THE UNIVERSITY OF EDINBURGH

SCHOOL OF MATHEMATICS

Tracking Freshwater Biodiversity Through Statistical Forecasting of Macroinvertebrate Indicators with GAM-based Modeling Regression

Master of Science in Statistics with Data Science

Author: Xu Guo
Student Number: S2743905
Instructor: Simon Wood and Nicole Augustin
Date: 15 Aug 2025
Total Words: 4972 (Body Text)

Exclusive Summary

This research overcomes the limitation of operationalizing long-term freshwater biodiversity trends from the UK Environment Agency's heterogeneous macroinvertebrate dataset. Historical data mix exact counts and interval-censored categories, usually with plentiful zeros, providing complexity for analysis [1]. To combine these formats and handle temporal, geographical, and seasonal variation, a single framework based on generalized additive models (GAMs) with censored Poisson likelihood [2] was created. Presence–absence was modeled by binomial GAMs with high predictive power (AUC ¿ 0.97). Abundance was modeled with censored Poisson GAMs, coupled modelling of counts and category data without midpoint bias. Validation with 2025 data verified the strong performance, and projections to 2030 signified taxon-specific trends—like declines of *Aphelocheiridae*—in response to environmental change. The framework retains the historical ecological signals with accuracy, providing a powerful, transferable tool for forecasting biodiversity and managing freshwater ecosystems.

Own Work Declaration

I declare that this thesis is entirely my own work, except where otherwise indicated. All sources have been acknowledged, and any help received has been fully credited.

Generative AI Acknowledgment

In this project, I acknowledge the use of GPT-4o (https://openai.com) for code modification and optimization, particularly in enhancing comments, checking and correcting errors, as well as improving code spacing and efficiency. Additionally, it assisted in generating ideas and aiding my learning of statistical concepts.

Contents

1. Introduction

1.1 Biological Background

Freshwater ecosystems are vital components of global biodiversity and play an essential role in maintaining ecological balance. Among these, riverine habitats are of particular importance, and the health status of these systems has profound implications for regional ecological security [3]. Benthic macroinvertebrates, as key biological indicators of water quality and ecological conditions, can integrate and reflect cumulative environmental pressures over time, and are therefore widely used in river monitoring and management programs [4, 5]. However, long-term monitoring programs are often subject to changes in sites, sampling methods, and recording protocols, particularly the shift from categorical abundance classes (interval data) to numerical counts. This variation introduces complexity and uncertainty in assessing long-term trends in species abundance [6].

Macroinvertebrate monitoring data are highly heterogeneous: some records provide exact count data, while others report abundance as intervals, and the interval boundaries themselves can differ between observations. Such mixed data often contain substantial numbers of zeros and involve censoring or interval uncertainty. Conventional generalized linear models or simple transformations (e.g., replacing interval data with the geometric mean of the class) can lead to bias and unstable trend estimation [1]. Therefore, there is a need to develop statistical models that can simultaneously handle exact counts and interval-censored observations, enabling more comprehensive use of historical monitoring data and improving inference accuracy.

Our work aims to build a statistical platform with generalized additive models (GAMs) and censored Poisson likelihood [2] to simultaneously investigate interval-censored abundance and accurate counts of target taxa with temporal trends. The new approach can readily incorporate flexible temporal, geographic, and environmental covariates and handle the complex distributional characteristics of monitoring datasets. The work will apply the new platform based on the long-term monitoring dataset of macroinvertebrates in order to provide more reliable quantitative evidence for riverine system ecological status assessment and offer methodological aid for integrating the information, forecasting the trend, and decision-making for water quality management.

1.2 Data Illustration

The datasets employed here come from the England Environment Agency's (EA) open-access macroinvertebrate monitoring datasets. The datasets contain three primary constituents: (1) biological recordings of benthic macroinvertebrates matched to the family level (e.g., *Brachycentridae*, *Odontoceridae*, *Cord-*

ulegastridae, and *Aphelocheiridae*) with linked abundance information, (2) sample-level metadata with information such as sample codes, dates of sampling, methods, and codes for analysis, and (3) site-level metadata with information on monitoring locations, including geographic attributes such as EA reporting region and waterbody type. The datasets cover multiple decades, the 1970s onwards to the current day, and are in the public domain and can be downloaded from the EA's Ecology and Fish Data Explorer portal [4].

A feature of the datasets is the wide temporal and methodological variation. During the period prior to the year 2000, abundance of the macroinvertebrates used to be frequently recorded in categorical abundance classes (e.g. A: 1–9 individuals, B: 10–99 individuals, C: 100–999 individuals), but thereafter the data were mostly recorded as numerical abundance estimates, usually given to one significant figure and subject to subsampling corrections in the event of the presence of highly-abundant samples [6]. These incompatibilities create problems for the determination of long-term trends since the categorical entries are interval-censored and numerical entries have uncertainty associated with them as well (e.g. "20" for 15–24 individuals).

The monitoring data are also distinguished by extreme spatial and sampling imbalance. Monitoring effort differs significantly throughout EA report regions, with some regions and taxonomic groups sampled more frequently than others, causing inequalities in the size of samples and temporal coverage [5]. Most of the taxa, particularly of the rarer families, consist of large percentages of zero observations, which could represent true absences or detectability limitations in the course of field surveys [1]. These traits in unison point to the necessity of statistical methods able to deal with the exact counts, interval-censored information, excess zeroes, and spatio-temporal variability all at once, for example, GAMs with grouped or censored Poisson likelihoods [2].

2. Data Process

2.1 Data cleaning and selection of EA regions

We extracted four ecologically relevant macroinvertebrate family records, *Brachycentridae*, *Odontoceridae*, *Cordulegastridae*, and *Aphelocheiridae*, from the EA harmonized family-level dataset. They were chosen for a range of ecological and practical reasons: each has a single representative in the UK dataset, to prevent taxonomic uncertainty when inferring abundance; they belong to the `WHPT_METRICS_B` harmonized metrics employed in EA surveys, to enable consistent monitoring over the assessment period; and cover different ecological and functional groups, from cased caddisflies to bugs and dragonflies. Furthermore, they have varying data densities and geographical distribution, facilitating sound modeling of temporal changes under varying sampling conditions (Table 2.1).

Table 2.1: Summary of the four selected macroinvertebrate families. Each family is represented by a single species in UK records and belongs to a distinct functional group. Sampling counts are based on harmonized EA data.

Family	UK Species	Functional Group	Sample Count
Brachycentridae	*Brachycentrus subnubilus*	Sensitive insect	13,223
Odontoceridae	*Odontocerum albicorne*	Sensitive insect	13,232
Cordulegastridae	*Cordulegaster boltonii*	Predatory insect	4,765
Aphelocheiridae	*Aphelocheirus aestivalis*	Aquatic bug	6,244

These files were combined with sample-level metadata (sample IDs, dates, methods, codes for analysis) and site-level metadata (site IDs and EA reporting regions). Samples collected with the S3PO method and one of three analyses (ANAA, ANLA, ANLE) were kept. Other temporal variables (year, month, season) were created for downstream modeling. The final dataset was in long format with a family ID. Monitoring effort was then calculated per EA region by the number of distinct sites and samples per family. Regions were sorted by the number of samples and displayed as stacked bar plots (Figure 2.1). From this, three regions—**Devon, Cornwall and the Isles of Scilly**, **Solent and South Downs**, and **Wessex**—were taken for additional analysis because of the large sampling effort and full coverage of all four families.

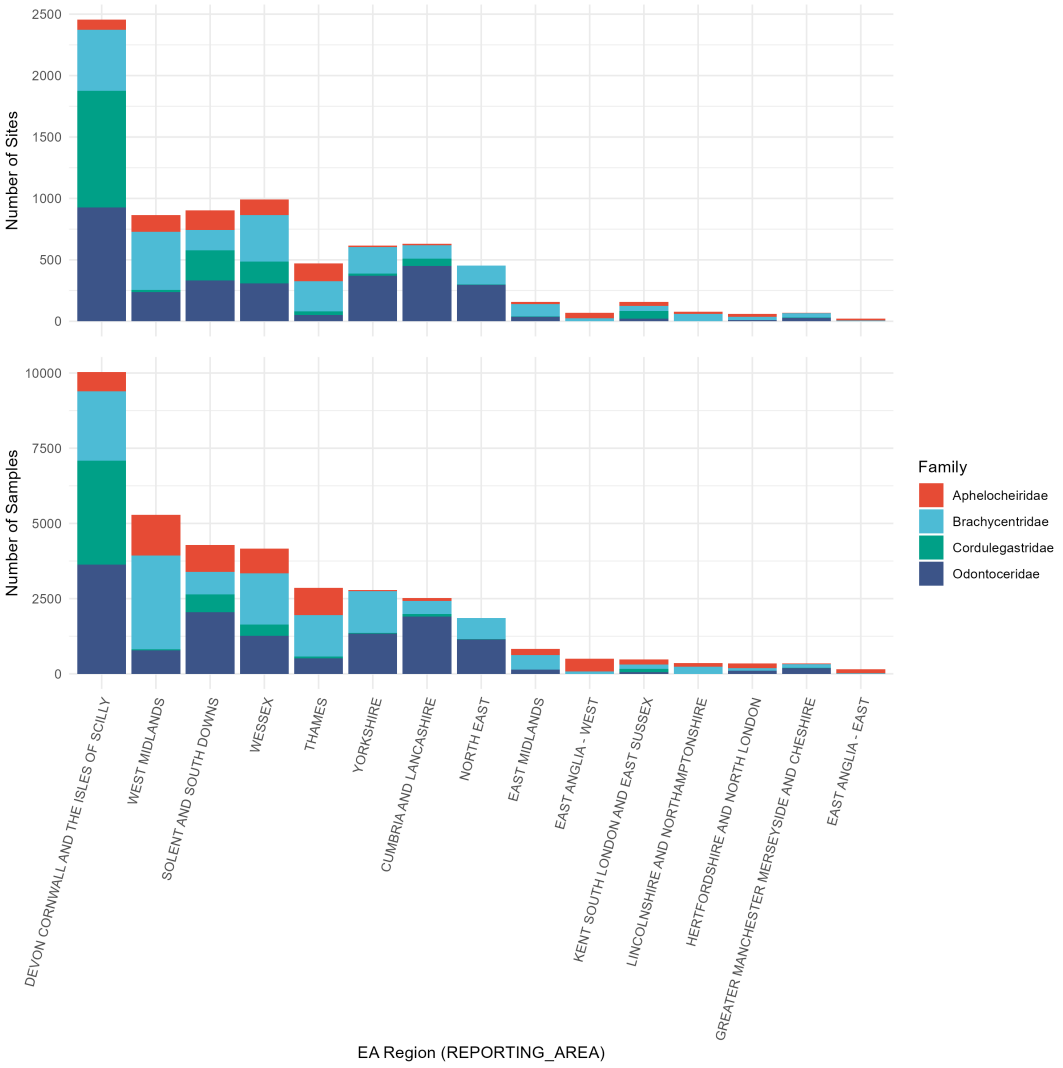

Figure 2.1: Number of monitoring sites (upper panel) and samples (lower panel) by EA reporting region for the four target families. Colors indicate families: *Aphelocheiridae* (red), *Brachycentridae* (light blue), *Cordulegastridae* (green), and *Odontoceridae* (dark blue). Regions are ordered by total sample count.

2.2 Annotation of Categorical and Numerical Abundance Values

To prepare the dataset for downstream analyses, all samples recorded prior to 1990 were removed due to the high prevalence of missing or inconsistent metadata. Furthermore, although the `ANLE` field is intended to contain categorical entries, we observed that it occasionally includes values resembling numerical abundance counts. Meanwhile, it is not possible to guarantee that the `ANAA` field is entirely free of categorical values. These inconsistencies likely reflect manual data entry errors or legacy recording practices. To minimize classification bias, we therefore adopted a unified classification scheme based solely on abundance values, independent of the field of origin.

Types of records were classified by a site-specific and rule-based process. To begin with, any abundance value of 3, 33, 333, or 3333 was classified provisionally as categorical, and all other values as numerical. This heuristic mirrors the EA monitoring protocol, in which such four values are commonly

used as the geometric placeholder in the evaluation of the categories. Nonetheless, since the agency had switched over to numerical recording sometime after 2000 and prior to 2001, a static rule alone is not adequate. We therefore implemented a site-level "conversion point"—the first sample date after which all subsequent records in a site are considered to have followed numerical recording. A site with no information prior to the year 2000 had all the samples automatically classified as numerical. Otherwise, every pre-2001 candidate numerical entry was tested with a hypothesis test in order to decide if such entry indicated true transition to numerical recording or merely indicated a spurious entry during a still-categorical regime. If the null-hypothesis could not be rejected, the candidate and subsequent samples were classified as numerical entries. The complete decision logic is given in the form of the Algorithm 1, and the technicalities of the test procedure are elaborated upon in **Appendix A.1**.

Algorithm 1: Inferring categorical and numerical abundance values

Input: Dataset with `SITE_ID`, `SAMPLE_DATE`, and `TOTAL_NUMBER`

Output: Updated `record_type` for all samples

foreach `site` *in unique sites* **do**

- Sort samples of `site` by `SAMPLE_DATE`;
- **if** *first sample date* $\geq$ *Jan 1, 2000* **then**
  - Mark all samples as **numerical**;
  - **continue**;
- Initially mark 3, 33, 333, 3333 as **categorical**, others as **numerical**;
- **foreach** *candidate sample marked* ***numerical*** *(chronologically)* **do**
  - **if** *candidate date* $\geq$ *Jan 1, 2001* **then**
    - Mark candidate and all subsequent samples as **numerical**;
    - **break**;
  - Define *check interval* = samples between candidate date and Jan 1, 2001 (excluding endpoints);
  - Count n_3, n_{33}, n_{333}, n_{3333} in the interval;
  - Compute

$$P = (1/10)^{n_3}(1/100)^{n_{33}}(1/1000)^{n_{333}}(1/10000)^{n_{3333}}$$

  - **if** $P > 1/1000$ **then**
    - Accept candidate as true conversion point;
    - Mark candidate and all subsequent samples as **numerical**;
    - **break**;

Merge all updated sites into a single dataset;

Following classification, we analyzed the temporal and regional distribution of record types. As shown

in Figure 2.2, the number of categorical records declined gradually throughout the 1990s, followed by an abrupt drop in 2001 across all three EA regions. Based on this observation, we reclassified all samples dated 2001 or later as numerical, even if their values matched the categorical placeholders (e.g., 3, 33, 333, or 3333). This override ensures consistency with post-2000 monitoring protocols and eliminates residual ambiguity in more recent records. The final classification result—after applying both the conversion point inference and the 2001 override—is summarized by taxonomic family and region in **Appendix A.1**.

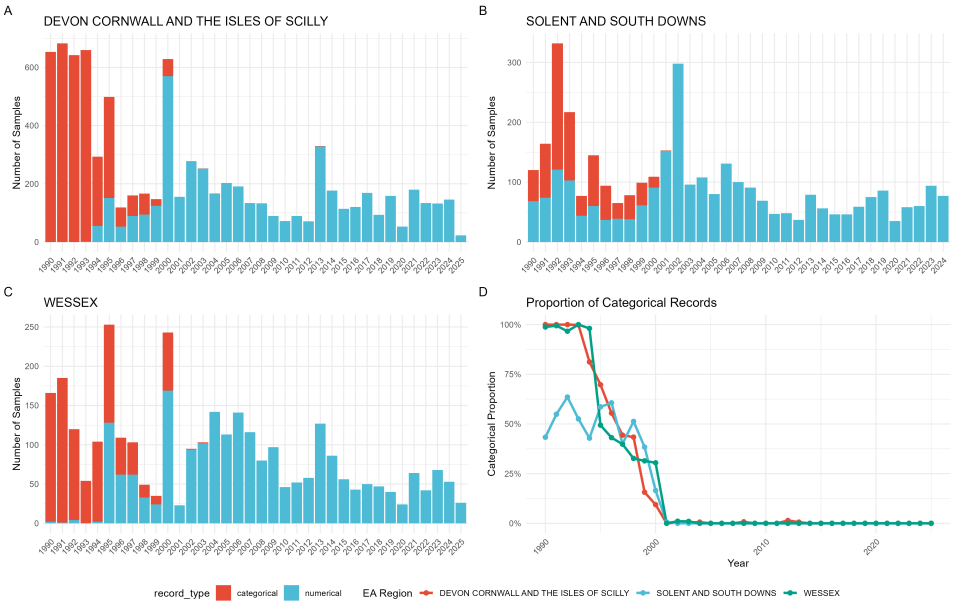

Figure 2.2: Proportion of categorical and numerical records across EA regions and years. Panels A–C show the annual number of biological samples for **Devon, Cornwall and the Isles of Scilly**, **Solent and South Downs**, and **Wessex**, respectively, with bars stacked by record type. Panel D shows the temporal decline in categorical records. The sharp drop after 2000 confirms that EA sites had almost universally transitioned to numerical recording by 2001.

3. Variable Analysis

3.1 Completing Zero-Count Records

In ecological monitoring datasets, the absence of a record for a given taxonomic family in a sample does not necessarily indicate that the family was not considered; rather, it usually means that no individuals from that family were observed. To ensure that zero counts are explicitly represented, we supplemented the dataset with additional records for all missing family–sample combinations, setting their `TOTAL_NUMBER` to zero.

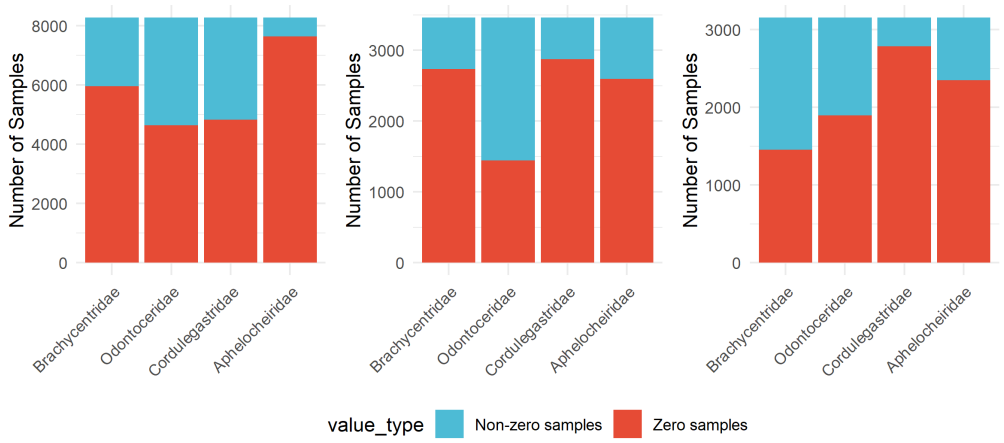

Figure 3.1: Number of zero and non-zero samples by family and EA region. Stacked bar plots show the number of samples where the total count for the family is zero (red) versus non-zero (blue). Data are shown for the four families across the three selected EA regions. These results illustrate the presence of zero-value and non-zero-value records within each family–region combination, and also demonstrate clear differences in sampling intensity across regions and families.

After filling in the zero counts, we investigated the distribution of zero-value and non-zero-value samples for the four target family groups (*Brachycentridae*, *Aphelocheiridae*, *Odontoceridae*, and *Cordulegastridae*) over the three sampled EA regions (**Devon, Cornwall and the Isles of Scilly**, **Solent and South Downs**, and **Wessex**). The stacked bar charts (Figure 3.1) indicate for each family in each region that there are both zero-value and non-zero-value records, indicating the natural variation in detections in the samples. Obvious differences are also evident in the aggregate counts of samples over the three regions, and there is considerable variation by family for the same region. This indicates heterogeneity in both taxonomic representation and geographical coverage. Note that the counts here are of samples where there could potentially be detection of taxa of the respective family, rather than the presence of abundance records.

3.2 Comparison of Categorical and Numerical Abundance Across Time and Region

In order to explore temporal and spatial patterns in both numerical and categorical abundance data, we prepared a 4-column figure panel for each of the four target family groups (*Brachycentridae*, *Aphelocheiridae*, *Odontoceridae*, and *Cordulegastridae*) (Figure 3.2). The leftmost column of each panel aggregates the distribution of categorical placeholder abundance values (`3`, `33`, `333`, `3333`) observed between 1990 and 2000. All four families exhibit a consistent declining trend in the proportion of the lowest category (`3`), suggesting a temporal increase in recorded abundance during the pre-numerical era. This trend may indicate progressive improvements in detectability or actual increases in population levels prior to the transition to numerical recording. The second column presents the annual mean numerical abundance for each family, visualized through point and line plots. Superimposed linear trend lines (red dashed, fitted via `method = "lm"` in `ggplot2`) confirm the presence of a general upward trend, reflecting either increasing abundance or enhanced recording sensitivity over time.

The third and fourth columns highlight regional comparisons among the three selected EA regions (**Devon, Cornwall and the Isles of Scilly**, **Solent and South Downs**, and **Wessex**). The third column displays stacked bar plots of regional categorical proportions, while the fourth column visualizes log-transformed numerical abundances using boxplots. Both data types exhibit substantial and statistically significant regional variation across all four families. Results from pairwise Wilcoxon tests [7] indicate that many of these differences are statistically significant ($p < 0.05$), suggesting that abundance levels may be systematically influenced by regional environmental conditions or differences in monitoring effort.

3.3 Effects of Site and Season on Abundance Patterns

In order to determine the appropriateness of including environmental and geographical aspects in abundance modeling, we explored the effect of *Site* and *Season* for the four target families (Figure **??**). Each family has four plots: site-level categorical abundance proportions and site-level mean–variance relationships (columns 1–2), and seasonal categorical proportions and seasonal numerical abundance distributions (columns 3–4).

The first two columns exhibit apparent spatial heterogeneity. Column 1 presents the distribution of site-level categorical abundance values (`3`, `33`, `333`, `3333`) with broad variation in category proportions across sites. Column 2 presents each site's mean and variance in log-transformed abundance with a regression line (red dashed) fit to the points, illustrating a positive mean–variance relationship characteristic of overdispersion. The large site-to-site variability here strongly supports the use of *Site* as a random effect in future mixed-effects modeling.

Columns 3 and 4 show seasonal patterns. Column 3 provides the seasonal distribution of the

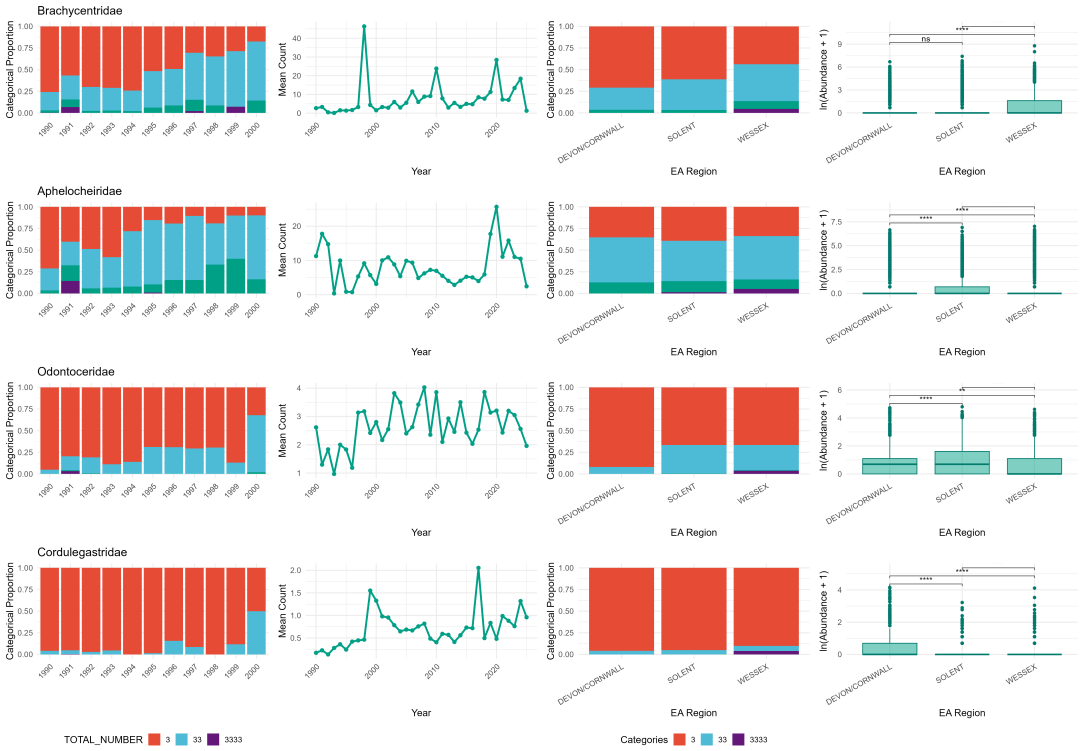

Figure 3.2: Four-panel comparison of categorical and numerical abundance patterns by family. Each row corresponds to a different taxonomic family. Column 1: proportions of categorical values (3, 33, 333, 3333) from 1990–2000. Column 2: annual mean numerical abundance with regression trend (red dashed line). Column 3: categorical placeholder proportions by EA region. Column 4: log-transformed numerical abundance by region with Wilcoxon test comparisons. Together, the plots reveal temporal increases in abundance and clear spatial variation across regions. Significance codes: "ns" = not significant ($p \geq 0.05$), * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$.

categorical values, demonstrating that there are some families where there are elevated proportions of larger categories in the warmer seasons. Column 4 shows log-transformed numerical abundance by season, with results of Wilcoxon rank-sum tests ($p < 0.05$, after Bonferroni adjustment) for pairwise comparisons. Significant differences are seen in several families, indicating that *Season* has a detectable impact on the level of abundance. These seasonal effects and the lateral variability combine to support the importance of including both in the models of abundance dynamics.

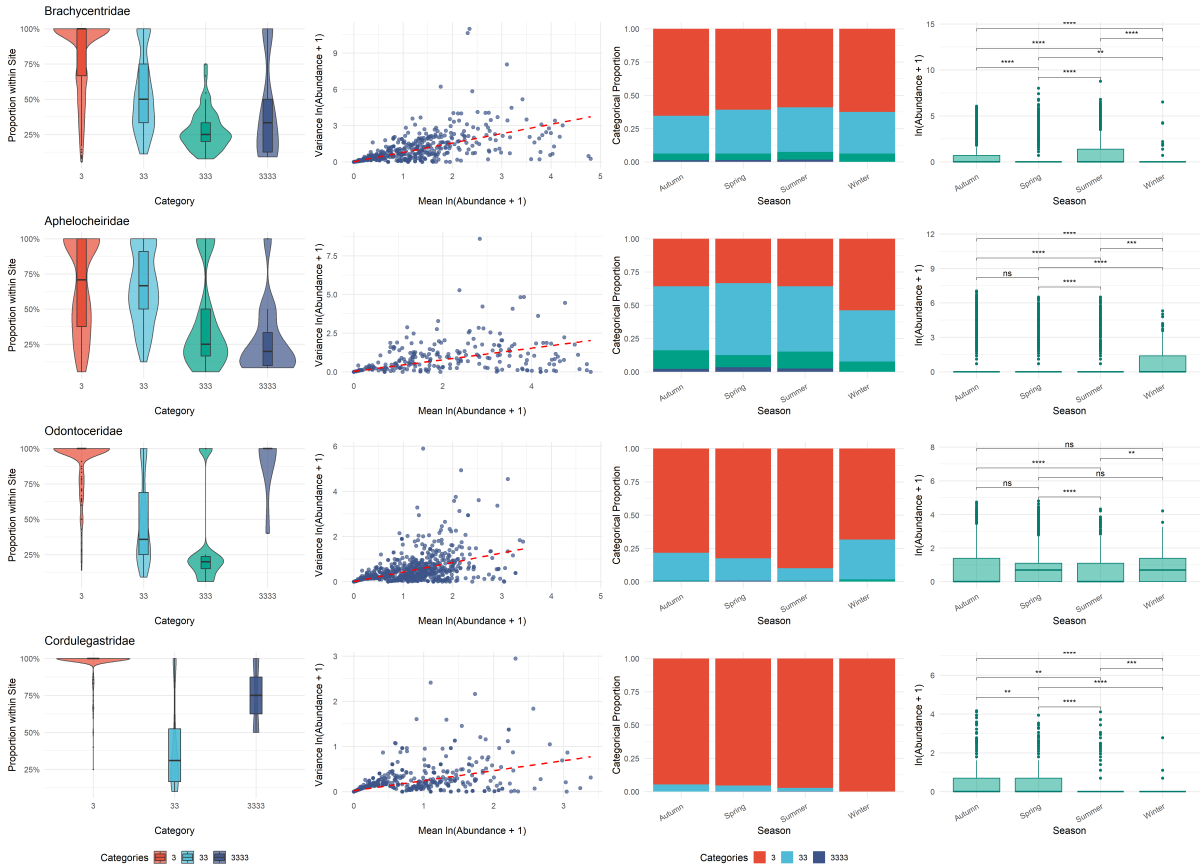

Figure 3.3: Spatial and seasonal variation in categorical and numerical abundance for four target families. Each row corresponds to one family. Column 1: violin and boxplots of site-level categorical proportions. Column 2: scatterplots of site-level mean vs. variance of log-transformed abundance with linear fit. Column 3: seasonal stacked bar charts of categorical proportions. Column 4: seasonal numerical abundance with Wilcoxon pairwise tests (Bonferroni-adjusted). Significance codes: "ns" = not significant ($p \geq 0.05$), * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$.

4. Regression Analysis

4.1 Conversion of Abundance Data to Interval Format

To incorporate measurement uncertainty and harmonise categorical and numerical records, all abundance observations were converted into integer-valued intervals for censored Poisson regression.

Categorical records were mapped to fixed integer ranges following the Environment Agency (EA) protocol:

Code	→	**Interval** [lower, upper]
`3`	→	$[1, 9]$
`33`	→	$[10, 99]$
`333`	→	$[100, 999]$
`3333`	→	$[1000, 9999]$

All other values, including `0`, were treated as exact counts.

Numerical records were transformed into fixed-width integer ranges according to abundance magnitude, reflecting increased counting error at higher counts during field sampling:

Value range	→	**Interval width**
[0, 14]	→	exact value (no range)
[15, 154]	→	step size = 10
[155, 999]	→	step size = 50
[1000, 9999]	→	step size = 200
$\geq$ 10000	→	$[10000, \infty)$

For each record, the `lower` and `upper` bounds and the censoring category (`"interval"` or `"none"`) were noted. This practice captures abundance magnitude information and explicitly codes uncertainty such that both category and numerical records are modeled uniformly under the censored Poisson paradigm. Furthermore, to enhance the stability of the subsequent regression analyses, records from before 1995 were omitted since this year corresponded with the first implementation of policies for the treatment of wastewaters and had seen particularly large interannual variability in the abundance of species, with 1991 presenting particularly extreme abundances. Records from 2025 were also omitted because of incompleteness.

4.2 Model specification

The model incorporates temporal, spatial, and seasonal covariates selected to represent key ecological and sampling structures in the macroinvertebrate monitoring data. Year is modeled using a penalized

cubic regression spline to capture nonlinear long-term trends without imposing a restrictive parametric form, accommodating gradual changes and multi-year cycles in abundance. Sampling site (SITE_ID) and reporting area (REPORTING_AREA) are included as random effect smooths to account for hierarchical spatial structure and unobserved heterogeneity; SITE_ID represents a finer spatial unit nested within REPORTING_AREA (e.g., one Environment Agency region may contain over 1000 sites, while another may contain around 800 sites). Modeling these factors as random effects enables partial pooling across groups and improves estimation stability in the presence of unbalanced or sparse sampling. Season is treated as a fixed effect because it comprises a small number of predefined categories with direct ecological interpretation, allowing for explicit estimation of level-specific contrasts relative to a reference category.

GAMs provide a flexible framework for combining smooth functions of continuous covariates with fixed and random effects for categorical and hierarchical factors [8, 2]. This flexibility is particularly suited to ecological time series, which often exhibit nonlinear responses to temporal and spatial gradients. The `mgcv` package in R [9] implements GAMs using penalized regression splines, supports a variety of basis types (including random effect smooths via `bs = "re"`), and allows joint modeling of multiple effect types within a single likelihood. Smoothing parameters are selected by restricted maximum likelihood (REML), balancing model fit and smoothness. In this study, smooth temporal trends, hierarchical spatial variation, and seasonal fixed effects are estimated simultaneously using `mgcv`, with technical details of model fitting and computational settings provided in **Appendix A.2**.

Formally, the model is expressed as

$$\log \mu_i = f(\text{year}_i) + b^{\text{site}}_{j(i)} + b^{\text{region}}_{k(i)} + \boldsymbol{x}_i^\top \boldsymbol{\beta}, \tag{4.1}$$

where $f(\text{year}_i)$ is a penalized cubic regression spline for year, $b^{\text{site}}_j \sim \mathcal{N}(0, \sigma^2_{\text{site}})$ are random intercepts for site, $b^{\text{region}}_k \sim \mathcal{N}(0, \sigma^2_{\text{region}})$ are random intercepts for reporting area, and $\boldsymbol{x}_i^\top \boldsymbol{\beta}$ represents fixed seasonal effects. This linear predictor is combined with the appropriate count distribution in the likelihood (Appendix A.2) to yield the full generalized additive mixed model used in the analysis.

4.3 Logistic regression analysis

We modelled presence–absence for each target family using binomial generalized additive models (GAMs) with an additive predictor for temporal and spatial structure (see Appendix A.3 for details on our logistic additive modelling) [10]. Let $Y_i \in \{0, 1\}$ denote occurrence and $p_i = \Pr(Y_i = 1)$ the occurrence probability. The distributional assumption is

$$\Pr(Y_i = y_i \mid p_i) = p_i^{y_i}(1 - p_i)^{1-y_i}, \qquad y_i \in \{0, 1\} \tag{4.2}$$

This presence model complements the count-based framework and directly supports the study goal of capturing temporal trends while flexibly incorporating temporal and spatial covariates for ecological status assessment [2, 10]. Models were estimated in `R` using `mgcv::bam()` with discrete smoothing and parallel computation to enable efficient fitting across all families.

Table 4.1: Main summary of fitted binomial GAMs (per family): deviance explained (%), Akaike Information Criterion (AIC), area under the ROC curve (AUC), edf of $s(\text{year})$.

Family	Dev. expl. (%)	AIC	AUC	edf(s(year))
Aphelocheiridae	86.2	2848.9	0.998	1.44
Odontoceridae	69.2	7391.8	0.982	5.35
Brachycentridae	63.9	6895.8	0.972	6.04
Cordulegastridae	66.8	6819.7	0.978	7.27

Table 4.1 shows that the fitted binomial GAMs perform strongly (AUC = 0.972–0.998; deviance explained = 63.9–86.2%; [11]), and that the *year* smooth is the dominant structured effect: edf of $s(\text{year})$ spans ≈ 1.44 (*Aphelocheiridae*, nearly monotone decline) to ≈ 7.27 (*Cordulegastridae*, sustained increase with curvature), with intermediate, gently undulating patterns in *Odontoceridae* and *Brachycentridae*. Figure 4.1 (Fig. 4.1) visualizes these results: column 1 compares observed yearly occurrence rates with mean predicted probabilities (showing close alignment); column 2 plots $s(\text{year})$ on the probability scale with 95% confidence bands (revealing a decline in *Aphelocheiridae*, a shallow hump then softening in *Odontoceridae*, a U-shaped pattern in *Brachycentridae*, and an increase through the 2010s–2020s in *Cordulegastridae*); column 3 highlights substantial dispersion in SITE_ID random effects, while column 4 shows more moderate contrasts across REPORTING_AREA; column 5 presents ROC curves consistent with the high AUCs. Seasonal fixed effects are relatively minor compared to the intercept; detailed estimates and term-level tests (via `mgcv::anova.gam`) are included in Appendix A.4. Standard diagnostics suggest approximately normal deviance residuals, no major structure in residuals versus fitted values, negligible autocorrelation, and good model calibration, supporting the adequacy of these models for inference on spatial–temporal occurrence patterns [2, 12].

In general, the binomial GAMs form a powerful baseline for species occurrence modeling. They exhibit high discriminatory power (AUC > 0.97 for all taxa) and successfully capture dominant temporal signals via the *year* smooth $s(\text{year})$, as visualized in Figure 4.1. The long-term trajectories revealed by this smooth indicate clear taxon-specific differences in ecological response to environmental change. For example, *Aphelocheiridae* exhibits a sustained decline in predicted occurrence since the early 2000s, likely reflecting reduced organic pollution and increased oxygen availability, as this family is relatively tolerant of low-oxygen environments. In contrast, *Odontoceridae* and *Brachycentridae*, which are more sensitive to pollution, show a decline in the early 2000s followed by partial recovery in recent years—potentially reflecting improvements in water quality and catchment management. Although *Cordulegastridae* has lower overall occurrence rates, it shows a recent increase in predicted presence. These patterns are further supported by the random-effect structure, where variation across SITE_ID exceeds that across

REPORTING_AREA, underscoring the influence of localized habitat conditions. While binomial GAMs effectively unify heterogeneous monitoring data and provide well-calibrated occurrence probabilities, they do not capture abundance gradients or interval-censored uncertainty. This limitation motivates the adoption of censored-Poisson GAMs for more detailed abundance modeling within the vector generalized additive modeling framework [13].

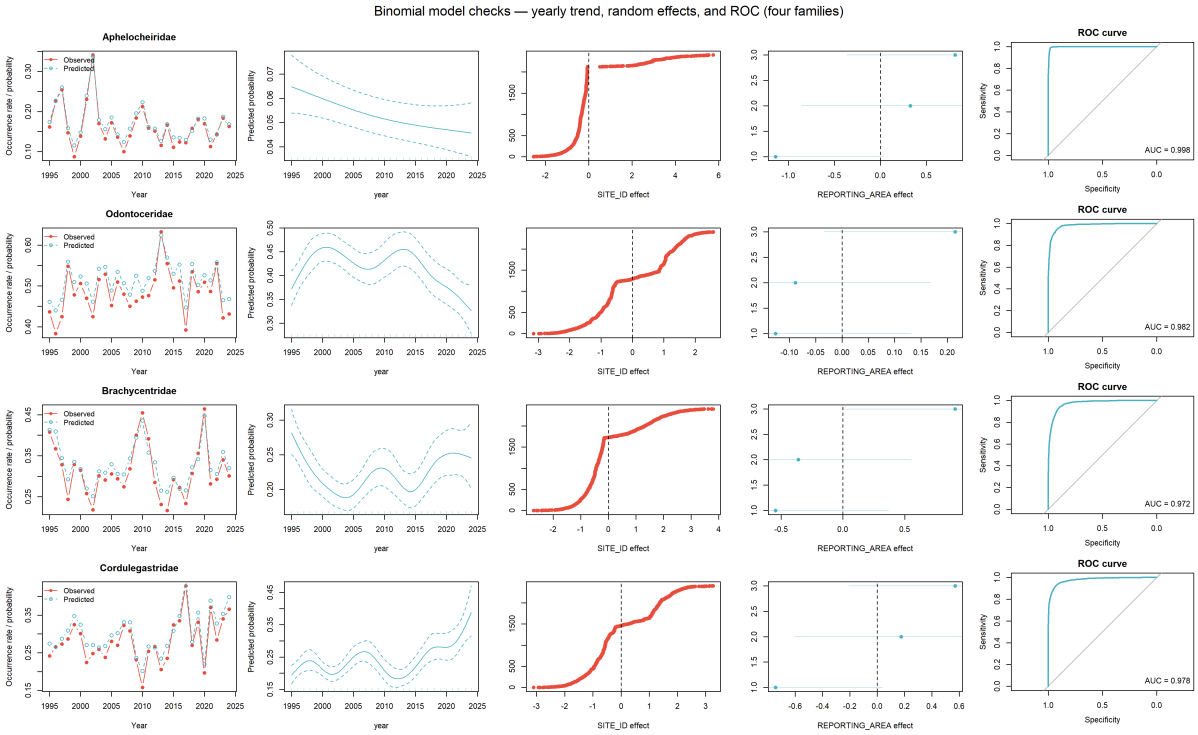

Figure 4.1: Binomial model checks—yearly trend, random effects and ROC for four families. Each row corresponds to one family. Column 1: observed vs. predicted yearly mean occurrence. Column 2: partial effect of *year* on the probability scale (smooth $s(\text{year})$ with 95% shading). Column 3: SITE_ID random effects (caterpillar plot). Column 4: REPORTING_AREA random effects (dot ± 95% CI). Column 5: ROC curve with AUC.

4.4 Censored Poisson regression analysis

We model abundance using a censored Poisson regression that jointly analyzes exact counts and interval-censored records, reflecting changes in recording protocols and counting uncertainty; by letting both data types contribute directly to the likelihood, this approach avoids midpoint–substitution bias and propagates measurement uncertainty into parameter estimates [1]. Specifically, the unobserved true abundance $N_i \sim \text{Poisson}(\mu_i)$, with μ_i linked to covariates via the additive predictor in Equation (4.1); for an observation with bounds (L_i, U_i), the likelihood contribution is

$$\Pr(L_i \le N_i \le U_i \mid \mu_i) = \sum_{n=L_i}^{U_i} \frac{e^{-\mu_i}\mu_i^n}{n!}, \tag{4.3}$$

where exact counts correspond to $L_i = U_i$. The full likelihood combining Equation (4.3) with the linear predictor appears in Appendix A.5; we implemented the model in R (`mgcv`; [2]) with custom likelihood evaluation for censored observations, enabling consistent estimation across heterogeneous

historical categorical records and modern numerical counts within a single framework.

Table 4.2: Main summary of fitted censored–Poisson GAMs (per family): deviance explained (%), Akaike Information Criterion (AIC), root mean squared error (RMSE), mean absolute error (MAE), and edf of $s(\text{year})$.

Family	Dev. expl. (%)	AIC	RMSE	MAE	edf(s(year))
Aphelocheiridae	86.1	64466.1	86.7	8.5	8.91
Odontoceridae	71.0	38157.4	11.5	2.7	8.76
Brachycentridae	73.0	131718.1	121.2	11.0	8.99
Cordulegastridae	78.9	15569.5	3.1	0.7	8.28

The censored–Poisson GAMs reveal detailed, nonlinear long-term patterns through the *year* smooth, with high effective degrees of freedom (EDF ~8.3–9.0) indicating multi-phase trends and family-specific turning points (Table 4.2; Fig. 4.2, cols. 1–2). *Aphelocheiridae* shows a sharp increase post-2010; *Odontoceridae* displays a slower, modest recovery; *Brachycentridae* exhibits a U-shaped trajectory dipping around 2005–2010; while *Cordulegastridae* shows steady rise from the early 2000s. These trajectories match the smoothed yearly means closely, underscoring model calibration. The large EDFs for `SITE_ID` (973–1774) versus much smaller values for `REPORTING_AREA` confirm strong site-level heterogeneity. Term-level Wald and ANOVA tests suggest seasonal fixed effects are weak for most families, and model diagnostics show residual normality and minimal autocorrelation (Appendix A.6). Notably, prediction error is markedly higher before 2000 (Fig. 4.2, col. 5), consistent with wide interval-censored categories dominating early records and limiting temporal precision.

The censored Poisson GAM approach offers a principled answer to one of the central issues in this project: the unification of decades of heterogenous abundance data gathered with changing recording protocols. While allowing for both true counts and interval-censored observations in a single likelihood (see Appendix A.6), the approach eschews ad-hoc transformations and ensures proper uncertainty propagation. Most importantly, the model exposes nuanced nonlinear temporal patterns through the smooths $s(\text{year})$, in turn picking up on gradual trends and instantaneous shifts (Figure 4.2). This is particularly useful in the case of pre-2000 category data, where early time points are dominated by such information and where the accuracy of subsequent predictions is constrained, but which are nevertheless rigorously included in order to extract all possible information. The use of random intercepts for `SITE_ID` and `REPORTING_AREA` allows for the effects of geographical clustering and regional variability, respectively; and the use of fixed seasonal effects further improves the estimation of the parameters of interest (Appendix A.6). Generally speaking, this approach provides a generalizable tool for the investigation of semi-quantitative biological monitoring data [1, 13].

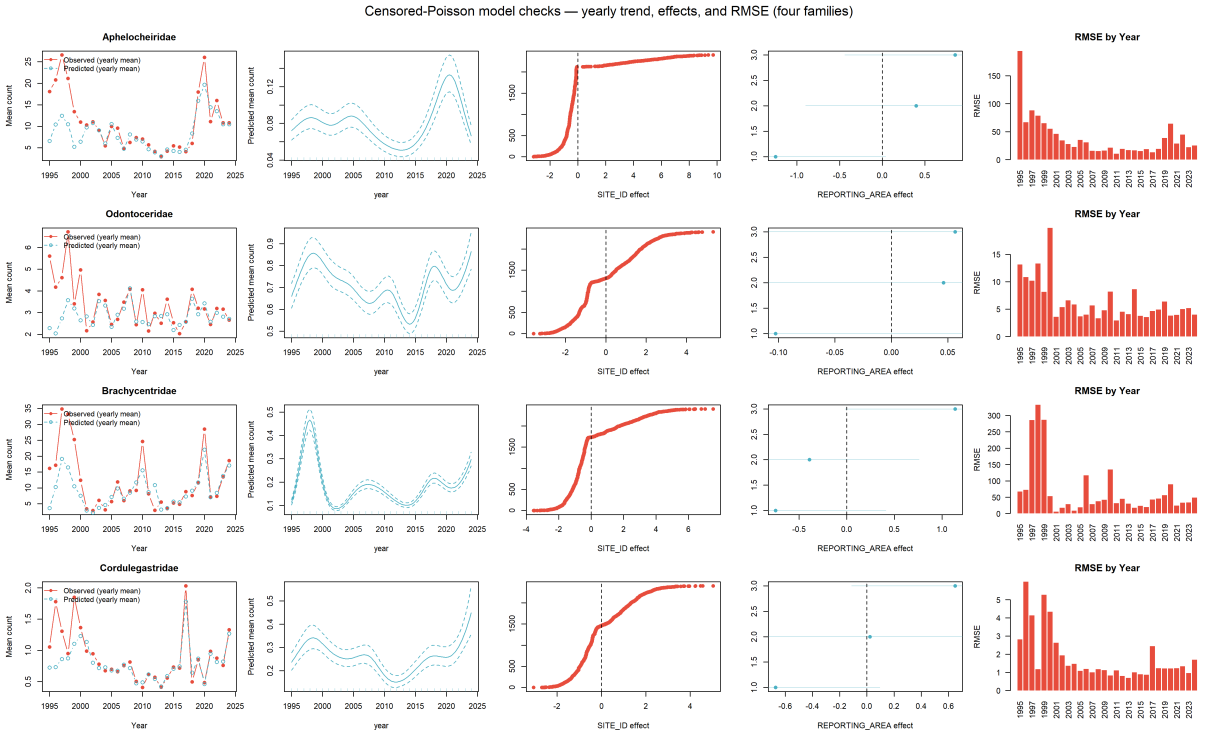

Figure 4.2: Censored–Poisson model checks—yearly trend, effects, and RMSE (four families). Each row is a family. Column 1: observed vs. predicted yearly mean counts. Column 2: partial effect of *year* on the mean-count scale (smooth $s(\text{year})$ with 95% bands). Column 3: SITE_ID random effects (caterpillar plot). Column 4: REPORTING_AREA random effects (dot ± 95% CI). Column 5: RMSE by year.

4.5 Numerical-Only Censored Poisson GAM

To determine if early categorical abundance reports improve model fit, we fit a censored Poisson GAM with numerical data only. The specification of the model–including temporal splines, space random effects, and seasonal fixed effects–follows the main fit (Appendix A.2, A.3) and is carried out with `mgcv`. Unlike the full model, however, this one has no interval-censored values based on placeholder codes (3, 33, 333, 3333) that Section 4.4 indicated as a possible source of early-year bias. By limiting to exact counts, we test the effect of eliminating uncertain historical data on fit and prediction and hence on the trade-off between data completeness and the precision of measurement.

Table 4.3: Comparison of fitted censored–Poisson GAMs (Full vs Numerical-only) by family. Each cell shows **Full / Numerical**. Metrics include deviance explained (%), AIC, RMSE, and MAE.

Family	Dev. expl. (%)	AIC	RMSE	MAE
Aphelocheiridae	86.10 / 87.60	64466.10 / 52957.10	86.70 / 27.99	8.48 / 5.05
Odontoceridae	71.00 / 71.20	38157.40 / 35684.00	11.48 / 5.14	2.65 / 1.92
Brachycentridae	73.00 / 73.40	131718.10 / 118898.70	121.16 / 68.41	11.00 / 9.00
Cordulegastridae	78.90 / 79.10	15569.50 / 14626.10	3.14 / 1.76	0.66 / 0.52

In order to assess the impact of adding categorical abundance information, we compared the full censored–Poisson GAM (with both numeric and categorically valued response) with a constrained model fit on numeric response only (Table 4.3, Table 4.4). In each of the four representative macoinvertebrate

Table 4.4: Comparison of smoothness and random effects in censored–Poisson GAMs (Full vs Numerical-only) by family. Each cell shows **Full / Numerical**. Metrics include the edf of $s(\text{year})$, SITE_ID, and REPORTING_AREA, as well as the overall p-value for the season fixed effect.

Family	edf(s(year))	edf(SITE_ID)	edf(REPORTING_AREA)	Season p-value
Aphelocheiridae	8.91 / 8.95	973.27 / 912.26	1.96 / 1.95	< 0.001 / < 0.001
Odontoceridae	8.76 / 8.85	1773.88 / 1675.64	1.29 / 1.35	≈ 0 / ≈ 0
Brachycentridae	8.99 / 9.00	1420.19 / 1323.28	1.98 / 1.97	< 0.001 / < 0.001
Cordulegastridae	8.28 / 8.35	1445.38 / 1328.78	1.97 / 1.96	3.8×10^{-5} / 1.7×10^{-5}

family-based analyses, the numeric-only model performed typically better in overall fit measures, for instance, with lesser AIC, RMSE, and MAE. But the difference was not drastic, and the full model still explained the vast majority of the deviance (e.g., 86.10% versus 87.60% for Aphelocheiridae). Furthermore, the effective degrees of freedom (edf) of the smooth term $s(\text{year})$ and of the random effects were highly similar across variants, and seasonal effects remained highly significant in both environments (Table 4.4). These results indicate that, despite numerical-only models having the advantage of more precise inputs, the inclusion of the categorical information in the full model does not interfere with the relationships fit and retains largely equal interpretability.

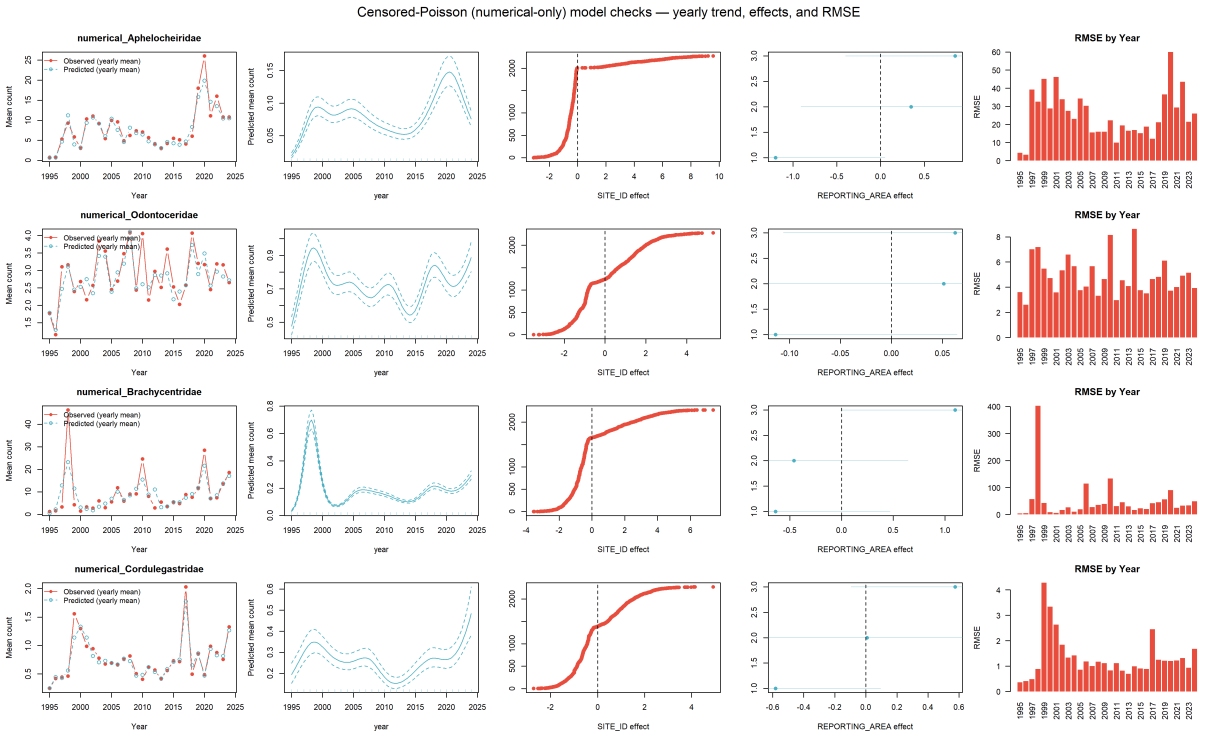

Figure 4.3: Censored–Poisson (numerical-only) model checks—yearly trend, effects, and RMSE (four families). Each row corresponds to one macroinvertebrate family. Column 1: observed vs. predicted yearly mean counts. Column 2: partial effect of *year* on the mean-count scale (smooth $s(\text{year})$ with 95% confidence bands). Column 3: SITE_ID random effects shown as caterpillar plots. Column 4: REPORTING_AREA random effects with point estimates and 95% confidence intervals. Column 5: root mean squared error (RMSE) by year.

While the numerical-only model performs better in aggregate statistics for some years, a closer examination of the annual RMSE trends paints a more complex picture. As indicated in Figure 4.4, since 2001—when numeric observations have become more reliably available—for a number of years through-

out families the full model performs equally well or even better than the numerical-only model. This pattern is particularly apparent for families like Aphelocheiridae and Cordulegastridae, where the RMSE across models aligns or varies without a clear predominance of one over the other. When considered in the light of diagnostic checks for each model (Figure 4.2, Figure 4.3) for each family, we see the full model captures the temporal trends and geographical heterogeneity (SITE_ID and REPORTING_AREA effects) well even when trained on numeric and categorical observations together. Considering the sparse and uncertain nature of the pre-2001 categorical observations, the full model's better performance since 2001 underscores the latter's strengths and real-world usefulness for characterizing the long-term abundance processes, particularly in the presence of numeric observations.

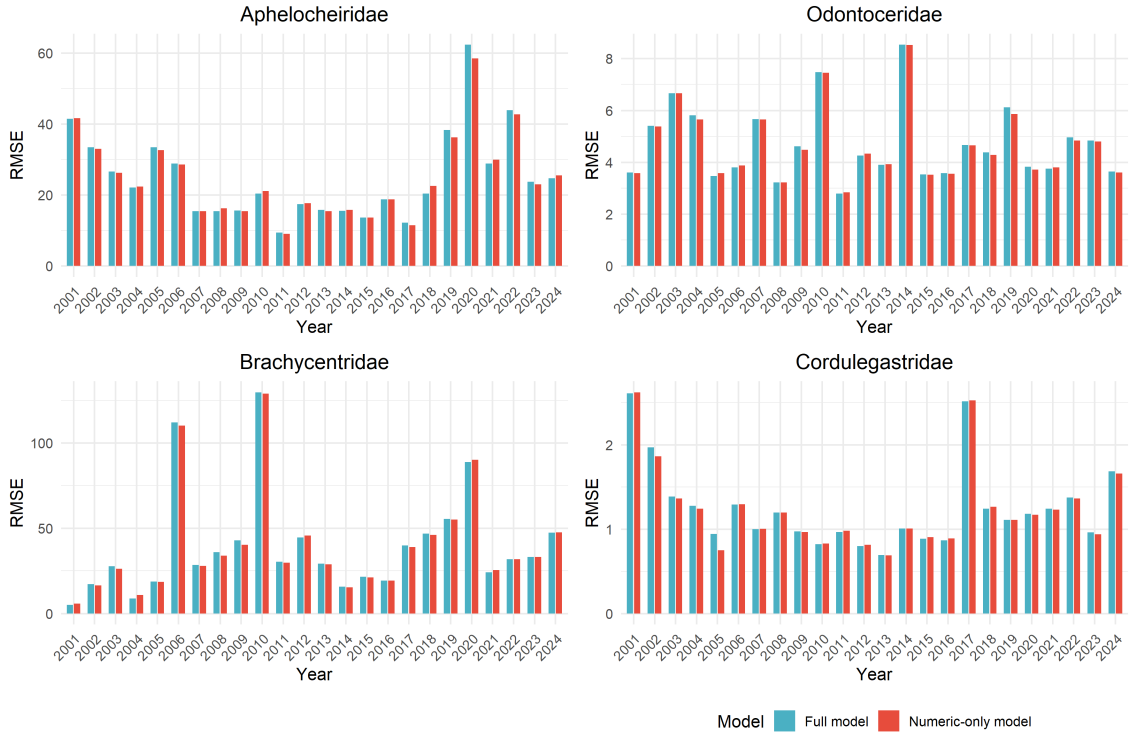

Figure 4.4: Yearly RMSE comparison from 2001 onward for four representative families (Aphelocheiridae, Odontoceridae, Brachycentridae, and Cordulegastridae). The bars compare the full model (which includes both categorical and numeric observations) with the numeric-only model. While the numeric-only model shows slightly better performance in some early years, the overall trend from 2001 onward indicates that the full model performs comparably and even better in several years. This suggests that, despite the greater uncertainty associated with categorical records before 2001, the full model remains robust and interpretable when focusing on the period with more reliable numeric data.

The smooth function $s(\text{year})$ in both models presents divergent long-term and recent reactions of various taxa to modifications in aquatic environments. With regards to the long-term pattern, the clean-water indicator families *Brachycentridae* and *Odontoceridae* have significantly diminished since the late 1990s and may have suffered from hydrological disruptions, habitat deterioration, or localized pollution demands. Both, however, exhibit signs of revival in the recent decade and indicate the new management actions may yet start showing their benefits. *Cordulegastridae* (a dragonfly family) exhibits the same pattern with a gentle initial decrease followed by recent stabilization or modest increase. *Aphelocheiridae*,

on the other hand, a low oxygen and organic pollution-tolerating group, has demonstrated a progressive increase since 2000 and may indicate the country-wide improvements of dissolved oxygen conditions. These results are clearly evident in Figures 4.2 and 4.3, highlighting inter-taxon variation in ecological sensibility and further underlining the necessity of moving on from binary presence/absence indicators toward abundance-based assessment regimes. With suitable modeling, even imprecise historical information can reveal the long-term ecological trends and form a sound basis for the appraisal and prediction of freshwater ecosystem condition.

5. Model External Validation

5.1 Logistic GAM Model Validation and Prediction

In order to assess the predictive ability of our binomial GAMs, we used them to predict a held-back test dataset for the year 2025 not used in the model fits. This helps us have an independent test of the ability of the models to generalize beyond the training dataset. For each of the four chosen macroinvertebrate families—*Odontoceridae*, *Brachycentridae*, *Aphelocheiridae*, and *Cordulegastridae*—we obtained the probabilities of occurrence under the respective pre-fitted models. We assessed the accuracy of the predictions in the test data by comparing the results with the true presence/absence results using the ROC curves and the respective AUC measures. We also calculated the log-loss for each family to express probabilistic prediction error [14, 13].

The test results, depicted in Figure 5.1, reveal good discriminative power for the majority of the taxa. *Brachycentridae* had the maximum AUC of 0.966 and minimum log-loss of 0.2112, indicating very good prediction accuracy. *Odontoceridae* closely followed with AUC = 0.907 and log-loss = 0.4223. *Aphelocheiridae*, being tolerant to environmental stress, had slightly lesser performance (AUC = 0.85; log-loss = 0.1174), and *Cordulegastridae* had the minimum AUC of 0.742 with maximum log-loss of 0.5626, which may perhaps result from the relatively few occurrences and larger site-specific variability of *Cordulegastridae*. These results support the model check patterns established in previous sections (Figure 4.1) and show the binomial GAM approach.

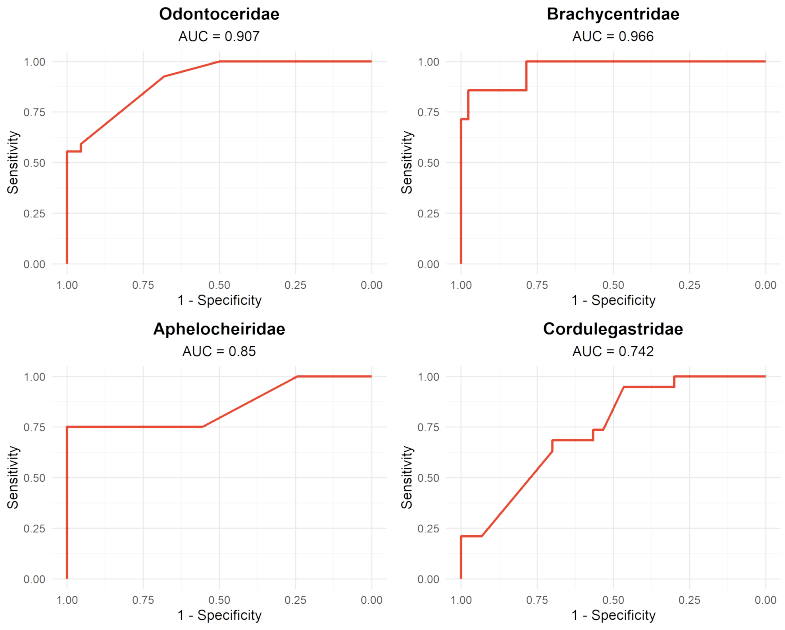

Figure 5.1: ROC curves for the binomial GAM predictions on the 2025 test dataset for four macroinvertebrate families. AUC values and log-loss metrics indicate strong predictive performance for most taxa.

To quantify future presence dynamics of some of the selected macroinvertebrate families, binomial GAMs were fit to both future and past sampling information. Presence probabilities for the future years 2001–2025 were directly predicted with the complete past dataset. The predictions for the future years 2026–2030 employed site-seasons sampled in 2024–2025 under realistic monitoring coverage. Each point's probabilities were thresholded by 0.5 (presence), 0.2 (lower), and 0.8 (upper) to estimate the mean annual occurrence rates and 95% confidence intervals [15]. Figure 5.2 results reveal diverging trajectories for the families. *Aphelocheiridae* reveals a slight downtrend after 2020, which may indicate environment sensitivity. *Odontoceridae* levels off to historic levels, and *Brachycentridae* reveals recent increase perhaps owing to the presence of eutrophic conditions. *Cordulegastridae* maintains the ascending trajectory, hinting of possible recovery.

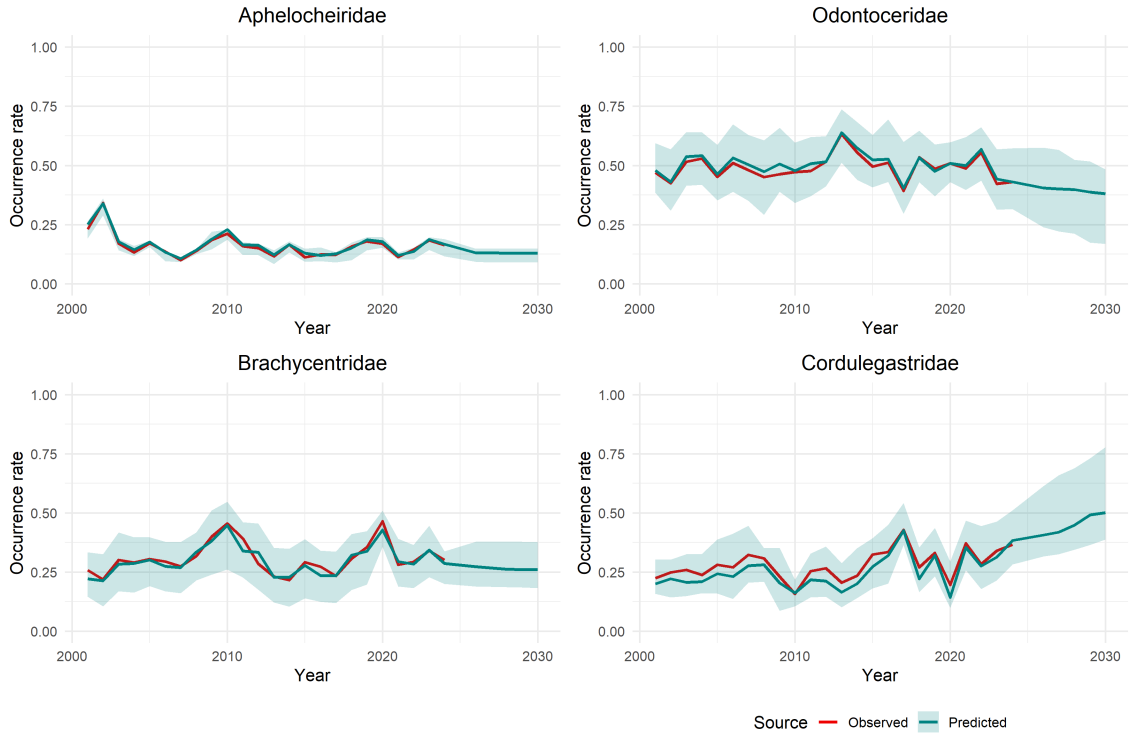

Figure 5.2: Predicted annual occurrence rates (with 95% intervals) for four macroinvertebrate families from 2001 to 2030. Red lines represent observed mean occurrence; blue lines represent thresholded predictions from binomial GAMs. Shaded ribbons indicate 95% prediction intervals based on presence probability thresholds of 0.2 and 0.8.

5.2 Censored Poisson GAM Validation

To further assess the extrapolative ability of the censored Poisson GAMs, we tested both the full model and the numeric-only model on an independent dataset collected in 2025. This dataset was not used during the model training phase and thus serves as a reliable holdout set for performance validation. For each macroinvertebrate family, we loaded the corresponding full model (which includes categorical predictors such as site and region) and numeric-only model (excluding such factors), and used them to predict the expected abundance (μ) for each site. We then compared the predicted values with the

observed abundance and computed the residuals. Four diagnostic plots were generated for each family: (1) predicted vs. observed abundance, (2) residuals vs. predicted values, (3) residual histograms, and (4) time series of predicted and observed abundance. These diagnostics collectively assess the accuracy and stability of each model in capturing ecological patterns across time and taxa [2, 16].

As shown in Figure 5.3, both the full model and the numeric-only model exhibited highly similar predictive behavior across all four macroinvertebrate families. The predicted values aligned closely with the observed abundances, and the residual distributions between the two models were nearly identical. Residuals were centered around zero and showed no systematic bias in either model, and the histograms revealed comparable spread. The time series plots further confirmed that both models captured the seasonal abundance patterns with similar temporal resolution and fluctuation intensity. These results suggest that removing categorical predictors such as site or region did not meaningfully degrade the predictive accuracy of the model.

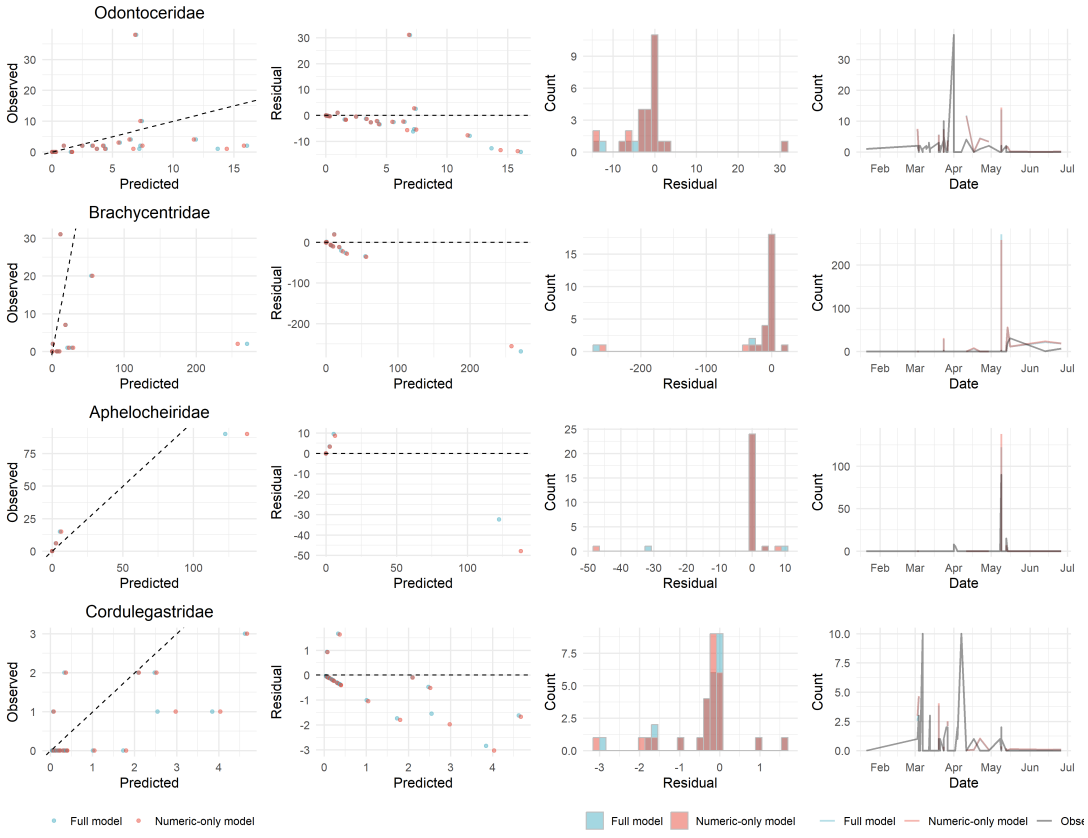

Figure 5.3: Model diagnostics for censored Poisson GAMs applied to 2025 test data. From left to right: predicted vs. observed abundance, residuals vs. predicted values, residual histograms, and predicted time series. Each row corresponds to a macroinvertebrate family. Blue represents the full model and red represents the numeric-only model.

Though the numeric-only model has similar predictive precision, the complete model is preferred for real-world use. With the addition of a larger group of predictors—including site- and region-level covariates—the complete model embodies spatial ecological heterogeneity without losing generalization [2, 17]. This is particularly useful for extrapolation, or where hierarchical structure embodies significant

ecological processes. As highlighted in recent ecological modeling frameworks, the incorporation of multiple data sources with retention of predictive robustness is essential for informing management and policy [18]. Thus, in spite of the parsimony of the numeric-only model, the complete model should serve as the benchmark, with the numeric-only model being suited for use in environmental monitoring and adaptive decision-making.

5.3 Forecasting Macroinvertebrate Abundance Using the Full Model

Based on the validation in Section 5.2, where the full Poisson GAM model had better or equal performance than the reduced model, the full model was used to predict macroinvertebrate abundance over time. During the historical period (2001–2025), the predictions were calculated directly with the use of observed values for all the sites. During the forecasting period (2026–2030), the predictions were created from sampling stations with documented observations for either the year 2024 or the year 2025. Four seasonal predictions each year per site and annual means were calculated to create the trajectories of abundance. Furthermore, 95% confidence intervals were calculated based on standard errors in the linear predictor of the model and back-transformed to the response variable scale. The trends of the predictions are depicted in the Figure 5.4.

Forecasts of abundance dynamics vary significantly for the macroinvertebrate families, as depicted in the figure. *Aphelocheiridae* falls off after a peak in the early 2020s, reflecting sensitivity to environmental decline. *Odontoceridae* levels off or rebounds, reflecting potential habitat enhancement. *Brachycentridae* surges strongly in increase, consonant with increased nutrient levels favoring collector-filterer groups. *Cordulegastridae* remains low but increases modestly. These forecasts are consistent with the smoothed temporal pattern of `s(year)` in Section 5.2 and validate the ecological appropriateness of the temporal structure of the model. Generally, the forecasts suggest a possible direction toward eutrophication or local water quality decline and highlight the importance of ongoing biological monitoring for early detection and management of ecological change.

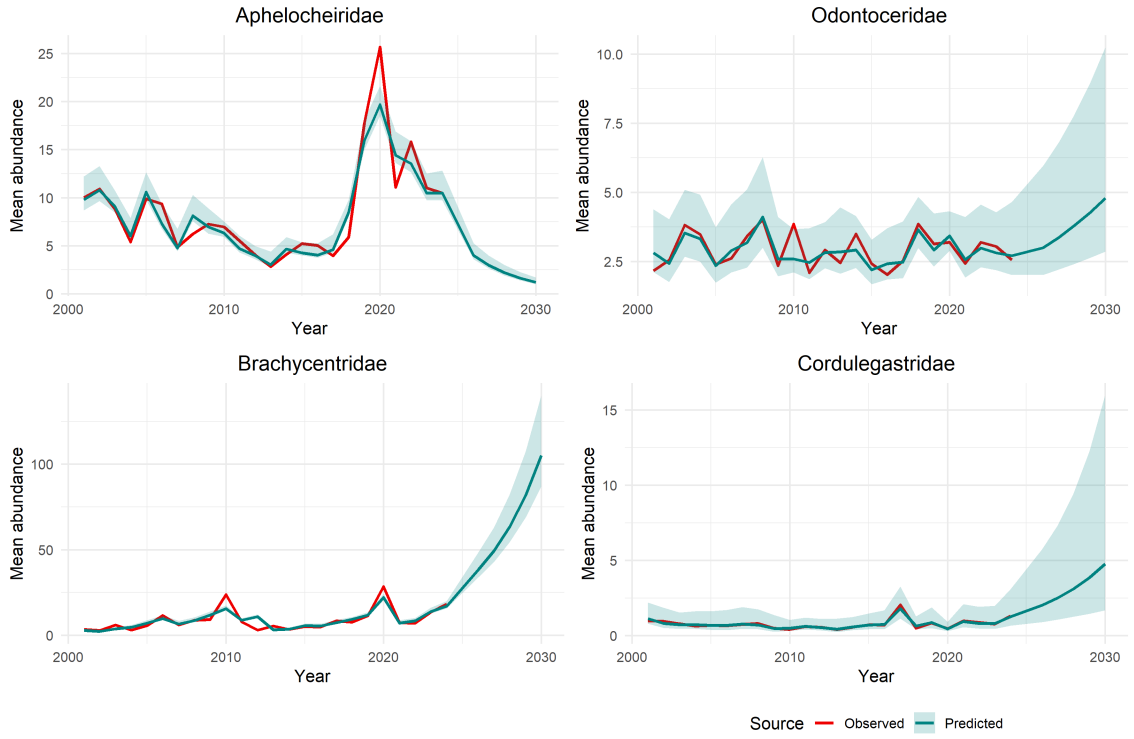

Figure 5.4: Forecasted mean abundance of four macroinvertebrate families from 2001 to 2030 using the full Poisson GAM model. Observed annual means (black lines) are shown for 2001–2025. Predictions from 2001–2025 use complete data; forecasts from 2026–2030 are based on sites sampled in 2024 or 2025. Shaded ribbons represent 95% confidence intervals.

6. Discussion

6.1 Conclusion

In this research, GAMs were used to describe and predict the distribution of macroinvertebrates with time, taking into account seasonal, geographical, and methodological heterogeneity. Validating censored Poisson and binomial models with historical observations ensured accurate and vigorous predictive performance for all taxa. The 2026–2030 predictions for occurrence probabilities thresholded to 0.5 (with uncertainty intervals of 0.1 and 0.9) indicate varied trends for the different families. While *Aphelocheiridae* is likely to decrease in the future, *Brachycentridae* indicates a rising occurrence, which may reflect the continuation of the process of eutrophication.

They confirm the ecological validity of the temporal smooths embodied by the models (see Section 5.2), and indicate that certain sensitive taxa could yet decline under the pressures of current environments. As a whole, the findings bolster the value of GAMs in the forecasting of ecology and the necessity of long-term monitoring of biological phenomena in informing freshwater management.

6.2 Limitation and Outlook

Though the fitted models uncover informative results on temporal trends of macroinvertebrate occurrence rates, the following weaknesses are to be recorded:

- **Zero inflation not modeled:** Zero-inflated binomial or negative binomial models would more adequately address excess zeros prevalent in ecological presence–absence data, particularly for sparse or irregularly distributed taxa.
- **Overdispersion not accounted for:** The binomial likelihood has constant variance; many ecologies have overdispersed ecological counts; a negative binomial distribution may offer a more flexible and robust fit.
- **Missing informative environmental covariates:** Informative covariates such as river flow velocity and river depth were not accounted for and could have otherwise enriched the explanatory power of the modeled relationships.
- **Spatial structure ignored:** Current models assume site-wide spatial independence. Geographic autocorrelation may be better accounted for with smoothers or geographic-based random effects based on geographic coordinates such as latitude and longitude.

To address poorer performing models and increased ecological realism in future research, the following avenues are possible:

- **Incorporate extra environmental information:** Utilization of habitat covariates such as velocity of flow, river depth, and location can improve the refinements on the predictions. Even with incomplete information, the use of imputation and partial pool methods can aid [19, 20].

- **Adopt Bayesian modeling:** Zero-inflated negative binomial (ZINB) models can be used to jointly model both the dropout probability (π) and count mean (μ), incorporating covariates. Temporal effects can be modeled using structured priors such as second-order random walk (RW2) for the zero-inflation component π, which assumes smooth and flexible trends over time, and first-order random walk (RW1) for the mean component μ, which allows for more localized temporal adaptation [21, 22, 23, 24]. See **Appendix A.7** for implementation details in `Stan`.

- **Handle spatial extrapolation:** To enable predictions at new locations, spatial Gaussian processes or conditional autoregressive (CAR) models [25, 26], as well as transfer learning approaches based on site similarity, may be used.

- **Model temporal dynamics more explicitly:** Predicting future year effects may be helped by time-series models, like autoregressive or state-space models [27, 28], that improve extrapolation performance beyond the observation window.

Although the current models have important temporal structures, the models are limited by the lack of adjustment for zero-inflation, overdispersion management, and large environmental or geographic covariates. Future improvement can include the application of ZINB models under Bayesian estimation, space structure, and time-series components for better forecasting accuracy. Enhancing coverage of the data and the complexity of the modeling will allow for stronger biodiversity forecasting for the long-term monitoring of freshwater ecosystems.

References

[1] E. Pharaoh, M. Diamond, S.J. Ormerod, G. Rutt, and I.P. Vaughan. Evidence of biological recovery from gross pollution in english and welsh rivers over three decades. *Science of the Total Environment*, 878:163107, 2023.

[2] Simon N. Wood. *Generalized Additive Models: An Introduction with R*. CRC Press, Boca Raton, 2nd edition, 2017.

[3] David Dudgeon et al. Freshwater biodiversity: importance, threats, status and conservation challenges. *Biological Reviews*, 81(2):163–182, 2006.

[4] J. Murray-Bligh, M. Griffiths, and M. Forshaw. *Freshwater Biology and Ecology Handbook*. Foundation for Water Research & Freshwater Biological Association, 2022.

[5] Y. Qu, V. Keller, N. Bachiller-Jareno, M. Eastman, F. Edwards, M.D. Jürgens, J.P. Sumpter, and A.C. Johnson. Significant improvement in freshwater invertebrate biodiversity in all types of english rivers over the past 30 years. *Science of the Total Environment*, 905:167144, 2023.

[6] M.A. Wilkes, M. Mungee, M. Naura, V.A. Bell, and L.E. Brown. Predicting nature recovery for river restoration planning and ecological assessment: A case study from england, 1991–2042. *River Research and Applications*, 2024.

[7] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[8] Trevor J. Hastie and Robert J. Tibshirani. *Generalized Additive Models*. Monographs on Statistics and Applied Probability. Chapman and Hall/CRC, London, 1990.

[9] S. N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36, 2011.

[10] Peter McCullagh and John A. Nelder. *Generalized Linear Models*, volume 37 of *Monographs on Statistics and Applied Probability*. Chapman and Hall/CRC, London, 2nd edition, 1989.

[11] Alan H Fielding and John F Bell. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, 24(1):38–49, 1997.

[12] Nicole H Augustin, Monica Musio, and Klaus von Wilpert. Model checking and goodness-of-fit. *Forest Biometry, Modelling and Information Sciences*, 1:45–66, 2009.

[13] Thomas W Yee. *Vector Generalized Linear and Additive Models: With an Implementation in R*. Springer, 2015.

[14] Thomas W Yee. The vgam package for categorical data analysis. *Journal of Statistical Software*, 23(10):1–36, 2008.

[15] Chunzhu Liu, Pam M Berry, Terence P Dawson, and Richard G Pearson. Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, 28(3):385–393, 2005.

[16] Martin Ridout, Clarice G B Demétrio, and John Hinde. Modeling censored count data with many zeros: A case study from microbiology. *Biometrical Journal*, 50(3):377–388, 2008.

[17] Liu Ye, Jun Li, Yuhui Cao, and Qiqi Liu. Modelling potential impacts of climate change on the spatial distribution of freshwater macroinvertebrate assemblages. *Ecological Modelling*, 221(6):933–940, 2010.

[18] John F. Carriger and Mace G. Barron. Toward the development of an ecological risk assessment framework for non-chemical stressors using ecosystem services. *Environmental Toxicology and Chemistry*, 38(8):1601–1614, 2019.

[19] Donald B Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, 1987.

[20] Stef van Buuren. *Flexible Imputation of Missing Data*. CRC press, 2018.

[21] Håvard Rue and Leonhard Held. Gaussian markov random fields: theory and applications. *Monographs on Statistics and Applied Probability*, 104, 2005.

[22] Stefan Lang and Andreas Brezger. Bayesian p-splines. *Journal of Computational and Graphical Statistics*, 13(1):183–212, 2004.

[23] H Rue, S Martino, and N Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B*, 71(2):319–392, 2009.

[24] Juho Piironen and Aki Vehtari. Bayesian predictive inference for general linear models with hierarchical shrinkage priors. *Journal of Statistical Computation and Simulation*, 90(4):703–727, 2020.

[25] Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. CRC press, 2 edition, 2014.

[26] Noel A C Cressie. *Statistics for Spatial Data*. John Wiley & Sons, 1993.

[27] James Durbin and Siem Jan Koopman. *Time Series Analysis by State Space Methods*. Oxford University Press, 2 edition, 2012.

[28] Jacques JF Commandeur and Siem Jan Koopman. *An Introduction to State Space Time Series Analysis*. Oxford University Press, 2007.

[29] Daniel Simpson, Håvard Rue, Andrea Riebler, Tiago G Martins, and Sigrunn H Sørbye. Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, 32(1):1–28, 2017.

[30] Julian Besag, Jeremy York, and Annie Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–20, 1991.

[31] Alain F Zuur, Elena N Ieno, Neil J Walker, Anatoly A Saveliev, and Graham M Smith. *Mixed Effects Models and Extensions in Ecology with R*. Springer, 2009.

Appendix

A Method

A.1 Categorical vs. Numerical Classification

We adopted a two-step strategy to classify abundance records into categorical and numerical types. In the first step, we applied a coarse rule that treats all values equal to 3, 33, 333, or 3333 as categorical and all other values as numerical. This classification is based on the EA's official monitoring protocol, which defines these four values as geometric means used for categorical abundance. However, because these same values may also appear in numerical datasets by chance or error, this rule alone is insufficient for accurately separating record types, especially during the 2000–2001 transition period when many sites gradually switched from categorical to numerical monitoring.

To address this, we introduced the concept of a "conversion point" for each monitoring site. If a site had no samples before January 1, 2000, we assumed it was numerical from the beginning, and all its samples were classified as numerical. For other sites, we scanned all provisionally numerical records dated before January 1, 2001 to find the earliest plausible conversion point. The key assumption is that a true conversion point must be followed by consistently numerical entries. Therefore, each candidate record is evaluated to determine whether it marks a genuine transition or is instead a misclassified numerical entry during a categorical period.

We formalized this as a hypothesis test. For each candidate numerical record dated before January 1, 2001, we defined a *check interval* as the period between the candidate's date and January 1, 2001, excluding endpoints. The null hypothesis (H_0) states that the candidate is the true conversion point, and all subsequent values—even if matching 3, 33, 333, or 3333—are genuine numerical entries. The alternative hypothesis (H_1) assumes that the candidate is a false positive. Under H_0, the probability of observing those four placeholder values as exact numerical results is given by:

$$P(3) = \frac{1}{10}, \quad P(33) = \frac{1}{100}, \quad P(333) = \frac{1}{1000}, \quad P(3333) = \frac{1}{10000}.$$

Let n_3, n_{33}, n_{333}, and n_{3333} be the number of times these values appear in the check interval. The joint probability of such a pattern occurring under H_0 is:

$$P = (1/10)^{n_3} \cdot (1/100)^{n_{33}} \cdot (1/1000)^{n_{333}} \cdot (1/10000)^{n_{3333}}.$$

If $P \leq 1/1000$, we reject H_0 and continue to the next candidate. If $P > 1/1000$, we accept H_0 and reclassify the candidate and all subsequent records at that site as numerical. If no valid candidate is

found before January 1, 2001, the first sample dated on or after that day is automatically considered the conversion point.

Once this classification protocol was completed for all sites, we additionally applied a hard override: any records dated 2001 or later were forcibly reclassified as numerical, regardless of their original values. This adjustment is motivated by the observation that categorical entries essentially vanish after 2001 (see Figure 2.2). After this reclassification, we summarized the final distribution of record types across four target macroinvertebrate families within the three selected EA regions. As shown in Figure 1, all post-2001 records are numerical, while pre-2001 data retain a mixture of categorical and numerical entries depending on site-specific transitions. This confirms that the conversion point algorithm effectively captured regional and taxonomic heterogeneity while ensuring consistency with EA's long-term protocol shift.

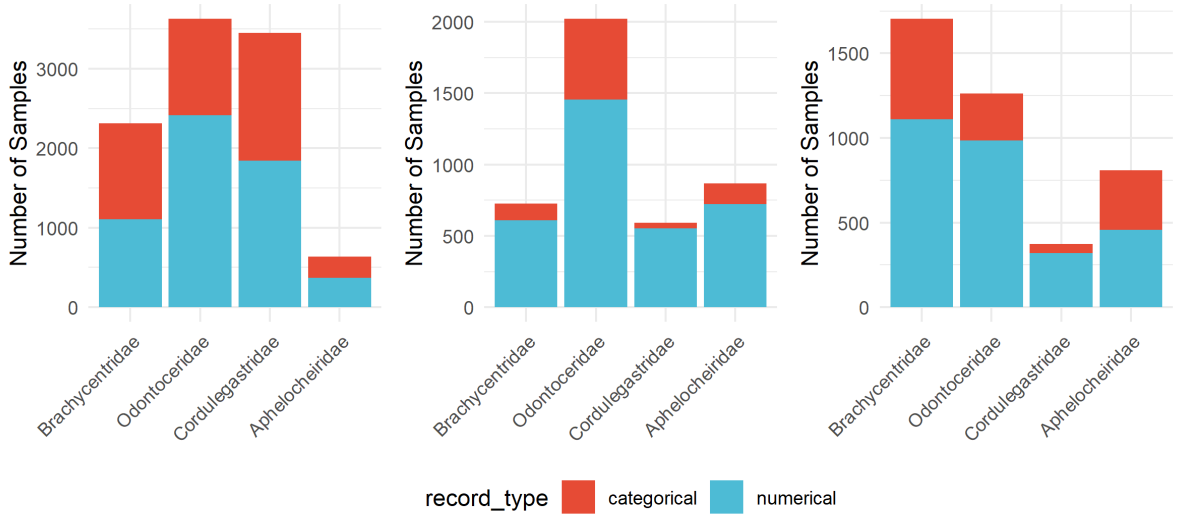

Figure 1: Number of categorical and numerical samples by family and EA region. Stacked bar plots show the number of samples classified as categorical (red) or numerical (blue) for each of the four families (*Brachycentridae*, *Odontoceridae*, *Cordulegastridae*, and *Aphelocheiridae*) across the three selected EA regions. These results reflect the final classification after applying site-specific conversion point inference and reclassifying all post-2001 samples as numerical.

A.2 Generalized additive models and the `mgcv` package

Generalized additive models (GAMs) extend generalized linear models (GLMs) by allowing smooth, potentially nonlinear functions of covariates while preserving an additive structure in the linear predictor [8, 2]. Formally, GAMs model the transformed mean response as

$$g(\mathbb{E}[Y_i]) = \beta_0 + \sum_{m=1}^{M} f_m(z_{im}), \tag{1}$$

where $g(\cdot)$ is a link function, β_0 is the intercept, z_{im} is the m-th covariate for observation i, and $f_m(\cdot)$ is a smooth function represented via basis expansion. Penalization of the smoothness of f_m controls

overfitting, with smoothing parameters typically estimated from the data using methods such as restricted maximum likelihood (REML).

The `mgcv` package [9, 2] provides a flexible and efficient framework for fitting GAMs in R. It offers a variety of smooth basis functions, including cubic regression splines and thin plate splines, and supports random effect smooths (`bs = "re"`), which enable the inclusion of hierarchical or grouped data structures within the GAM framework. For large datasets, the `bam` function implements computational optimizations such as discrete approximation, parallelization, and chunk-wise processing to improve efficiency without sacrificing accuracy.

In this study, GAMs are well suited for modeling long-term macroinvertebrate monitoring data. Temporal changes in ecological time series are rarely linear, and spline-based smooths allow flexible estimation of such trends. The hierarchical sampling design, involving multiple sites and reporting regions, can be accommodated through random effect smooths, capturing unobserved heterogeneity while borrowing information across groups. Fixed effects are used for factors such as season, which have a small number of predefined categories with direct ecological interpretation.

A.3 Binomial presence–absence modeling

To model species occurrence, we employed family-specific binomial generalized additive mixed models (GAMMs) with a complementary log–log (cloglog) link, which is well-suited for rare events. Let $Y_i \in \{0, 1\}$ denote the presence (1) or absence (0) of a given taxonomic family for observation i, and let $p_i = \Pr(Y_i = 1)$ be the corresponding occurrence probability. Under the binomial assumption, the likelihood for the i-th observation is

$$\Pr(Y_i \mid p_i) = p_i^{Y_i}(1 - p_i)^{1-Y_i}. \tag{2}$$

We link p_i to covariates through the cloglog transformation,

$$\log\big(-\log(1 - p_i)\big) = f(\text{year}_i) + b^{\text{site}}_{j(i)} + b^{\text{region}}_{k(i)} + \boldsymbol{x}_i^\top \boldsymbol{\beta}, \tag{3}$$

where:

$$f(\text{year}_i) : \text{penalized cubic regression spline capturing nonlinear temporal trends,} \tag{4}$$

$$b^{\text{site}}_j \sim \mathcal{N}(0, \sigma^2_{\text{site}}), \tag{5}$$

$$b^{\text{region}}_k \sim \mathcal{N}(0, \sigma^2_{\text{region}}), \tag{6}$$

and $\boldsymbol{x}_i$ encodes the fixed seasonal effects with a chosen reference category.

The smooth term $f(\text{year})$ captures long-term temporal patterns in occurrence probability, while the random intercepts b^{site} and b^{region} account for spatial heterogeneity at the site and reporting-area levels,

respectively. The seasonal fixed effects quantify phenological differences between sampling periods.

The complete likelihood for n observations is

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} p_i^{Y_i} (1 - p_i)^{1-Y_i}, \tag{7}$$

where $\boldsymbol{\theta}$ collects the spline coefficients defining $f(\cdot)$, the fixed effect coefficients $\boldsymbol{\beta}$, and the variance components σ^2_{site} and σ^2_{region}.

This formulation allows for flexible modeling of presence–absence data, capturing both nonlinear temporal patterns and hierarchical spatial structure. The cloglog link ensures that small occurrence probabilities are modeled on an appropriate scale for rare species, while the GAMM framework facilitates efficient estimation using penalized likelihood and smoothing parameter selection.

A.4 Binomial presence–absence GAM results and diagnostics

The fitted binomial GAMs for the four focal families achieved high predictive performance (AUC = 0.972–0.998; [11]) and explained 63.9–86.2% of the deviance (Table 2). Seasonal fixed effects (Table 1) were generally small relative to the intercept, with a clear negative summer effect for *Odontoceridae* ($p = 2.7 \times 10^{-4}$) and strong seasonality for *Brachycentridae* (Spring −, Summer +; both $p < 10^{-10}$). Seasonal effects were weak or non-significant for *Aphelocheiridae* and *Cordulegastridae*. Term-level tests for the season factor based on `mgcv::anova.gam` (approximate χ^2/F) were significant for *Brachycentridae* ($p < 2 \times 10^{-16}$) and *Odontoceridae* ($p \approx 0.001$), non-significant for *Aphelocheiridae* ($p \approx 0.18$), and modest for *Cordulegastridae* ($p \approx 0.004$) [2].

The smooth for *year* showed effective degrees of freedom (EDF) from $\sim$ 1.44 (*Aphelocheiridae*) to $\sim$ 7.27 (*Cordulegastridae*), indicating near-linear to moderately non-linear temporal structure. Very large EDFs for `SITE_ID` ($\sim$769–1465) confirmed strong site-level heterogeneity dominating over regional effects, consistent with standard GAM practice [2].

Model checks (Figure 2) indicated approximately normal residuals, little structure against fitted probabilities, minimal autocorrelation, and good calibration across the probability range [2, 12]. These diagnostics support the adequacy of the fitted models for inference on spatial–temporal occurrence patterns.

Table 1: Fixed effect estimates from logistic GAMs for four focal families. The rightmost column reports the per-family *term-level* p for the seasonal factor from `mgcv::anova.gam` (approximate χ^2/F; not a true LRT). For brevity, this term-level p is shown once per family (first row) and applies to all season coefficients in that family.

Family	Term	Estimate	SE	Statistic	Wald p-value	p (season)
Aphelocheiridae	(Intercept)	-2.880	0.599	-4.805	1.5443e-06	0.1834
	seasonSpring	0.030	0.071	0.424	0.6716	
	seasonSummer	0.245	0.130	1.883	0.0597	
	seasonWinter	0.332	0.234	1.418	0.1561	
Odontoceridae	(Intercept)	-0.614	0.125	-4.896	9.7987e-07	0.0011
	seasonSpring	-0.190	0.042	-4.499	6.8301e-06	
	seasonSummer	-0.363	0.090	-4.022	5.7623e-05	
	seasonWinter	0.131	0.282	0.467	0.6403	
Brachycentridae	(Intercept)	-1.690	0.610	-2.772	0.0056	0.0000
	seasonSpring	0.485	0.074	6.550	5.6602e-11	
	seasonSummer	1.432	0.122	11.767	$< 2 \times 10^{-16}$	
	seasonWinter	0.849	0.225	3.776	1.5927e-04	
Cordulegastridae	(Intercept)	-1.718	0.377	-4.558	5.1529e-06	0.0043
	seasonSpring	-0.073	0.061	-1.202	0.2293	

Table 2: Smooth terms, model fit statistics, and predictive performance for binomial GAMs. EDF = estimated degrees of freedom.

Family	Smooth term	EDF	Ref.df	F	p-value	Dev %	AUC
Aphelocheiridae	s(REPORTING_AREA)	1.965	2.000	13940.334	$< 2 \times 10^{-16}$	86.2	0.998
	s(SITE_ID)	769.409	2400.000	25664.401	$< 2 \times 10^{-16}$	86.2	0.998
	s(year)	1.439	1.738	8.221	0.0310	86.2	0.998
Odontoceridae	s(REPORTING_AREA)	1.742	2.000	908.818	9.3×10^{-5}	69.2	0.982
	s(SITE_ID)	1464.656	2400.000	8044.968	$< 2 \times 10^{-16}$	69.2	0.982
	s(year)	5.351	6.325	32.341	1.5×10^{-5}	69.2	0.982
Brachycentridae	s(REPORTING_AREA)	1.974	2.000	13239.744	$< 2 \times 10^{-16}$	63.9	0.972
	s(SITE_ID)	1114.842	2400.000	10452.775	$< 2 \times 10^{-16}$	63.9	0.972
	s(year)	6.039	7.079	40.272	$< 2 \times 10^{-16}$	63.9	0.972
Cordulegastridae	s(REPORTING_AREA)	1.967	2.000	2881.347	$< 2 \times 10^{-16}$	66.8	0.978
	s(SITE_ID)	1347.461	2400.000	7855.020	$< 2 \times 10^{-16}$	66.8	0.978
	s(year)	7.269	8.123	60.776	$< 2 \times 10^{-16}$	66.8	0.978

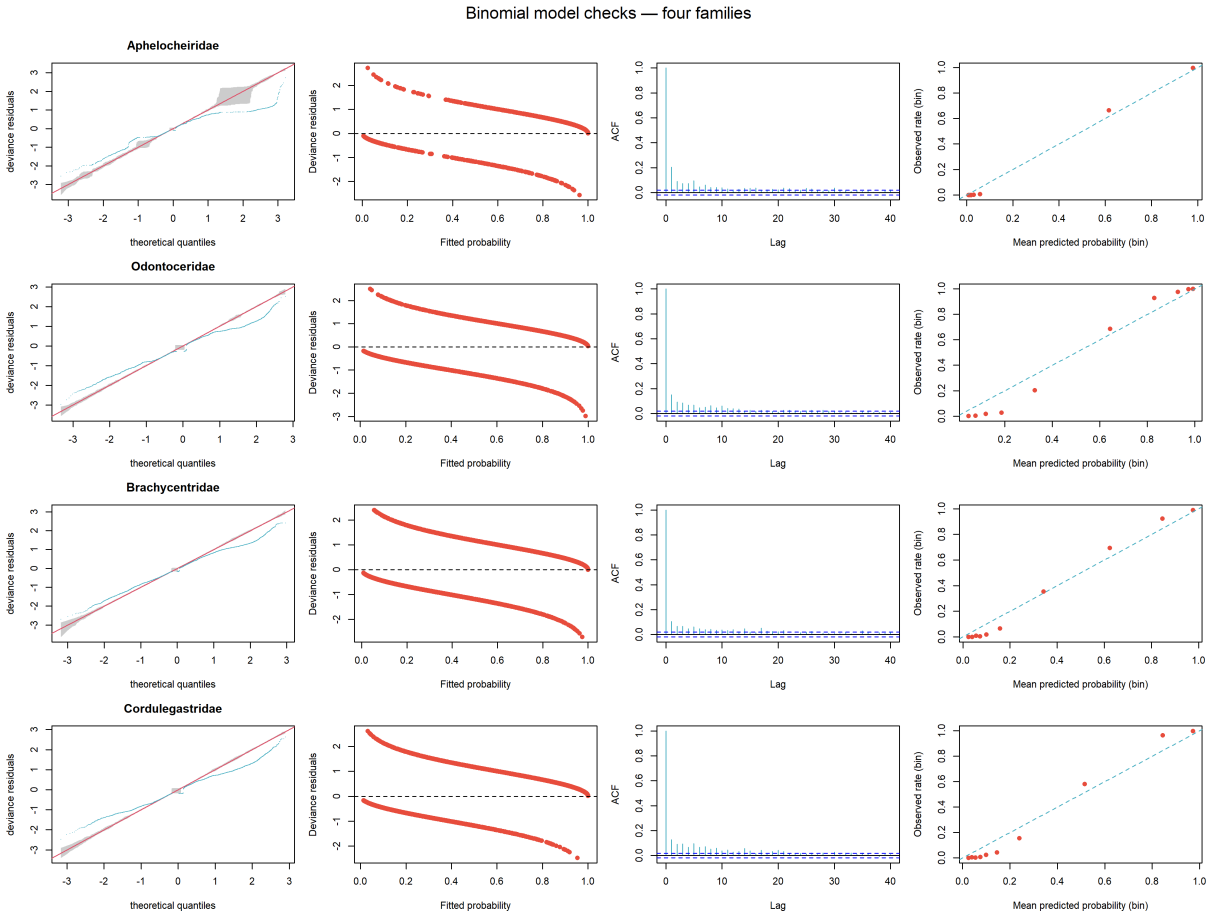

Figure 2: Model diagnostics for four focal families. From left to right: Q–Q plots of deviance residuals; residuals vs. fitted probability; autocorrelation function (ACF) of residuals; calibration plots comparing observed occurrence rate to mean predicted probability. All four families show approximate residual normality, minimal residual autocorrelation, and good calibration across the probability range.

A.5 Censored Poisson modeling

Censored count models extend standard count distributions to cases where the observed value is only known to fall within an interval. Let N_i denote the unobserved true abundance for observation i. In the case of exact counts, the observation equals N_i; for interval-censored data, the record provides bounds $[L_i, U_i]$ such that $L_i \le N_i \le U_i$. Under a Poisson distribution with mean μ_i, the probability of observing the i-th datum is

$$\Pr(L_i \le N_i \le U_i \mid \mu_i) = \sum_{n=L_i}^{U_i} \frac{e^{-\mu_i} \mu_i^n}{n!}. \tag{8}$$

When $L_i = U_i$, this reduces to the standard Poisson probability mass function.

In the present study, the mean μ_i is linked to covariates through a generalized additive mixed model (GAMM) structure,

$$\log \mu_i = f(\text{year}_i) + b^{\text{site}}_{j(i)} + b^{\text{region}}_{k(i)} + \boldsymbol{x}_i^\top \boldsymbol{\beta}, \tag{9}$$

where:

$$f(\text{year}_i) : \text{penalized cubic regression spline}, \tag{10}$$

$$b_j^{\text{site}} \sim \mathcal{N}(0, \sigma^2_{\text{site}}), \tag{11}$$

$$b_k^{\text{region}} \sim \mathcal{N}(0, \sigma^2_{\text{region}}), \tag{12}$$

and $\boldsymbol{x}_i$ encodes the seasonal factor levels with a reference category. The penalized spline term captures smooth temporal trends. The random intercepts for site and reporting area model hierarchical spatial variation, where SITE_ID represents a finer spatial unit nested within REPORTING_AREA.

The full likelihood for all observations is then

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{n=L_i}^{U_i} \frac{e^{-\mu_i} \mu_i^n}{n!}, \tag{13}$$

where $\boldsymbol{\theta}$ collects the fixed effect coefficients $\boldsymbol{\beta}$, smooth coefficients defining $f(\cdot)$, variance components σ^2_{site} and σ^2_{region}, and smoothing parameters.

This formulation allows exact counts and interval-censored observations to be analyzed jointly under a coherent likelihood, avoiding information loss from midpoint substitution and propagating measurement or counting uncertainty into parameter estimates. The GAMM structure accommodates nonlinear temporal patterns and hierarchical spatial design, capturing variation at both regional and site levels. However, the censored Poisson assumption implies equidispersion (variance equal to the mean) for the latent counts, which may be violated in ecological data exhibiting over- or under-dispersion. In such cases, extensions to more flexible count families (e.g., negative binomial, Conway–Maxwell–Poisson) may be required, and computation can become more intensive when intervals are wide or sample size is large.

A.6 Censored Poisson GAM results and diagnostics

The fitted censored Poisson GAMs for the four focal families explained 71.0–86.1% of the deviance (Table 4). Seasonal fixed effects (Table 3) were generally small relative to the intercept, with strong seasonality for *Brachycentridae* (Spring −, Summer +; both $p < 10^{-15}$) and *Aphelocheiridae* (Spring −, Summer +; both $p < 10^{-4}$), while *Odontoceridae* and *Cordulegastridae* showed weaker or inconsistent seasonal signals. Overall tests for the season factor based on `mgcv::anova.gam` (approximate χ^2/F; not a true LRT) were highly significant for *Brachycentridae* and *Aphelocheiridae* ($p < 10^{-15}$), but non-significant for *Odontoceridae* and *Cordulegastridae* [2].

The smooth for *year* showed high effective degrees of freedom (EDF), from ∼ 8.28 (*Cordulegastridae*) to ∼ 8.99 (*Brachycentridae*), indicating strongly non-linear temporal variation. Very large EDFs for `SITE_ID` (973–1774) confirmed substantial site-level heterogeneity that dominated over regional effects

(REPORTING_AREA) [2].

Model checks (Figure 3) indicated approximately normal residuals, minimal residual structure, and no strong temporal autocorrelation, supporting the adequacy of the fitted models for inference on spatial–temporal count patterns [12, 2].

Table 3: Fixed effect estimates from censored Poisson GAMs for four focal families. The rightmost column reports the per-family *term-level p* for the seasonal factor from `mgcv::anova.gam` (approximate χ^2/F; not a true LRT). For brevity, this term-level p is shown once per family (first row) and applies to all season coefficients in that family.

Family	Term	Estimate	SE	Statistic	Wald p-value	p (season)
Aphelocheiridae	(Intercept)	-2.554	0.654	-3.907	9.39e-05	0.00e+00
	seasonSpring	-0.425	0.008	-50.052	0.00e+00	
	seasonSummer	-0.381	0.014	-27.608	2.07e-161	
	seasonWinter	-0.661	0.036	-18.176	1.39e-72	
Odontoceridae	(Intercept)	-0.348	0.080	-4.363	1.30e-05	1.37e-58
	seasonSpring	-0.148	0.014	-10.881	2.15e-27	
	seasonSummer	-0.377	0.031	-12.313	1.50e-34	
	seasonWinter	0.244	0.084	2.920	3.51e-03	
Brachycentridae	(Intercept)	-1.763	0.585	-3.012	2.60e-03	0.00e+00
	seasonSpring	0.321	0.009	35.092	6.31e-253	
	seasonSummer	1.440	0.011	128.149	0.00e+00	
	seasonWinter	0.852	0.040	21.091	2.03e-96	
Cordulegastridae	(Intercept)	-1.373	0.388	-3.541	4.01e-04	2.12e-06
	seasonSpring	-0.096	0.025	-3.857	1.15e-04	
	seasonSummer	0.100	0.051	1.955	5.06e-02	
	seasonWinter	0.122	0.225	0.541	5.88e-01	

Table 4: Smooth terms, model fit statistics for censored Poisson GAMs. EDF = estimated degrees of freedom.

Family	Smooth term	EDF	Ref.df	χ^2/F	p-value	Dev %
Aphelocheiridae	s(REPORTING_AREA)	1.957	2.0	146041488.084	$< 2 \times 10^{-16}$	86.1
	s(SITE_ID)	973.272	2400.0	217.007	$< 2 \times 10^{-16}$	
	s(year)	8.910	8.997	330.373	$< 2 \times 10^{-16}$	
Odontoceridae	s(REPORTING_AREA)	1.287	2.0	12579.524	0.0922	71.0
	s(SITE_ID)	1773.876	2400.0	34.450	$< 2 \times 10^{-16}$	
	s(year)	8.764	8.980	30.603	$< 2 \times 10^{-16}$	
Brachycentridae	s(REPORTING_AREA)	1.975	2.0	39585964.297	$< 2 \times 10^{-16}$	73.0
	s(SITE_ID)	1420.195	2400.0	71.930	$< 2 \times 10^{-16}$	
	s(year)	8.993	9.000	1112.460	$< 2 \times 10^{-16}$	
Cordulegastridae	s(REPORTING_AREA)	1.969	2.0	25635.153	$< 2 \times 10^{-16}$	78.9
	s(SITE_ID)	1445.376	2400.0	16.182	$< 2 \times 10^{-16}$	
	s(year)	8.284	8.821	22.650	$< 2 \times 10^{-16}$	

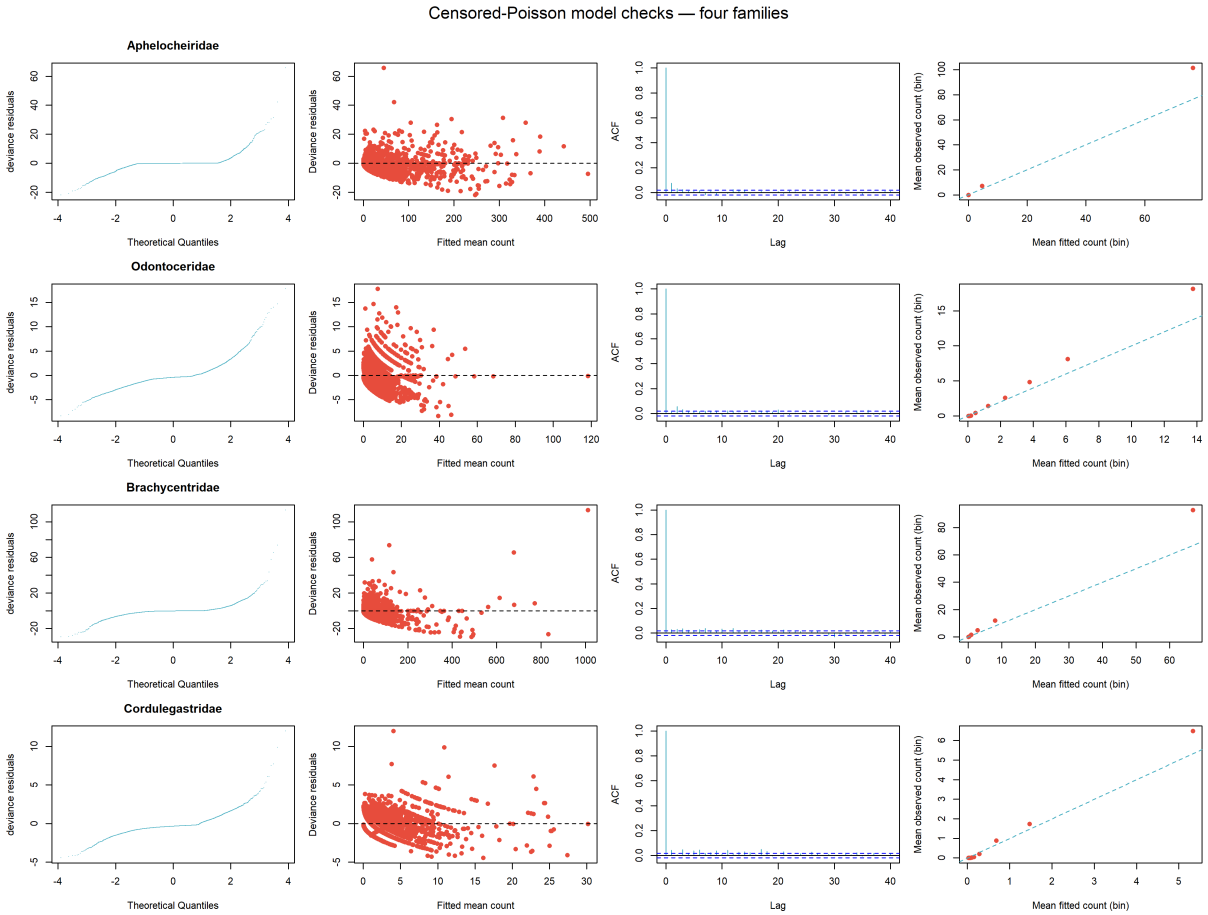

Figure 3: Censored-Poisson model diagnostics for four focal families. From left to right: Q–Q plots of deviance residuals; residuals vs. fitted mean count; autocorrelation function (ACF) of residuals; calibration plots comparing observed to binned mean fitted counts. Residuals are approximately normal with minimal autocorrelation and good calibration across fitted-count ranges.

A.7 Extension: Zero-inflated Negative Binomial Bayesian Random Walk Model

To address zero inflation and temporal smoothness in ecological presence–absence data, we propose a Bayesian ZINB model. This approach models the zero-inflation component (π) using a second-order random walk (RW2) prior and the count mean (μ) using a first-order random walk (RW1) prior, capturing both long-term structural dropout and short-term abundance fluctuations. This hierarchical framework improves predictive performance and uncertainty quantification, especially for rare taxa and future extrapolations [23, 29, 30, 31].

The model assumes that the true latent count Y_i follows a ZINB distribution, and the observations are either exact or interval-censored, denoted by bounds $[L_i, U_i]$ for each observation $i = 1, \ldots, N$. When $L_i = U_i$, the observation is exact.

Likelihood. The likelihood of the data under the ZINB distribution is given by:

$$Y_i \sim \begin{cases} 0 & \text{with probability } \pi_i + (1-\pi_i) \cdot \mathrm{NB}(0 \mid \mu_i, \phi) \\ y \sim \mathrm{NB}(\mu_i, \phi) & \text{with probability } 1-\pi_i \end{cases} \tag{14}$$

For censored observations, the likelihood contribution is:

$$\mathbb{P}(Y_i \in [L_i, U_i]) = \begin{cases} \pi_i + (1-\pi_i) \cdot F_{\text{NB}}(U_i \mid \mu_i, \phi) & \text{if } L_i = 0 \\ (1-\pi_i) \cdot [F_{\text{NB}}(U_i) - F_{\text{NB}}(L_i - 1)] & \text{if } L_i > 0 \end{cases} \tag{15}$$

Linear Predictors. The model uses separate linear predictors for the count mean μ_i and the zero-inflation probability π_i, defined as:

$$\log \mu_i = \alpha_\mu + \theta_{t[i]} + a_{j[i]} + d_{k[i]} + \beta_{\text{season}[i]} \tag{16}$$

$$\text{logit}(\pi_i) = \alpha_\pi + \zeta_{t[i]} + u_{j[i]} + v_{k[i]} \tag{17}$$

where:

- α_μ and α_π are intercept terms;
- $\theta_t \sim$ RW2 is a second-order random walk over time t for the log-mean component;
- $\zeta_t \sim$ RW1 is a first-order random walk over time t for the logit-zero-inflation component;
- $a_j \sim \mathcal{N}(0, \sigma^2_{\mu,\text{area}})$, $d_k \sim \mathcal{N}(0, \sigma^2_{\mu,\text{site}})$ are area and site random effects for μ;
- $u_j \sim \mathcal{N}(0, \sigma^2_{\pi,\text{area}})$, $v_k \sim \mathcal{N}(0, \sigma^2_{\pi,\text{site}})$ are area and site random effects for π;
- β_{season} is a seasonal fixed effect with sum-to-zero constraint;
- $t[i], j[i], k[i]$ index the year, area, and site for observation i.

Random Walk Priors. The temporal effects follow the following difference priors:

$$\theta_t - 2\theta_{t-1} + \theta_{t-2} \sim \mathcal{N}(0, \sigma^2_{\mu,\text{rw}}) \quad \text{for } t = 3, \ldots, T \tag{18}$$

$$\zeta_t - \zeta_{t-1} \sim \mathcal{N}(0, \sigma^2_{\pi,\text{rw}}) \quad \text{for } t = 2, \ldots, T \tag{19}$$

Priors. The prior distributions are set as follows:

$$\alpha_\mu \sim \mathcal{N}(0, 5), \quad \alpha_\pi \sim \mathcal{N}(0, 2) \tag{20}$$

$$\beta_s \sim \mathcal{N}(0, 1), \quad \sum_{s=1}^{S} \beta_s = 0 \tag{21}$$

$$\log \phi \sim \mathcal{N}(0, 1) \tag{22}$$

where ϕ is the dispersion parameter of the negative binomial distribution. All standard deviations (for random effects and random walks) are assigned weakly informative half-Student-$t(3, 0, 1)$ or half-normal priors, depending on the component.

Model Summary. This model is well-suited for ecological monitoring data with excess zeros, overdispersion, spatial structure (area and site), seasonal variation, and smooth temporal trends. It can handle both exact and interval-censored observations using a flexible zero-inflated negative binomial likelihood.

Stan Codes for ZINB-BRWM

```
data {
  int<lower=1> N;                          // number of observations
  int<lower=0> L[N];                       // lower bound (when exact: L == U)
  int<lower=0> U[N];                       // upper bound
  int<lower=1> T;                          // number of distinct years
  int<lower=1, upper=T> year_idx[N];       // year index (1..T) for each observation

  int<lower=1> J_area;                     // number of areas
  int<lower=1> K_site;                     // number of sites
  int<lower=1, upper=J_area> area_id[N];   // area index for each observation
  int<lower=1, upper=K_site> site_id[N];   // site index for each observation

  int<lower=1> S_season;                   // number of season levels
  int<lower=1, upper=S_season> season_id[N]; // season index for each observation

  // Optional: centered time index used to remove intercept/slope in RW projections
  vector[T] t_centered;                    // centered (mean-zero) time covariate
}

parameters {
  // Intercepts
  real alpha_mu;
  real alpha_pi;

  // Year effects (raw): RW2 for mu, RW1 for pi (to be projected/centered later)
  vector[T] theta_mu_raw;                  // for mu (RW2)
  vector[T] zeta_pi_raw;                   // for pi (RW1)
```

```
28
29    // Season fixed effects (raw, will be sum-to-zero centered later)
30    vector[S_season] beta_season_raw;
31
32    // i.i.d. random effects (non-centered parameterization)
33    vector[J_area] area_mu_raw;
34    vector[K_site] site_mu_raw;
35    vector[J_area] area_pi_raw;
36    vector[K_site] site_pi_raw;
37
38    // Hyper-parameters (standard deviations / scales)
39    real<lower=0> sigma_mu_rw;                // RW2 curvature scale for mu
40    real<lower=0> sigma_pi_rw;                // RW1 increment scale for pi
41    real<lower=0> sigma_mu_area;
42    real<lower=0> sigma_mu_site;
43    real<lower=0> sigma_pi_area;
44    real<lower=0> sigma_pi_site;
45
46    // Negative binomial shape parameter; use log scale to enforce positivity
47    real log_phi;
48  }
49
50  transformed parameters {
51    // ----- Construct transformed structures -----
52
53    // (1) Year effect for mu (theta_mu): remove mean and linear trend for RW2 identifiability
54    vector[T] theta_mu_tilde = theta_mu_raw;
55    real a_mu = mean(theta_mu_tilde);                                  // intercept component
56    real b_mu = dot_product(theta_mu_tilde, t_centered)
57                / dot_product(t_centered, t_centered);                 // slope component
58    vector[T] theta_mu = theta_mu_tilde - a_mu - b_mu * t_centered;
59
60    // (2) Year effect for pi (zeta_pi): mean-center only for RW1 identifiability
61    vector[T] zeta_pi = zeta_pi_raw - mean(zeta_pi_raw);
62
63    // (3) Season fixed effects: impose sum-to-zero for interpretability
```

```
  vector[S_season] beta_season = beta_season_raw - mean(beta_season_raw);

  // (4) Non-centered random effects (i.i.d.)
  vector[J_area] area_mu = sigma_mu_area * area_mu_raw;
  vector[K_site] site_mu = sigma_mu_site * site_mu_raw;
  vector[J_area] area_pi = sigma_pi_area * area_pi_raw;
  vector[K_site] site_pi = sigma_pi_site * site_pi_raw;
}

model {
  // ===== Priors =====
  // Intercepts
  alpha_mu ~ normal(0, 5);
  alpha_pi ~ normal(0, 2);

  // Season (only in the mu component)
  beta_season_raw ~ normal(0, 1);

  // Bases for non-centered random effects
  area_mu_raw ~ normal(0, 1);
  site_mu_raw ~ normal(0, 1);
  area_pi_raw ~ normal(0, 1);
  site_pi_raw ~ normal(0, 1);

  // Hyper-parameters (weakly-informative shrinkage)
  sigma_mu_area ~ student_t(3, 0, 1);     // half-t(3, 1)
  sigma_mu_site ~ student_t(3, 0, 1);
  sigma_pi_area ~ student_t(3, 0, 1);
  sigma_pi_site ~ student_t(3, 0, 1);

  sigma_mu_rw ~ student_t(3, 0, 0.5);     // RW2 curvature prior (tune 0.1{0.5)
  sigma_pi_rw ~ normal(0, 0.3);           // half-normal(0, 0.3) via <lower=0> constraint

  // Negative binomial shape (log scale)
  log_phi ~ normal(0, 1);
  real phi = exp(log_phi);
```

```
100
101   // ===== Random-walk priors (difference form) =====
102   // RW2 for theta_mu
103   for (t in 3:T) {
104     real dd = theta_mu[t] - 2 * theta_mu[t - 1] + theta_mu[t - 2];
105     dd ~ normal(0, sigma_mu_rw);
106   }
107   // Optional weak priors at boundaries to stabilize endpoints
108   theta_mu[1] ~ normal(0, 10);
109   theta_mu[2] ~ normal(0, 10);
110
111   // RW1 for zeta_pi
112   for (t in 2:T) {
113     real d = zeta_pi[t] - zeta_pi[t - 1];
114     d ~ normal(0, sigma_pi_rw);
115   }
116   // Weak prior at the first endpoint
117   zeta_pi[1] ~ normal(0, 5);
118
119   // ===== Likelihood (ZINB with interval censoring) =====
120   for (i in 1:N) {
121     // Linear predictors
122     real eta_mu = alpha_mu
123                   + theta_mu[year_idx[i]]
124                   + area_mu[area_id[i]]
125                   + site_mu[site_id[i]]
126                   + beta_season[season_id[i]];
127     real mu_i = exp(eta_mu);
128
129     real eta_pi = alpha_pi
130                   + zeta_pi[year_idx[i]]
131                   + area_pi[area_id[i]]
132                   + site_pi[site_id[i]];
133     real pi_i = inv_logit(eta_pi);
134
135     if (L[i] == U[i]) {
```

```
      // Exact observation
      if (L[i] == 0) {
        // Exact zero: log( pi + (1 - pi) * P_NB(0) )
        target += log_sum_exp( log(pi_i),
                               log1m(pi_i) + neg_binomial_2_lpmf(0 | mu_i, phi) );
      } else {
        // Exact positive count: log( (1 - pi) * P_NB(y) )
        target += log1m(pi_i) + neg_binomial_2_lpmf(U[i] | mu_i, phi);
      }
    } else {
      // Interval-censored: [L, U], inclusive
      if (L[i] == 0) {
        // Interval includes zero: Pr = pi + (1 - pi) * F_NB(U)
        target += log_sum_exp( log(pi_i),
                               log1m(pi_i) + neg_binomial_2_lcdf(U[i] | mu_i, phi) );
      } else {
        // Positive-only interval: Pr = (1 - pi) * (F(U) - F(L-1))
        target += log1m(pi_i)
                  + log_diff_exp( neg_binomial_2_lcdf(U[i] | mu_i, phi),
                                  neg_binomial_2_lcdf(L[i] - 1 | mu_i, phi) );
      }
    }
  }
}

```

Although the primary objective of this study is to model macroinvertebrate abundance using a censored Poisson GAM, we have additionally constructed a Bayesian zero-inflated negative binomial (ZINB) model with temporal random walk priors in `Stan`. This model addresses several key limitations of the GAM framework: it explicitly models the zero-inflation probability (π) and the count mean (μ) as functions of covariates, uses a negative binomial distribution to account for overdispersion, and incorporates dual temporal structure via a first-order random walk (RW1) on μ and a second-order random walk (RW2) on π. While not the focus of this paper, the model has been fully implemented and is ready for posterior inference and validation. In addition to offering narrower and more interpretable credible intervals through Bayesian estimation, this framework is expected to substantially improve model fit and forecasting performance in future work.

B Matericals

B.1 Data Sources

Macroinvertebrate sample metadata: Sample-level information was obtained from `INV_OPEN_DATA_METRICS.parquet`, an Environment Agency (EA) open dataset containing invertebrate monitoring records across England. Fields include sampling date (`SAMPLE_DATE`), sample identifier (`SAMPLE_ID`), sampling method (`SAMPLE_METHOD`), and analysis method (`ANALYSIS_METHOD`). For this study, samples were filtered to retain only those collected using the `S3PO` sampling method and analysed with one of three methods: `ANAA`, `ANLA`, or `ANLE`.

Site metadata: Geographical and waterbody-type information for each site was retrieved from `INV_OPEN_DATA_SITE.parquet`, which includes the reporting area (`REPORTING_AREA`) and site identifier (`SITE_ID`). This dataset was used to link biological samples to their respective locations and waterbody categories.

Biological records: Taxonomic abundance data were obtained from `R_INV_WHPT_METRICS_B.parquet`, a harmonised EA dataset where abundance records from different sampling protocols and identification resolutions have been standardised to the family level. Each record corresponds to a `SAMPLE_ID`–family combination, with abundance counts aggregated accordingly. For the present analysis, only four target families were selected: *Brachycentridae*, *Odontoceridae*, *Cordulegastridae*, and *Aphelocheiridae*.

B.2 R Packages

[1] arrow	21.0.0	[2] cowplot	1.2.0
[3] dplyr	1.1.5	[4] ggplot2	3.5.2
[5] ggpubr	0.6.1	[6] ggsci	3.1.0
[7] lubridate	1.9.4	[8] mgcv	1.9-4
[9] pROC	1.18.6	[10] patchwork	1.2.1
[11] scales	1.3.1	[12] tidyr	1.3.2