxes the penalty constraint, thereby increasing localized flexibility to resolve rapid structural changes. Conversely, in regions featuring high data density and uniform variance, the framework enforces a strict penalization, generating a highly regularized and smooth trajectory. Consequently, adaptive smoothing mechanisms dynamically conform to localized data topologies, mitigating the dual risks of localized overfitting and the truncation of geographically or temporally restricted ecological signals.

Mathematically, this non-parametric framework is operationalized by replacing the invariant global smoothing scalar (λ) with a spatially or temporally varying penalization structure applied to the basis function coefficients. By incorporating a localized penalty matrix that modulates across the predictor domain, the estimator achieves precise locational control over curve flexibility. The objective function for an adaptive penalized spline modeling the smooth function $f(x)$ is formulated through the minimization of the following expression:

$$\min \sum_{i=1}^n (y_i - f(x_i))^2 + \sum_{j=d+1}^K \lambda_j (\Delta^d \beta_j)^2 \quad (6)$$

Where $\Delta^d \beta_j$ is the finite difference operator of order d (typically $d = 2$) and λ_j represents the adaptive smoothing penalty coefficient specific to each respective spline segment, which is automatically estimated from the empirical data.

This adaptive framework is exceptionally effective for modeling structural data heterogeneity, including irregular seasonal time series, non-uniform functional responses, and high-variance spatial data. By modulating the localized smoothing parameter vector λ_j via smooth functions over the covariate space (e.g., a regression on log-penalties) or through Restricted Maximum Likelihood (REML), Maximum Likelihood (ML), and Empirical Bayes estimation, the model dynamically balances regularization. Consequently, it decreases

penalty constraints in high-variability regions to allow greater flexibility while increasing penalization across homogeneous segments to enforce structural smoothness.

Within the scope of this study, the application of adaptive smoothers is based on replacing the single penalization parameter λ with a variable penalization λ_i , specific to each spline segment. This approach allows reduced penalization in regions of high variability and increased penalization in sections exhibiting uniform behavior. This mathematical plasticity, estimated through Restricted Maximum Likelihood (REML) or Empirical Bayes approaches, ensures that the modeling of *Chelonia mydas* abundance captures local nuances and irregular seasonal patterns, while simultaneously mitigating the risk of overfitting in sparse-data regions and underrepresentation of ecologically relevant patterns.

To effectively harness this mathematical plasticity and address the systematic challenges inherent in overdispersed ecological data, the general objective of this work is to investigate the structural interaction between smoother types and error distribution families in Generalized Additive Models (GAMs), thereby establishing a robust inferential framework for modeling the abundance of juvenile *Chelonia mydas*. This overarching investigation aims to determine how the coupling of specific response distributions and advanced spline configurations can stabilize parameter estimations under high residual noise and variance inflation.

To operationalize this comprehensive evaluation, the subsequent Materials and Methods section outlines a systematic, three-tiered analytical workflow. First, it describes the parametric benchmarking designed to identify the optimal probability distribution family that accommodates the intrinsic variance structure of the empirical data. Second, it establishes the competitive framework used to evaluate the robustness and numerical stability of alternative smoothing classes against localized overfitting. Finally, this structured methodology details how non-parametric estimators are deployed to resolve complex, nonlinear relationships between environmental predictors and count responses, providing a transparent, reproducible roadmap that bridges theoretical statistical optimization and robust ecological interpretation.

3 Material and Methods

3.1 Study Area

Fieldwork was executed along the rocky shore of Itapirubá Norte Beach 28° 20' 14.431" S, 48° 42' 7.031" W; Imbituba, Santa Catarina, Brazil), a region integrated into the Federal Environmental Protection Area of Baleia Franca (*APA Baleia Franca*). Composed of an igneous and metamorphic matrix, this high-productivity rocky shore provides critical structural complexity, serving as a key foraging ground and shelter for juvenile green turtles (*Chelonia mydas*). The local oceanographic regime—modulated by water mass convergence and prevailing winds—enhances nutrient deposition and sustains dense macroalgal communities essential for these chelonians during their ontogenetic development.

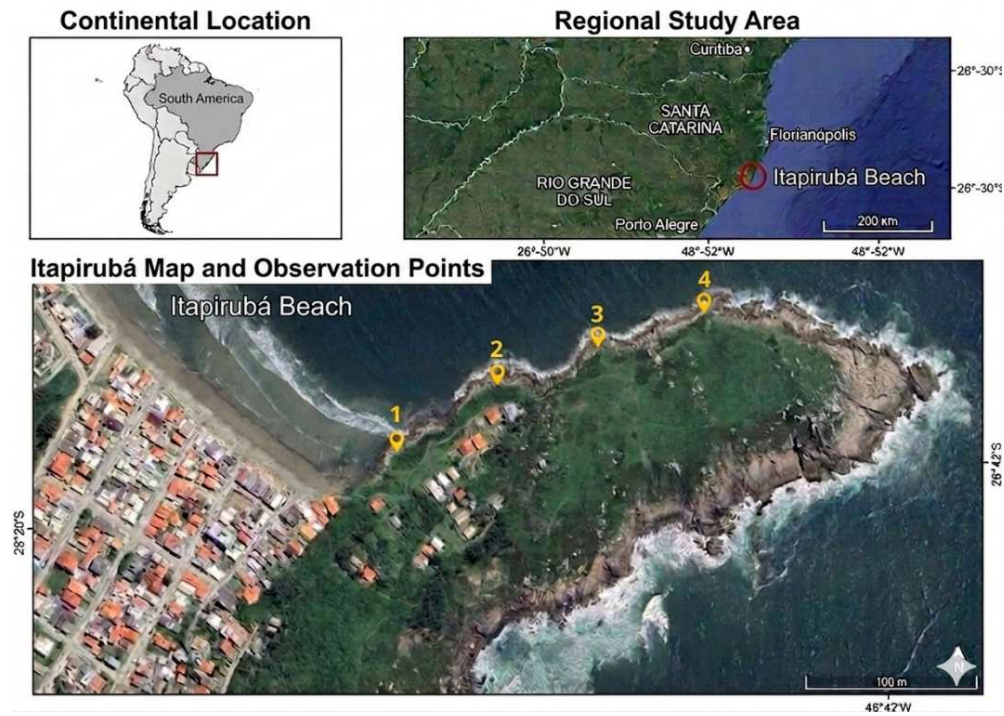


Figure 1. Geographic location of the study area. **(A)** Continental location within South America; **(B)** Regional coastal context of Santa Catarina State; **(C)** High-resolution map of Itapirubá Norte rocky shore indicating the four systematic observation stations (1-4). **Source:** Prepared by the author (2026), based on Google Earth imagery.

3.2 Free-Ranging Population Sightings

The discrete response variable was derived from visual counts of sea turtles monitored continuously since 2021 along the rocky shore of Itapirubá Beach by the *Instituto Caminho Marinho (ICMar)*. Sampling was conducted from four fixed, categorical observation stations distributed along an onshore-offshore gradient from the beach entry to the outer point facing the Atlantic Ocean: Station 1 (28° 20' 16.8" S, 48° 42' 11.5" W), Station 2 (28° 20' 14.5" S, 48° 42' 07.7" W), Station 3 (28° 20' 13.5" S, 48° 42' 03.7" W) and Station 4 (28° 20' 10.3" S, 48° 41' 57.4" W). At each station, a pair of experienced observers performed systematic naked-eye visual scans divided into four consecutive 10-minute intervals (totaling 40 minutes per sampling session), recording individuals surfacing during foraging or resting periods. Surveys spanned the entire diurnal cycle year-round, with sampling intensity modulated by team availability. This collaborative dataset was provided by the "*No Caminho das Tartarugas*" project, a joint conservation network comprising ICMar, BIOCO, IMEF, and FURG.

3.3 Meteoceanographic Data Acquisition

To model the environmental drivers of sea turtle abundance, ten candidate predictor variables were compiled from open-access, replicable environmental monitoring platforms: temporal tracking (1) Days, (2) Temperature, (3) Tidal Height, (4) Wave Height, (5) Wave

Period, (6) Pressure), hydrodynamics (3) Tidal Height, (4) Wave Height, (5) Wave Period, (7) Time of Day, (8) Wind Speed, along with the calculated (9) North-South and (10) East-West wind vectors. Hourly atmospheric temperature, surface pressure, and meteorological tide data were extracted from the *Centro de Informações de Recursos Ambientais e de Hidrometeorologia de Santa Catarina* .

Coastal hydrodynamic parameters (wave height, period, and direction) and wind field vectors (speed in knots and geographic direction) were acquired from hindcast models hosted by the Windguru platform . Complementary meteorological and climatological validation records were retrieved from the *Instituto Nacional de Meteorologia* and the *Instituto Nacional de Pesquisas Espaciais* , ensuring robust geospatial alignment for the targeted coastal zone.

3.4 Use of Artificial Intelligence Tools

During the development of this study, Generative Artificial Intelligence (AI) frameworks—specifically the ChatGPT and Gemini platforms—were utilized exclusively as auxiliary instruments for textual organization, grammatical refinement, and the synthesis of methodological graphical layouts. In strict accordance with the core principles of scientific integrity and academic ethics, the intellectual execution of the mathematical architectures, the biological interpretation of the ecological results, and the technical validation of all written content remain entirely under the responsibility of the authors, thereby guaranteeing the originality, authenticity, and reliability of the evidence presented herein.

3.5 Exploratory Analysis

Ecological and biological processes—such as birth, growth, decay, mortality, abundance, production, and competition—are increasingly recognized as inherently non-linear functions of environmental predictors . Recently, the rapid evolution of open-source statistical software and integrated development environments (IDEs), particularly the R computing environment deployed via RStudio has expanded accessibility to complex, non-parametric analytical architectures. In this study, sea turtle abundance data derived from visual sightings were structurally integrated into data frames, undergoing rigorous quality control, missing-data filtering, and variable transformations. The final high-resolution dataset comprised $n = 4042$ observations spanning from June 9, 2022, to July 22, 2024. Preliminary exploratory data analysis quantified the degree of zero-inflation and cumulative response metrics, alongside spatial-temporal clustering evaluations stratified across months, years, and discrete sampling stations, both independently and aggregated over the entire multi-year monitoring period.

3.6 Statistical Modeling

This study adopted a full modeling approach wherein all predictor variables were simultaneously retained within the global model structure. This strategy was intentionally designed to isolate and systematically evaluate two fundamental dimensions of the optimization process: (i) the structural adequacy of the response variable's probability distribution family and (ii) the relative performance of alternative smoothing mechanics. Unlike conventional variable selection protocols that rely on sequential parameter pruning, no traditional stepwise elimination was conducted. Instead, shrinkage and regularization

approach inherent to the smoothers were utilized to handle model complexity automatically, as the arbitrary exclusion of covariates would alter the functional geometry of the smooths and compromise cross-model comparability. Consequently, the adoption of a fixed global model architecture guaranteed rigorous consistency throughout the comparative benchmarking.

All statistical computations were executed within the R environment utilizing the `mgcv` library, the standard framework for fitting additive architectures via automated roughness penalization . Model specifications explicitly accommodated the discrete count properties of the response vector and the presence of overdispersion, both typical features of empirical ecological data (O'Hara & Kotze, 2010). The selection of the smoothing parameter (λ) was operationalized via Restricted Maximum Likelihood (REML) estimation, a framework that provides superior numerical stability and minimized bias relative to alternative objective functions such as Generalized Cross-Validation (GCV) .

3.7 Determination of the Response Variable Probability Distribution

Given that the response variable comprises discrete sea turtle sighting counts, specifying the probability distribution *a priori* is a fundamental requirement for the structural integrity of Generalized Additive Models (GAMs) . Conducting this selection prior to model fitting ensures that the chosen statistical family accurately reflects the data's error structure, thereby providing a consistent baseline to evaluate smoother performance . Candidate distributions (Normal, Log-normal, Poisson, and Negative Binomial) were evaluated based on:

- Mean–variance relationship;
- Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) ;
- Graphical diagnostics, including Q-Q plots and empirical density distributions.

To prevent analytical distortions stemming from spatial heterogeneity, descriptive statistics—focusing on measures of central tendency and dispersion—were initially calculated across each discrete "sampling site." Subsequently, candidate probability distributions were systematically fitted and evaluated within the generalized additive modeling framework using the `mgcv` package . The following distributions were tested:

- Normal;
- Log-normal;
- Poisson;
- Negative Binomial.

3.8 Smoothing Basis Selection and Regularization

To evaluate how basis function mechanics influence model architecture, distinct GAM structures were fitted by systematically varying the smoothing basis classes while maintaining a fixed predictor matrix. This controlled benchmarking isolated the trade-offs

between localized flexibility and roughness penalization across different basis specifications. To restrict model complexity without explicit variable elimination, double-penalty regularization was operationalized via the `select = TRUE` argument in `mgcv`. This integrated approach forces the singular values of non-informative smooth terms toward zero, enabling automated parameter shrinkage and yielding a more parsimonious multivariate topology .

Complementarily, a preliminary exploratory phase assessed the individual performance of alternative smoothing bases for each environmental predictor. Univariate GAMs were systematically constructed and ranked using the corrected Akaike Information Criterion (AICc). While this univariate screening optimizes localized basis fit, it serves an exclusively exploratory purpose as it ignores higher-order interactions and multi-collinear dependencies within the complete multivariate framework .

In contrast to localized smoothers (e.g., LOESS/LOWESS span parameters), kernel density estimators (e.g., bandwidth boundary constraints), or random-effect variances in hierarchical mixed models, splines within the Penalized Quasi-Likelihood/REML framework optimize the global-local smoothness balance directly through localized penalty matrices. Based on these mathematical properties, four distinct smoothing approaches—previously detailed in the theoretical framework—were comprehensively evaluated:

- **Thin Plate Regression Splines (tp):** Operationalized as the baseline isotropic smoother due to its numerical robustness in resolving multi-dimensional patterns without the subjective bias of manual knot placement;
- **Cubic Regression Splines (cr):** Constrained via piecewise third-degree polynomials to guarantee continuous, smooth transitions across discrete data segments;
- **Penalized Splines (ps):** Implemented to optimize the explicit trade-off between structural flexibility and B-spline roughness penalization, thereby suppressing spurious oscillations and local overfitting;
- **Adaptive Smoothing Splines (ad):** Deployed to allow the smoothing parameter to vary functionally across the predictor domain. This adaptive allocation yields high localized flexibility across non-linear ecological gradients while maintaining structural rigidity across stable intervals—a critical property for accommodating meteoceanographic variables prone to abrupt physical regime shifts (e.g., sharp sea surface temperature anomalies or localized wind vector fluctuations).

To guarantee that non-linear relationships between environmental predictors and juvenile sea turtle abundance are resolved with optimal precision, a systematic comparative evaluation among alternative smoothing bases was conducted. For each continuous covariate, candidate architectures were benchmarked by shifting the basis function specifications within a controlled framework, keeping the baseline parametric and random-effect structures invariant.

Model parsimony, goodness-of-fit, and structural complexity were simultaneously assessed using a multi-metric criteria matrix, wherein model selection relied on the Akaike Information Criterion (AIC), the corrected Akaike Information Criterion (AICc) to penalize potential overparameterization, and Restricted Maximum Likelihood (REML) scores. To safeguard against localized overfitting, cross-validation procedures were implemented to

explicitly evaluate out-of-sample predictive capacity, while the estimated Effective Degrees of Freedom (EDF) for each smooth term were rigorously monitored as a proxy for functional complexity, where $\text{EDF} > 1$ identifies non-parametric non-linearities and $\text{EDF} \sim 1$ indicates convergence toward a standard linear relationship.

4 Results

4.1 Descriptive Baseline and Convergence Dynamics

Initial diagnostic screening of the sea turtle sighting records revealed an empirical structure typical of opportunistic ecological field surveys, characterized by an overall zero-inflation rate of 28.65% and severe overdispersion, with a global variance-to-mean ratio exceeding unity ($\sigma^2 = 13.65$ vs. $\mu = 2.93$). This overdispersion pattern was consistently maintained across all spatial strata, though the localized frequency of structural zeros varied substantially among sectors: **Station 1** exhibited the lowest absence rate (20.7%), followed by **Station 2** (31.3%) and **Station 3** (32.4%), whereas **Station 4** displayed the highest zero-inflation, with absences spanning 43.6% of its operational time series. This fine-scale spatial heterogeneity in zero distribution underscores the necessity of implementing robust mixture or count-family GAM formulations capable of explicitly accommodating overdispersion and the semicontinuous properties of localized rocky shore abundance matrices.

Under the fixed global architecture—wherein all environmental predictors were evaluated simultaneously—the generalized modeling framework demonstrated high numerical stability under alternative smoothing configurations and error distributions. Parameter optimization via Restricted Maximum Likelihood (REML) achieved stable convergence across all candidate specifications, yielding consistent information criteria scores and orthogonal, homoscedastic residual topologies. Diagnostic profiles confirmed that the REML objective function effectively mitigated localized overfitting risks, successfully optimizing the mathematical equilibrium between multi-dimensional spline flexibility and roughness penalization.

The general model using the default smoothing function (s), which adopts the thin plate basis, is described as follows:

```
general_model <- gam( Sightings ~ Station + s(Days) + s(Temperature) + s(Tide) + s(Wave_Height) + s(Wave_Period)
+ s(Pressure) + s(Time_of_the_day) + s(North-South_Component) + s(East-West_Component), data = dad, family
= nb(link = "log"))
```

Preserving the global predictor matrix invariant across all candidate models facilitated a direct, controlled benchmarking of alternative architectures, demonstrating that variations in goodness-of-fit were predominantly driven by smoothing basis mechanics rather than variable omission or inclusion. This outcome validates the structural consistency of the modeling protocol, successfully isolating the mathematical effects of different spline bases on ecological model calibration.

4.2 Spatial Distribution and Dispersion Properties

Spatial stratification of the sighting data revealed strong positive skewness and distinct leptokurtic profiles across all monitoring sectors, reflecting the sporadic nature of high-abundance aggregation events typical of heterogeneous marine habitats. **Station 1** exhibited the highest demographic density, with a mean abundance of 4.36 (± 4.85) individuals and a

median of 3. This central tendency discrepancy underscores a highly right-skewed allocation (skewness = 2.02) and a prominent heavy-tailed behavior, characterized by an expansive operational range extending up to 12 individuals and sharp tail inflation (kurtosis = 6.79). This pattern of localized concentration at low values punctuated by rare, high-density pulses was closely mirrored at **Station 2** (mean = 2.36 ± 2.75 ; median = 1; skewness = 1.90; kurtosis = 6.03) and **Station 3** (mean = 2.36 ± 2.75 ; median = 1; skewness = 2.08; kurtosis = 6.13), both capturing restricted baseline presence coupled with occasional foraging influxes.

Conversely, **Station 4** registered the lowest baseline density among the analyzed sectors (mean = 1.78 ± 2.63 ; median = 1), bounded by a narrower maximum range of 6 individuals. Although Station 4 retained a right-skewed geometry (skewness = 1.93), its substantially lower kurtosis (3.20) indicates a muted influence of extreme aggregation anomalies relative to the inner rocky shore sectors. Taken together, these fine-scale asymmetric distributions and localized dispersion variances strongly suggest that the targeted rocky shore functions as a heterogeneous-use habitat, where juvenile green turtle patchiness is modulated by underlying spatial gradients in temperature and primary productivity.

4.3 Diagnostic Selection of Probability Distribution Families

To identify the underlying stochastic process governing the sighting frequency data, a formal distributional benchmarking pipeline was executed to directly evaluate four candidate probability density functions: Normal, Lognormal, Poisson, and Negative Binomial. These families were selected as they constitute the standard parametric frameworks for resolving the inherent complexities of ecological count structures, ranging from continuous baselines to highly overdispersed and zero-inflated abundance matrices. By fit-testing these distinct statistical paradigms against the empirical juvenile green turtle (*Chelonia mydas*) datasets, this framework provides a rigorous, likelihood-based approach to isolate the optimal error distribution without relying on preliminary non-parametric skewness-kurtosis heuristics.

The comparative evaluation of the candidate distributions was operationalized using Maximum Likelihood Estimation (MLE) to compute stable parametric vectors for each distribution family. To ensure structural adequacy and prevent inferential bias caused by variance misspecification—such as treating highly asymmetric, heavy-tailed count structures under symmetric Gaussian assumptions—the models were subjected to a multi-faceted validation protocol. Model optimization and goodness-of-fit benchmarking were systematically executed through a comprehensive framework incorporating the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and log-likelihood scores.

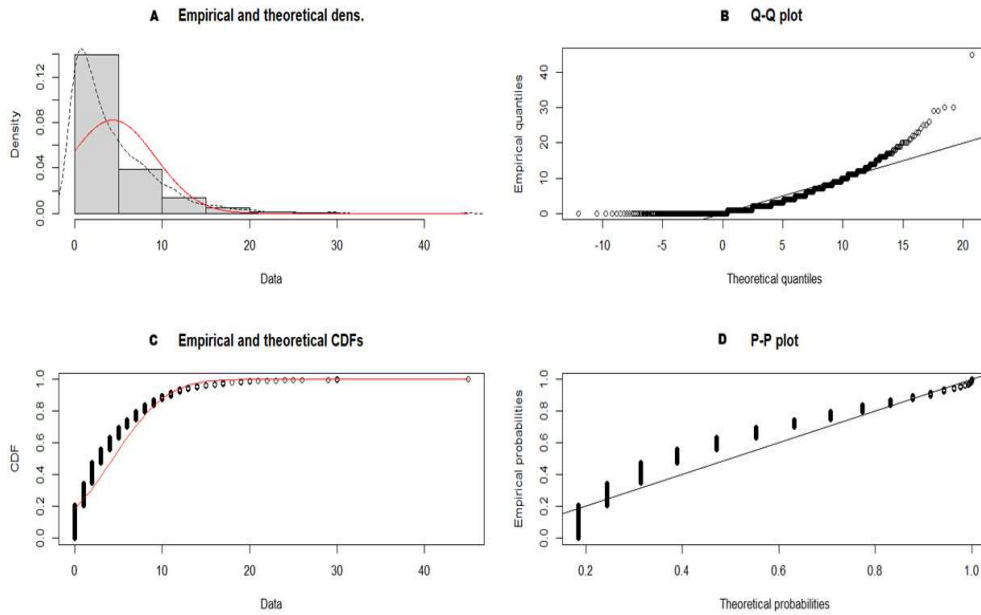
Following these information-theoretic criteria, a rigorous residual diagnostic workflow was implemented to verify pressupositional compliance. This phase utilized empirical Quantile-Quantile (Q-Q) plots, residual-versus-fitted scatterplots, and specific dispersion parameter (ϕ) assessments to identify potential overdispersion or localized autocorrelation. The results of this competitive parametric benchmarking, along with the formal selection metrics for the Normal, Lognormal, Poisson, and Negative Binomial configurations, are presented and analyzed in the following subsection.

Analysis of Sighting Station 1

Normal Distribution

The structural diagnostic profiles for the Gaussian (Normal) distribution family fitted to the juvenile green turtle sighting data at Station 1 are presented in Figure 2. Empirical versus theoretical density comparisons, alongside corresponding quantile-quantile (Q-Q) and probability-probability (P-P) spaces, revealed severe structural mismatches, demonstrating the mathematical inability of the classical Gaussian architecture to accommodate the zero-inflated, highly asymmetric properties of the observed abundance matrix.

Figure 2. Diagnostic plots of the Normal distribution family fitted to juvenile *Chelonia mydas* sighting data at Station 1. **(A)** Empirical versus theoretical density histogram; **(B)** Quantile–Quantile (Q-Q) plot; **(C)** Empirical versus theoretical Cumulative Distribution Function (CDF); **(D)** Boxplot. **Source:** Prepared by the author (2026).



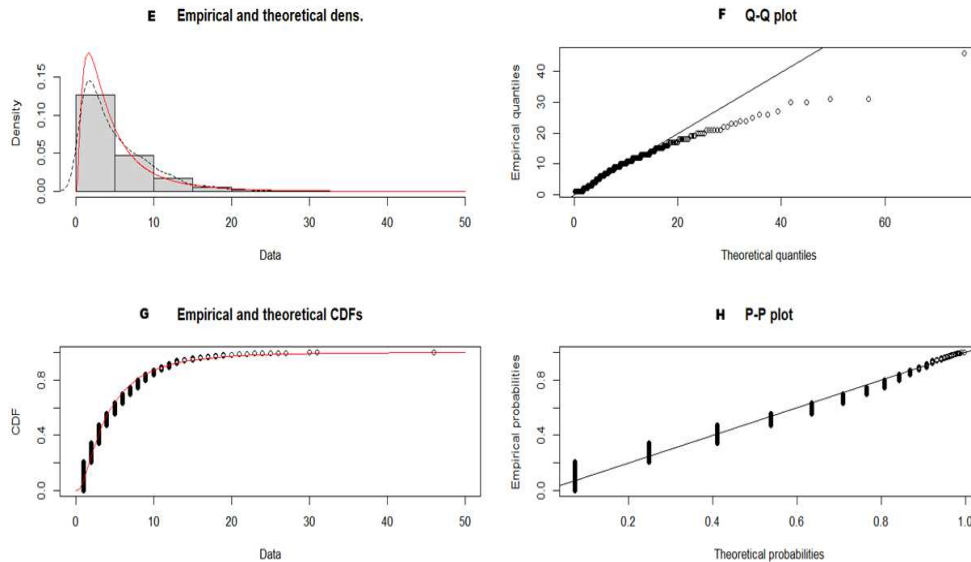
The Gaussian distribution fit exhibited poor performance, as evidenced by high AIC (8231.516) and BIC (8241.964) values, alongside pronounced visual discrepancies in the diagnostic plots.

Log-normal Distribution

The structural diagnostic profiles for the Lognormal distribution family fitted to the juvenile green turtle sighting data at Station 1 are detailed in Figure 3. Systematic comparisons between empirical and theoretical density spaces, alongside corresponding quantile-quantile (Q-Q) and probability-probability (P-P) plots, demonstrated a marked improvement in capturing the severe positive skewness of the dataset compared to previous frameworks. Nevertheless, localized structural limitations persist, indicating that while the Lognormal architecture partially accommodates the long-tailed profile, it fails to achieve complete statistical adequacy across the full gradient of the observed abundance matrix.

Figure 3. Diagnostic plots of the Lognormal distribution family fitted to juvenile *Chelonia mydas* sighting data at Station 1. **(E)** Empirical versus theoretical density histogram; **(F)** Quantile–Quantile (Q–Q) plot; **(G)** Empirical versus theoretical Cumulative Distribution Function (CDF); **(H)** Boxplot. **Source:** Prepared by the author (2026).

The Log-normal distribution provided a substantial improvement in model fit, yields a

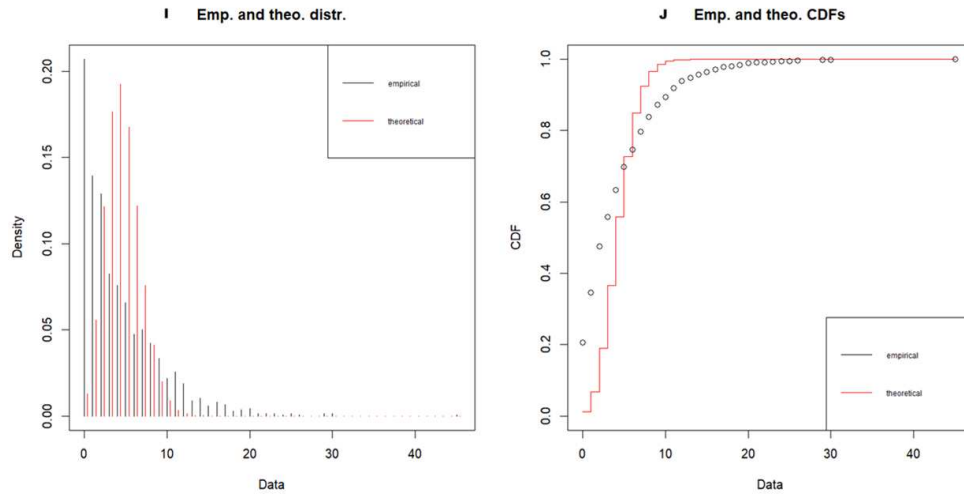


significant reduction in AIC (7159.99) and BIC (7170.438) values, which reflects its superior capacity to capture data asymmetry.

Poisson Distribution

The diagnostic profiles evaluating the performance of the Poisson distribution family fitted to the juvenile green turtle sighting data at Station 1 are illustrated in Figure 4. Systematic comparisons between empirical and theoretical probability densities, alongside their respective cumulative distribution functions (CDFs), highlighted severe structural discrepancies. The classical Poisson architecture failed to adequately replicate the observed ecological variability, primarily due to its restrictive equidispersion assumption, which proved mathematically incompatible with the highly overdispersed nature of the abundance matrix.

Figure 4. Diagnostic plots of the Poisson distribution family fitted to juvenile *Chelonia mydas* sighting data at Station 1. **(I)** Empirical versus theoretical density comparison; **(J)** Empirical versus theoretical Cumulative Distribution Function (CDF). **Source:** Prepared by the author (2026).

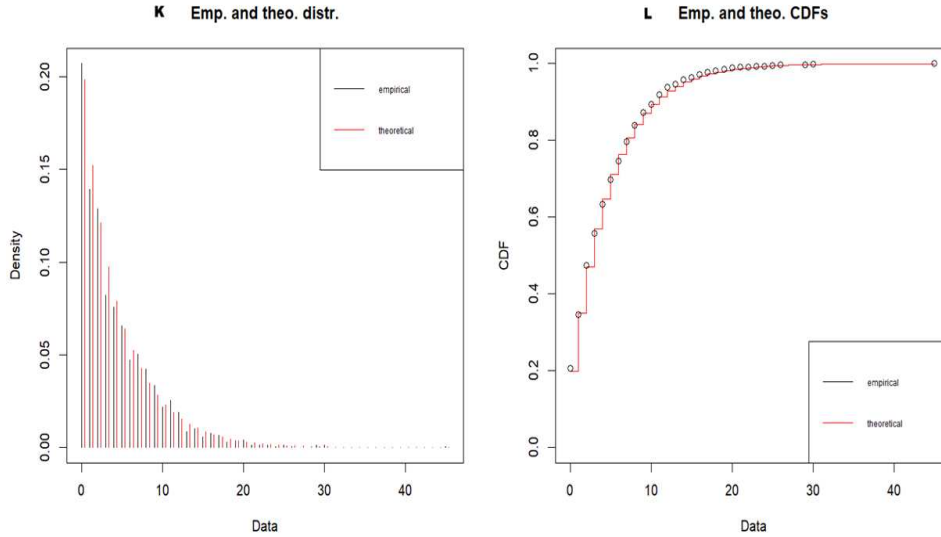


The Poisson distribution exhibited inferior performance, as evidenced by elevated AIC (10186.97) and BIC (10192.19) values, alongside strong discrepancies in the diagnostic plots

Negative Binomial Distribution

The diagnostic profiles evaluating the Negative Binomial distribution family fitted to the juvenile green turtle sighting data at Station 1 are illustrated in Figure 5. Systematic comparisons between empirical and theoretical probability densities, alongside their respective cumulative distribution functions (CDFs), demonstrated an exceptionally high degree of alignment. The Negative Binomial architecture exhibited superior performance in replicating both the severe positive skewness and the overdispersion structure of the dataset, successfully accommodating the ecological variability and tail inflation that constrained the previous candidate frameworks.

Figure 5. Diagnostic plots of the Negative Binomial distribution family fitted to juvenile *Chelonia mydas* sighting data at Station 1. **(K)** Empirical versus theoretical density comparison; **(L)** Empirical versus theoretical Cumulative Distribution Function (CDF). **Source:** Prepared by the author (2026).



In contrast, the Negative Binomial distribution showed the best performance among the evaluated models, presenting the lowest AIC (7079.15) and BIC (7089.598) values, indicating superior fit to the data.

Remaining fixed sighting stations: Stations 2, 3, and 4

Consistently, the remaining fixed monitoring stations exhibited distributional topologies highly congruent with the patterns observed at Station 1 across all evaluated probability frameworks. Consequently, the rejection of the classical Poisson architecture—owing to its restrictive equidispersion constraint in favor of the Negative Binomial distribution represents a critical methodological pivot. By integrating an independent dispersion parameter, the Negative Binomial framework successfully accommodates the severe overdispersion and tail inflation inherent to the empirical sighting matrix, thereby ensuring substantially more reliable parameter estimations, stabilized standard errors, and statistically robust model diagnostics across the entire spatial domain.

4.4 Smoother Performance and Model Selection Matrix

Count data characterized by structural overdispersion and a substantial zero-inflation rate (~28.65% absences), as documented in this empirical baseline, strongly influence the numerical stability of localized smoothers, potentially amplifying overfitting artifacts when coupled with highly flexible basis expansions. In this framework, isolating the mathematically appropriate probability distribution family constituted a fundamental prerequisite for subsequent generalized additive architecture. Transitioning to a Negative Binomial error structure effectively stabilized the smoothing parameter estimation vector, successfully optimizing the bias-variance trade-off by decoupling the mean-variance relationship. Consequently, the systematic benchmarking of smoother performance executed throughout this section was parameterized exclusively under the Negative Binomial distribution framework, ensuring strict mathematical coherence between the empirical data topology and the operationalized statistical modeling suite.

The candidate GAM structures fitted with alternative smoothing bases exhibited distinct performance variations, highlighting the pivotal role of basis function selection on global model adequacy, alongside nuanced differences in localized flexibility across specific environmental covariates. Operationalizing double-penalty regularization via the `select = TRUE` constraint yielded highly parsimonious multivariate topologies. This shrinkage mechanism effectively compressed the mathematical complexity of weakly informative or non-significant smooth terms toward zero, driving structural parsimony without inducing adverse adjustments in goodness-of-fit, as validated by the optimization scores.

Accordingly, final model selection was adjudicated via the corrected Akaike Information Criterion (AICc). For each environmental predictor variable, candidate smoothers were systematically ranked, isolating the baseline configuration exhibiting the lowest objective score as the optimal fit ($\Delta\text{AICc} = 0$). Table 1 presents a summarized matrix containing the top two best-performing smoothing bases for each continuous and categorical covariate, facilitating a direct visualization of parametric consistency and structural selection dynamics prior to final model integration.

Table 1: Model selection matrix showcasing the two best-performing smoothing bases tested across the selected environmental predictor variables.

bs	AICc	delta_AICc	rank	Predictor Variable
ad	22407.70	0	1	Days
tp	22458.71	51.00508	2	Days
ad	23199.48	0	1	Temperature
cr	23208.04	8.55937	2	Temperature
ad	23830.35	0	1	Tide
cr	23835.23	4.87276	2	Tide
cr	23953.47	0	1	Wave Height
cs	23953.62	0.15702	2	Wave Height
cr	23927.72	0	1	Wave Period
cs	23928.06	0.34249	2	Wave Period
cr	24039.31	0	1	Pressure
cs	24039.43	0.12377	2	Pressure
cr	23517.04	0	1	Period of the Day
cs	23517.24	0.19815	2	Period of the Day
cr	24229.47	0	1	Wind Speed
cs	24229.63	0.16380	2	Wind Speed
tp	24229.10	0	1	North–South Component (v)
cr	24234.68	5.58178	2	North–South Component (v)
tp	24069.22	0	1	East–West Component (u)
cr	24109.42	40.19259	2	East–West Component (u)

Source: Prepared by the author based on the reviewed literature (2026).

Likelihood (REML) play a critical role. Generally, REML is less prone to overfitting—making it the preferred choice for smaller sample sizes—whereas GCV offers greater computational efficiency when dealing with large environmental datasets.

The empirical optimization matrix indicated that the adaptive smoother (ad) achieved superior predictive performance for the temporal and thermodynamic covariates—specifically *Days*, *Temperature*, and *Tide*—where localized fluctuations in variance necessitate a flexible, spatially varying smoothing parameter. Conversely, the cubic

regression spline (cr) outperformed alternative bases for Wave Height, Wave Period, Atmospheric Pressure, Time of Day, and Wind Speed, demonstrating optimal efficiency in resolving continuous, smooth macro-environmental gradients. For the directional wind vectors, the thin plate regression spline (tp) provided the most robust isotropic fit for both the North–South wind and East–West wind geographic components, highlighting the necessity of unconstrained multidimensional smoothing to map complex atmospheric circulation dynamics.

These diagnostic variances underscore that the underlying ecological structure of the juvenile green turtle (*Chelonia mydas*) abundance matrix is highly heterogeneous, exhibiting non-linear spatial and temporal dependencies. Based on this optimized smoother framework, the final fitted global GAM architecture is specified as follows:

```
smoothing_model <- gam( Sightings ~ s(Days, bs="tp", k=60) + s(Temperature, bs="tp", k=30) + s(Tide, bs="tp", k=30) + s(Wave_Height, bs="cr", k=15) + s(Wave_Period, bs="cr", k=15) + s(Pressure, bs="cr", k=15) + s(Time_of_the_day, bs="cr", k=10) + s(North–South_Component, bs="tp", k=15) + s(East–West_Component, bs="tp", k=15), data = dad, family = nb(link = "log"))
```

5 Discussion

5.1 Ecological and Statistical Implications of Overdispersion

Overdispersion constitutes a pervasive structural feature in marine ecological datasets, particularly within population assessments of juvenile sea turtles, whose foraging aggregations are inherently characterized by high spatial patchiness and episodic localized influxes driven by micro-environmental fluctuations, transient thermal anomalies, and primary productivity pulses ; . Within an inferential framework, neglecting to explicitly account for this variance inflation would violate key parametric assumptions, inducing a severe artificial deflation of standard errors and a subsequent overestimation of test statistics. Consequently, failing to resolve overdispersion risks drastically inflating Type I error rates, leading to the erroneous identification of non-significant environmental covariates as critical ecological drivers when their apparent predictive influence is merely an artifact of stochastic noise or unmodeled spatial clustering.

5.2 Ecological Synthesis of Sighting Mechanics and Macro-Environmental Dynamics

The exploratory baseline of nearshore foraging activity across the designated monitoring stations revealed prominent asymmetric geometries, structural zero-inflation, and severe overdispersion—distributional landmarks widely documented in opportunistic ecological macro-fauna count matrices ; . These robust mathematical patterns reflect the highly stochastic, patchily aggregated nature of biological processes operationalized within dynamic, open coastal boundary layers. Mechanistically, the positive skewness and steep leptokurtic profiles act as quantitative indicators of episodic, high-density foraging pulses. At Station 1, the sharp tail inflation ($K = 6.79$) stations to specialized macro-aggregations driven by localized microhabitat resource pulses or transient fine-scale environmental optimization windows . A highly congruent intermittent habitat usage scheme was captured at Station 2 ($K = 6.03$), where the high-amplitude sighting anomalies indicate that juvenile green turtle (*Chelonia mydas*) densities fluctuate as a direct function of rapid, ephemeral primary productivity shifts or seasonal algal/macroalgal availability ; .

Spatial variations in the underlying zero-inflation rates further map the behavioral heterogeneity across the rocky shore landscape. At Station 3, the structural dominance of zero or near-zero blocks (first quartile = 0) isolates this sector as a highly irregular, condition-dependent foraging zone, rendering its utilization patterns strictly contingent upon localized physical thresholds or shifting prey arrays . Conversely, the lower kurtosis benchmark recorded at Station 4 ($K = 3.20$) denotes a more homogenous, low-intensity baseline utilization profile. This structural dampening suggests a structurally stable micro-environment, albeit one constrained by a systematically lower carrying capacity or diminished benthic food resources. Across all monitoring sectors, however, the persistent variance-to-mean inflation ($\sigma^2 > \mu$) mathematically invalidates the restrictive equidispersion assumptions of classical Poisson count models . This systemic overdispersion underscores the vital role of the Negative Binomial parameterization, which successfully absorbs the extra-Poisson biological noise and spatial clustering inherent to the target population .

Ultimately, the co-occurrence of strong non-linear distribution trends and heavy tail inflation across all monitoring sectors confirms that the operationalization of Generalized Additive Models (GAMs) equipped with multi-dimensional smoothing bases is uniquely suited to capture the underlying environmental heterogeneity without inducing predictive bias . From a macro-ecological perspective, these findings demonstrate that the target rocky shore functions as a highly heterogeneous mosaic habitat. Juvenile green turtle residency within this system is defined by prolonged periods of low-density baseline occupancy punctuated by localized, high-frequency aggregation pulses, a spatial patchiness fundamentally regulated by non-linear physical and biological drivers such as sea surface temperature, tidal dynamics, and primary macroalgal productivity gradients .

5.3 Methodological Validation of Distributional Adequacy

The localized spatial density of the bootstrapped stochastic cloud around the empirical sample coordinates, indicates high parametric stability and asymptotic stability in the estimation of statistical moments within the skewness–kurtosis domain. This non-parametric resampling protocol serves as a rigorous framework to quantify the structural uncertainties inherent to empirical estimators, a prerequisite particularly vital when navigating ecological data matrices characterized by severe variance inflation . The derived skewness and kurtosis coordinates formalize a highly right-skewed, heavy-tailed geometry, mathematically validating a baseline configuration wherein a dense concentration of structural zero or near-zero observations is interspersed with rare, high-magnitude aggregation anomalies. This structural topology is a signature of heterogeneous marine processes, where non-linear environmental gradients and patchiness give rise to extreme ecological count variances .

The comparative performance of the candidate distribution families underscores the structural suitability of asymmetric, heavy-tailed probability density functions for capturing overdispersed count data with prominent low-value inflation . Conversely, the numerical criteria and residual diagnostics definitively demonstrate the mathematical inadequacy of symmetric distribution families, such as the Normal framework, for modeling the sighting frequency matrix (Sightings). The substantial overdispersion and variance inflation observed

in the empirical data highlight the necessity of operationalizing error frameworks equipped to accommodate strong positive skewness and expansive tail behaviors, justifying the implementation of Poisson and Negative Binomial formulations under specialized mean-variance constraints. This strict agreement between the information-theoretic criteria and the observed empirical moments reinforces the diagnostic robustness of the benchmarking pipeline, confirming the utility of pairing likelihood-based verification with rigorous parameter optimization prior to final model specification.

Ultimately, the selection of the global stochastic architecture was guided not merely by descriptive adherence, but by a rigorous information-theoretic criteria framework balancing goodness-of-fit with mathematical parsimony. The comparative benchmarking of competing distribution candidates utilized the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) to penalize overparameterization and avoid localized likelihood inflation. This integrated cross-validation framework is crucial in complex ecological studies, where model specifications must adequately represent latent biological variability and the non-linear structure of underlying population dynamics without inducing predictive bias.

Consequently, to establish an unassailable foundation for the response variable across all monitoring stations, candidate family optimization was structured via maximum likelihood estimation (MLE) profiles, formal information criteria indexing, and multi-panel graphical diagnostics to explicitly resolve overdispersion and ensure compliance with modeling assumptions.

Sighting Station 1

Normal Distribution

The assumption of symmetry inherent to this continuous distribution proved completely incompatible with the empirical dataset. The target abundance matrix is characterized by a discrete domain, a high zero-inflation rate, and severe positive skewness, all of which systematically violate the core parametric assumptions of the Gaussian architecture, confirming its inadequacy for ecological count data.

Log-normal Distribution

The operationalization of an *ad hoc* data transformation (addition of a constant) was required to bypass structural zeros and execute the optimization pipeline. However, despite achieving intermediate descriptive advancements over the Normal distribution, this framework remains fundamentally suboptimal. Because it treats transformed counts as continuous values, it fails to explicitly preserve the discrete properties of the sighting process, nor does it possess an intrinsic mathematical mechanism to resolve structural overdispersion.

Poisson Distribution

The failure of the classical Poisson architecture is mathematically explained by the systematic violation of its restrictive equidispersion constraint, which assumes exact equality between the mean and the variance. Because the empirical sighting data exhibit a variance substantially higher than the mean, the Poisson model artificially deflates standard errors and drastically underestimates the true frequency of extreme aggregation events, generating unrealistic adjustments from an ecological perspective.

Negative Binomial Distribution

The integration of an independent dispersion parameter allows the Negative Binomial

framework to explicitly decouple the mean-variance relationship, successfully absorbing the excess biological noise and the severe overdispersion that collapsed the Poisson model. Multi-panel graphical diagnostics evidenced an exceptionally high degree of alignment within both the empirical density space and the quantile-quantile (Q-Q) coordinates, demonstrating that this mixture distribution is uniquely capable of simultaneously capturing the dual topology of high zero frequency and heavy-tailed aggregation anomalies .

Spatial Consistency Across Remaining Monitoring Stations (Stations 2, 3, and 4)

This competitive diagnostic hierarchy was consistently replicated across the remaining fixed monitoring sectors (**Stations 2, 3, and 4**). In all spatial strata, the Gaussian and Poisson formulations underperformed due to skewness constraints and overdispersion artifacts, whereas the Lognormal approach provided only localized, intermediate fits. The systematic superiority of the Negative Binomial architecture across the entire spatial domain was validated by minimized information-theoretic criteria (AIC and BIC) and superior tail alignment, showcasing its ability to represent the concurrent occurrence of low-frequency baselines and extreme foraging pulses .

Concurrently, these findings corroborate established ecological literature, confirming that the Negative Binomial framework constitutes the mathematical benchmark for resolving overdispersed population counts generated by behavioral clustering and environmental patchiness. Ultimately, pairing this robust error structure with the non-linear flexibility of Generalized Additive Models (GAMs) guarantees a statistically consistent, unassailable representation of juvenile green turtle (*Chelonia mydas*) foraging dynamics across the targeted rocky shore landscape .

5.4 Critical Evaluation of Basis Function Performance and Ecological Thresholds

The comparative assessment of smoothing basis performance detailed in Table 1 demonstrates that the ecological response of juvenile green turtles (*Chelonia mydas*) to environmental predictors is highly non-uniform, reflecting the multi-layered processes that regulate their distribution and foraging behavior within nearshore ecosystems. The structural dominance of the adaptive smoother (**ad**) for the temporal and thermodynamic covariates—specifically *Days*, *Temperature*, and *Tide*—indicates that these parameters exert non-linear, highly heterogeneous influences on abundance, characterized by abrupt shifts over localized intervals and relative stability across others. This pattern aligns with marine megafauna literature, which highlights that juvenile sea turtles exhibit intense sensitivity to macro-environmental variations, particularly thermal regimes and tidal cycles, both of which directly modulate benthic food availability and localized diving kinetics .

Crucially, this optimization protocol was executed within the framework of a fixed global model architecture where all predictor covariates were retained simultaneously. This methodological constraint ensures that the documented variations in performance are exclusively attributable to the mathematical properties of the smoothing bases and error distributions, rather than artifacts of variable inclusion or exclusion, thereby guaranteeing internal consistency across the benchmarking process.

The selection of adaptive smoothers for these core drivers confirms that the relationship between the physical environment and juvenile occurrence does not follow simple

monotonic trends but is governed by localized ecological thresholds and regime shifts. For instance, fine-scale thermal variations alter both reptilian metabolic rates and spatial habitat selection, whereas tidal cycles regulate access to shallow, macroalgae-rich rocky reefs, triggering non-linear behavioral responses along the environmental gradient. In contrast, variables such as wave height, wave period, atmospheric pressure, time of day, and wind speed achieved optimal calibration when coupled with cubic regression splines (*cr*). This indicates more stable, continuous, and gradual relationships with the response matrix. These factors appear to operate as environmental modulators with lower structural variance, influencing juvenile presence in a highly predictable manner. Nevertheless, the marginal differences in information criteria scores among competing smoothers for these parameters suggest that while these oceanographic gradients are influential, they exert a secondary, modulating impact compared to the direct trófico-thermal constraints imposed by temperature and tide.

For directional wind vectors, thin plate regression splines (*tp*) provided the most robust fit, highlighting the necessity of global isotropic flexibility when mapping these circulation components. From an ecological perspective, this requirement reflects the complex, indirect pathways through which atmospheric forcing impacts the nearshore environment, such as driving surface currents, altering localized nutrient transport, and dictating the spatial aggregation of drifting prey patches—factors that perturb juvenile spatial distribution in a highly non-linear fashion.

In synthesis, these dynamics reinforce that juvenile *Chelonia mydas* distribution across shallow coastal matrices emerges from the simultaneous interplay of multiple environmental forces operating across distinct scales. Integrating an ensemble of specialized smoothers within a unified generalized additive model proved essential to resolve this structural complexity, successfully capturing both threshold-type and gradual biological responses. Such a strategy is vital in landscape-scale ecology, where oversimplifying covariate relationships risks masking key demographic processes or generating misleading inferential conclusions. The finalized model optimizes the trade-off between capturing non-linear biological trends and maintaining structural parsimony, validating this variable-specific selection matrix as an effective routine for model optimization. Ultimately, these results confirm that no single smoothing basis is universally superior; rather, the selection is strictly contingent upon the stochastic properties of each individual predictor. Coupled with the Negative Binomial error distribution, this rigorous framework ensures that the identified abundance patterns are both statistically robust and ecologically plausible, accurately reflecting the dynamic habitat-use strategies of juvenile sea turtles within open coastal boundary layers.

6 Final Considerations

The integrated analysis of juvenile green turtle (*Chelonia mydas*) abundance matrix across the designated monitoring stations demonstrated that the underlying response architecture is characterized by strong positive asymmetry, severe tail inflation, and systematic overdispersion. These structural configurations constitute diagnostic landmarks typical of opportunistic megafauna counts within highly dynamic, open coastal boundary layers. The environmental consistency of these distributional patterns across all spatial strata indicates that the observed variance in abundance is not stochastically random; rather, it emerges as a direct consequence of multi-layered ecological processes coupled with fine-scale

environmental heterogeneity and episodic resource availability pulses.

The diagnostic phase dedicated to probability distribution family selection proved to be a critical prerequisite to guarantee mathematical coherence between the empirical sampling topology and the statistical modeling suite. The complete inadequacy of classical Gaussian and Poisson formulations—driven by the discrete, heavy-tailed, and overdispersed nature of the abundance counts—underscores the necessity of operationalizing more flexible parametric frameworks. In this context, the Negative Binomial distribution family achieved superior predictive performance, successfully accommodating the concurrent occurrence of high zero-inflation baselines and extreme foraging aggregation events while absorbing the extra-Poisson biological noise via its independent dispersion parameter. This structural calibration ensured heightened inferential robustness, stabilized standard errors, and maximized model generalization capacity, aligning with established guidelines in environmental and ecological statistics .

The systematic benchmarking of basis function performance revealed that the relationship between meteoceanographic predictors and juvenile abundance is highly variable, contingent upon the physical nature of each environmental forcing vector. The structural superiority of adaptive smoothers (*ad*) for the thermodynamic, tidal, and temporal coordinates indicates the presence of non-linear ecological thresholds and sharp behavioral regime shifts. Conversely, cubic regression splines (*cr*) proved mathematically sufficient to resolve more stable, continuous, and gradual macro-environmental gradients. Furthermore, thin plate regression splines (*tp*) demonstrated optimal efficiency in mapping the continuous isotropic behavior of multidimensional directional wind components, capturing indirect atmospheric forcing without inducing predictive bias.

From a macro-ecological perspective, these findings demonstrate that the occupancy dynamics of juvenile *Chelonia mydas* across shallow coastal matrices are regulated by a synergetic complex of environmental drivers operating at decoupled scales and intensities. Thermic and tidal variables, which directly modulate metabolic kinetics, macroalgal resource availability, and spatial access to nearshore rocky reefs, play a central role in dictating population aggregation patterns, whereas secondary oceanographic covariates act as continuous environmental modulators. This systemic complexity reinforces the vital role of advanced generalized additive architectures capable of preserving non-linear, heterogeneous relationships, thereby avoiding oversimplifications that could mask underlying demographic processes or compromise ecological interpretations.

Ultimately, the analytical cross-validation combining an appropriate mixture distribution error structure with localized smoothing parameter optimization enabled the formulation of statistically consistent, ecologically plausible models. This unified framework not only minimizes predictive artifacts but also provides a highly refined understanding of the latent biological processes regulating habitat-use strategies and site fidelity in juvenile sea turtles. Consequently, the outputs derived from this study establish an unassailable foundation for both methodological advancements in high-variance ecological data modeling and the broader spatial conservation ecology of marine megafauna in vulnerable nearshore ecosystems.

Acknowledgements

The author expresses sincere gratitude to the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), the Universidade Federal do Rio Grande (FURG), the Programa de Pós-Graduação em Ambientometria (PPG Amb), the Instituto Caminho Marinho (ICMar), the for the essential institutional, academic, and financial support provided throughout this research. Special thanks are extended to the Laboratório de Biometria & Conservação (BIOCO) for its invaluable contribution to the analyses, experiments, and activities that were fundamental to conducting this study. Finally, the author is deeply grateful to friends and loved ones for their endless patience, listening, and unwavering support during every crazy stage of this work.

Declarations

The authors declare no conflict of interest.

References

- Akaike H (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19(6):716–723.
- Bastos KV, Machado LP, Joyeux J-C, Ferreira JS, Militão FP, Fernandes VO, Santos RG (2022) Coastal degradation impacts on green turtle's (*Chelonia mydas*) diet: Nutritional richness and health. *Science of the Total Environment* 823:153593. <https://doi.org/10.1016/j.scitotenv.2022.153593>
- Bjorndal KA, Bolten AB, Chaloupka MY (2003) Survival probability estimates for immature green turtles (*Chelonia mydas*) in the Bahamas. *Marine Ecology Progress Series* 252:273–281.
- Bookstein FL (1989) Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11(6):567–585. <https://doi.org/10.1109/34.24792>
- Brasil (2000) Decreto de 14 de setembro de 2000. Dispõe sobre a criação da Área de Proteção Ambiental da Baleia Franca, no Estado de Santa Catarina, e dá outras providências. *Diário Oficial da União*, Brasília, DF, 15 Sep 2000.
- Burnham KP, Anderson DR (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, New York.
- Cameron AC, Trivedi PK (2013) *Regression Analysis of Count Data*, 2nd edn. Cambridge University Press, Cambridge.
- Centro de Informações de Recursos Ambientais e de Hidrometeorologia de Santa Catarina (EPAGRI/CIRAM) (2024) Meteorological and environmental data portal. Available at: <https://ciram.epagri.sc.gov.br/>. Accessed 16 Jul 2024.
- Dunson DB, Maity A (2015) Statistics in the big data era: Challenges and opportunities. *Statistics & Probability Letters* 101:1–2.
- Efron B, Tibshirani RJ (1993) *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Eilers PHC, Marx BD (1996) Flexible smoothing with B-splines and penalties. *Statistical Science* 11(2):89–121.

Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. *Journal of Animal Ecology* 77(4):802–813.

Fuentes MMPB, Limpus CJ, Hamann M, Dawson J (2011) Potential impacts of projected sea-level rise on sea turtle rookeries. *Aquatic Conservation: Marine and Freshwater Ecosystems* 21(2):213–222. <https://doi.org/10.1002/aqc.1198>

Fundação Projeto TAMAR (2024) Projeto TAMAR: Tartaruga-verde (*Chelonia mydas*). Available at: <https://www.tamar.org.br/>. Accessed 17 Jun 2025.

Google LLC (2024) Gemini. Available at: <https://gemini.google.com/>. Accessed 09 Jan 2026.

Hastie T, Tibshirani R (1986) Generalized additive models. *Statistical Science* 1(3):297–318. <https://doi.org/10.1214/ss/1177013604>

Hastie T, Tibshirani R (1990) *Generalized Additive Models*. Chapman & Hall, London.

Hays GC, Broderick AC, Godley BJ (2010) Climate change and marine turtles. *Endangered Species Research* 7(2):137–154.

Hilbe JM, Robinson AP, Myers RA (2011) (A) Modeling count data with overdispersion: An overview. *Ecological Modelling* 222(7):1333–1344.

Hilbe JM, Robinson AP, Bradley AL (2011) (B) *Negative Binomial Regression*, 2nd edn. Cambridge University Press, Cambridge.

Instituto Nacional de Meteorologia (INMET) (2024) Meteorological data portal. Ministério da Agricultura e Pecuária, Brazil. Available at: <https://portal.inmet.gov.br/>. Accessed 1 Oct 2023.

Instituto Nacional de Pesquisas Espaciais (INPE) (2024) CPTEC – Centro de Previsão de Tempo e Estudos Climáticos. Available at: <https://www.cptec.inpe.br/>. Accessed 1 Oct 2023.

James G, Witten D, Hastie T, Tibshirani R (2013) *An introduction to statistical learning: with applications in R*. Springer, New York.

Krivobokova T, Crainiceanu CM, Kauermann G (2008) Fast adaptive penalized splines. *J Comput Graph Stat* 17(1):1–20. <https://doi.org/10.1198/106186008X298440>

Magurran AE (2004) *Measuring biological diversity*. Blackwell Publishing, Oxford.

Marra G, Wood SN (2011) Practical variable selection for generalized additive models. *Comput Stat Data Anal* 55(7):2372–2387. <https://doi.org/10.1016/j.csda.2011.02.004>

McCullagh P, Nelder JA (1989) *Generalized linear models*, 2nd edn. Chapman & Hall, London.

O'Hara RB, Kotze DJ (2010) Do not log-transform count data. *Methods Ecol Evol* 1(2):118–122. <https://doi.org/10.1111/j.2041-210X.2010.00021.x>

OpenAI (2024) ChatGPT. Available at: <https://chatgpt.com/>. Accessed 9 Jan 2026.

Pedersen EJ, Miller DL, Simpson GL, Ross N (2019) Hierarchical generalized additive models in ecology: an introduction with mgcv. *PeerJ* 7:e6876. <https://doi.org/10.7717/peerj.6876>

Polis GA, Anderson WB, Holt RD (1997) Toward an integration of landscape and food web ecology: the dynamics of spatially subsidized food webs. *Annu Rev Ecol Syst* 28:289–316. <https://doi.org/10.1146/annurev.ecolsys.28.1.289>

Polovina JJ, Howell E, Kobayashi DR, Seki MP (2004) The transition zone chlorophyll front, a dynamic global feature defining migration and forage habitat for marine resources. *Prog Oceanogr* 61(2–4):257–264. <https://doi.org/10.1016/j.pocean.2004.06.001>

Posit Team (2024) RStudio: Integrated development environment for R. Posit Software, PBC, Boston. Available at: <https://posit.co/download/rstudio-desktop/>. Accessed 18 Jun 2024.

R Core Team (2024) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. Available at: <https://www.R-project.org/>. Accessed 18 Jun 2024.

Schabenberger O, Pierce FJ (2002) Contemporary statistical models for the plant and soil sciences. CRC Press, Boca Raton.

Seminoff JA (2004) *Chelonia mydas*. In: The IUCN Red List of Threatened Species 2004:e.T4615A11037468. International Union for Conservation of Nature (IUCN). Available at: <https://www.iucnredlist.org/species/4615/11037468>

Ver Hoef JM, Boveng PL, Quinn TJ (2007) Quasi-Poisson vs. negative binomial regression: ecological applications. *Ecology* 88(11):2766–2772. <https://doi.org/10.1890/07-0043.1>

Westfall PH, Henning KS, Howell JP (2014) Theoretical foundations of skewness and kurtosis in applied statistics. *Am Stat* 68(3):191–200.

Windguru (2024) Windguru weather forecasts and wind statistics. Windguru s.r.o. Available at: <https://www.windguru.cz/>. Accessed 17 Jun 2025.

Wood SN (2003) Thin plate regression splines. *J R Stat Soc Ser B Stat Methodol* 65(1):95–114. <https://doi.org/10.1111/1467-9868.00374>

Wood SN (2011) Fast stable direct fitting and gradient objective function evaluation of generalized additive models. *J R Stat Soc Ser B Stat Methodol* 73(1):3–36. <https://doi.org/10.1111/j.1467-9868.2010.00749.x>

Wood SN (2017) Generalized additive models: an introduction with R, 2nd edn. CRC Press, Boca Raton.

Zar JH (2010) Biostatistical analysis, 5th edn. Pearson Prentice Hall, Upper Saddle River.

Zuur AF, Ieno EN, Walker N, Saveliev AA, Smith GM (2009) Mixed effects models and extensions in ecology with R. Springer, New York.

Zuur AF, Ieno EN, Elphick CS (2010) A protocol for data exploration to avoid common statistical problems. *Methods Ecol Evol* 1(1):3–14. <https://doi.org/10.1111/j.2041-210X.2009.00001.x>

Zuur AF, Hilbe JM, Ieno EN (2013) A beginner's guide to GLM and GLMM with R: A frequentist and Bayesian perspective for ecologists. Highland Statistics Ltd, Newburgh.