School of Geosciences

Dissertation
for the degree of

MSc in Geographical Information Science

Ava Corry-Roberts

August 2025

Statement of Copyright and Originality

I declare that this dissertation represents my own work, and that where the work of others has been used it has been duly accredited. I further declare that the length of the components of this dissertation is **5,006** words for the Research Paper and **7,160** words for the Technical Report.

Copyright of this dissertation is retained by the author and The University of Edinburgh. Ideas contained in this dissertation remain the intellectual property of the author and their supervisors, except where explicitly otherwise referenced.

All rights reserved. The use of any part of this dissertation reproduced, transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise or stored in a retrieval system without the prior written consent of the author and The University of Edinburgh (Institute of Geography) is not permitted.

- I agree that this dissertation and associated electronic documents, web pages, data, files and computer programs can be retained by the University. YES
- I agree that, with the permission of my supervisor(s) or the Programme Director, these materials be made available for the purposes of preparing a publication. YES
- I agree that, with the permission of my supervisor(s) or the Programme Director, these materials can be used within the University of Edinburgh for continued research or teaching. YES

SIGNATURE:

DATE: 7th August 2025

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Dr. Neil Stuart and Peter-John Meynell, for their patience, insightful comments, and thoughtful recommendations throughout the development of this dissertation. I am also deeply thankful to for Peter-John's invaluable local knowledge. I gratefully acknowledge Dr. Kongmeng Ly and Dr. Phan Nam Long of the MRC for their generosity in compiling and sharing essential data. Finally, I want to express my heartfelt thanks to my family and friends for their unwavering support during this intense academic year.

Evaluating the Proxy Potential of Land Use and Macroinvertebrates for Water Quality Assessment in the Lower Mekong Basin

Part I: Research Paper

Abstract

Effective water quality monitoring in the Lower Mekong Basin (LMB) is challenged by the region's vast spatial extent, limited station coverage, and the logistical complexity of chemical sampling. Yet, given the basin's ecological significance and the millions of people who depend on its freshwater systems, scalable approaches to aquatic assessment are critically important. This study evaluates whether upstream land use patterns and downstream macroinvertebrate communities can reliably serve as proxy indicators for surface water quality in the LMB.

Using dry-season data from 34 monitoring stations spanning tributary and mainstem sites, the study integrated geospatial land cover metrics, physicochemical water quality variables, and benthic and littoral macroinvertebrate metrics to explore proxy relationships. Generalized Additive Models (GAMs), Redundancy Analysis (RDA), and variance partitioning were employed to assess predictive performance across parameters including total suspended solids, pH, dissolved oxygen, nutrient concentrations, and chemical oxygen demand.

Results indicate that both land use, particularly urban and agricultural cover, and macroinvertebrate metrics, such as ATSPT and richness, are independently and complementarily associated with water quality variation. Integrated models (with land use and macroinvertebrates as predictors) showed the strongest explanatory power, with proxy strength varying by water quality parameter. Temporal analysis revealed changing macroinvertebrate composition and increasing anthropogenic land use, underscoring the importance of multi-year monitoring.

By demonstrating the viability of spatial and biological proxies, this study supports a more scalable framework for freshwater assessment. It offers basin-scale insight into ecological condition and lays the groundwork for proxy-based monitoring strategies that are accessible, cost-effective, and applicable to other data-limited river systems.

1. Introduction

1.1 Study Region

Spanning over 2,300 km across Vietnam, Thailand, Laos, and Cambodia, the Lower Mekong Basin (LMB) is one of Southeast Asia's most ecologically important and socioeconomically significant freshwater systems (Figure 1) (Mekong River Commission, 2024). Nearly 65 million people depend directly on the basin's resources for agriculture, fisheries, and domestic water access, with approximately 80% living in proximity to its rivers and wetlands (Mekong River Commission, 2024). However, this lifeline is increasingly threatened by intensifying anthropogenic pressures, including rapid urbanization, agricultural expansion, and hydropower development (Whitehead *et al*., 2019). These drivers contribute to rising eutrophication, sedimentation, and pollution events, which degrade water quality and undermine aquatic ecosystem resilience (Chiarelli *et al.*, 2020; Dickens *et al*., 2018; Mekong River Commission, 2021; Tromboni *et al*., 2021).

To facilitate regional cooperation and environmental regulation, the Mekong River Commission (MRC) developed a transboundary monitoring network of 48 water quality (WQM) stations and 41 ecological health (EHM) stations across mainstem and tributary rivers (Figure 1) (Mekong River Commission, 2010, 2019). Despite these efforts, there are still too few stations to adequately cover the basin's spatial and ecological complexity. Further, significant discrepancies remain between formal water quality reports and academic field studies; while MRC data often report good water quality, independent research highlights degradation and ecological stress (Sor *et al*., 2021; Whitehead *et al*., 2019). These contradictions suggest the need for more nuanced, integrated approaches to monitoring that combine physiochemical and biological indicators with spatial analysis.

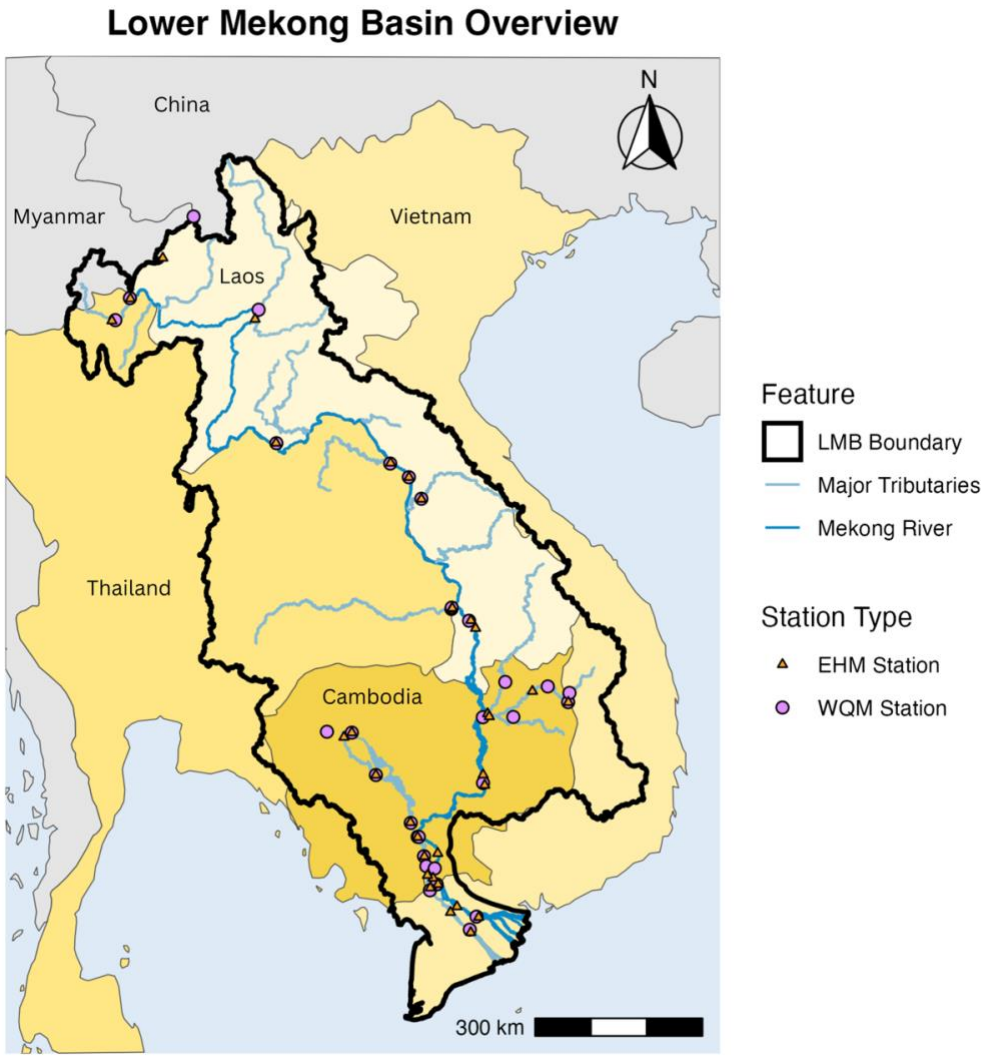

Figure 1: Map of the Lower Mekong Basin showing the spatial distribution of Mekong River Commission monitoring stations: for water quality and for ecological health. *These stations support transboundary assessments across mainstem and tributary rivers, forming the basis for regional ecological and physiochemical monitoring. Note the uneven spatial distribution, which contributes to data gaps and challenges in basin-wide assessments.*

1.2 Potential of Proxies for Water Quality Indicators

Relying solely on chemical monitoring may overlook episodic disturbances and lacks basin-wide coverage, while requiring considerable resources (Sor *et al*., 2017). In contrast, biological indicators, particularly aquatic macroinvertebrates, may reflect water quality conditions more completely (De Pauw, Gabriels, & Goethals, 2006; Błachuta *et al*., 2014). Aquatic macroinvertebrates occupy diverse ecological niches and display varied pollution tolerance, making them sensitive to both short- and long-term environmental stress (Sripanya *et al*., 2023; Tampo *et al.*, 2021). Metrics based on community composition, abundance, richness, and pollution tolerance scores (e.g., ATSPT) provide powerful insights into aquatic conditions over time (Dickens *et al*., 2018).

Upstream land use and land cover (LULC) are also recognized as key spatial determinants of water quality (Pakoksung *et al*., 2025; Yao *et al.*, 2023*)*. Urban landscapes increase impervious surfaces and point-source pollutants, while agricultural areas increase nutrient loads, sedimentation, and habitat disruption (Azrina *et al*., 2006; Ongley, 2009). In contrast, forested regions often act as natural buffers, stabilizing hydrological regimes and mitigating chemical disturbances (Cheng *et al.*, 2022). These land use patterns underscore the potential for LULC to serve as a spatial proxy for water quality. Landscape metrics, such as vegetative buffer width, impervious surface coverage, and land use proportions, have been associated with variations in pH, total nitrogen (TN), total phosphorus (TP), total suspended sediment (TSS), and dissolved oxygen (DO) (Tampo *et al*., 2021; Fierro *et al*., 2017; Yao *et al*., 2023).

GIS technologies help link upstream land use dynamics to downstream water quality trends and enable the development of predictive models (Pakoksung *et al*., 2025; Sor *et al*., 2017; Yao *et al.*, 2023). The integration of landscape-water quality relationships into GIS workflows supports scalable assessments that are both cost-effective and ecologically meaningful (Dickens *et al*., 2018; Sor *et al*., 2017).

1.3 Research Questions and Aims

Against this backdrop, this dissertation investigates whether land use composition and macroinvertebrate metrics can serve as reliable proxies for water quality in the LMB. Drawing on a decade of monitoring data (2011–2021), the study applies geospatial analysis and ecological modeling to explore how upstream LULC patterns and biological communities correspond with key water quality parameters.

The research is guided by two key questions:

1. Can land use composition serve as a strong predictor of in-situ water quality across the LMB?
2. Do macroinvertebrate community metrics act as reliable biological indicators of water quality?

It is hypothesized that catchments dominated by urban or agricultural land will exhibit elevated nutrients, altered pH, increased sediment loads, and reduced dissolved oxygen. In contrast, forested catchments are expected to buffer pollutants and maintain more stable water chemistry.

Ecologically, poor water quality is anticipated to correspond with reduced macroinvertebrate richness and increased dominance of pollution-tolerant taxa, as measured through ATSPT scores.

Ultimately, this dissertation seeks to strengthen proxy-based monitoring frameworks for tropical freshwater systems through interdisciplinary analysis. While proxies are not standard practice within the MRC's current monitoring regime, research has shown that combining geospatial data, ecological indicators, and water chemistry metrics can produce reliable

assessments of water quality (Iwasaki, Suemori, & Kobayashi, 2024; Locke, 2024; Tromboni *et al*., 2021). If consistent relationships can be demonstrated in the LMB, proxy indicators may offer results comparable to conventional in-situ measurements and serve as practical alternatives. This approach supports a more scalable, replicable, and policy-relevant framework for monitoring and managing freshwater ecosystems in the LMB and similar data-limited regions.

2. Methods

2.1 Study Area

The LMB is characterized by diverse hydro-ecological conditions, with differences in river morphology, land use, and ecological integrity between the Mekong mainstem and its tributaries (Mekong River Commission, 2024). To reflect this diversity, the MRC selected WQM and EHM sites across both river types to capture broad spatial and environmental gradients (Mekong River Commission, 2010, 2019).

WQM sites collect physiochemical water data monthly, while EHM sites assess biological indicators (littoral and benthic macroinvertebrates, benthic diatoms, and zooplankton) annually (Mekong River Commission, 2010, 2019).

Temporal analysis was conducted on six biennial years based on data availability (2011, 2013, 2015, 2017, 2019, and 2021), allowing for assessment of macroinvertebrate composition condition, land use change, and water quality variation across the basin.

2.2 Site Pairing Framework

The flowchart of data analysis is shown in Figure 2. Aligned study sites were selected where both macroinvertebrate and water quality data were available. Each EHM site was matched to the nearest upstream WQM site within 15 km using Euclidean distance (Figure 3) (Iwasaki, Suemori & Kobayashi, 2024). If multiple EHM sites were downstream of the same WQM site, all valid pairings were retained.

Site pairs were then examined to check for suitability in ArcGIS Online. If the two sites did not occur at the same location, their suitability was assessed based on whether the two sites were located within the same river, the absence of inflow of major tributaries and changes in land use between the sites, and the availability of other more suitable WQM sites.

This process yielded 34 matched site pairs: 14 on the mainstem and 20 on tributary rivers.

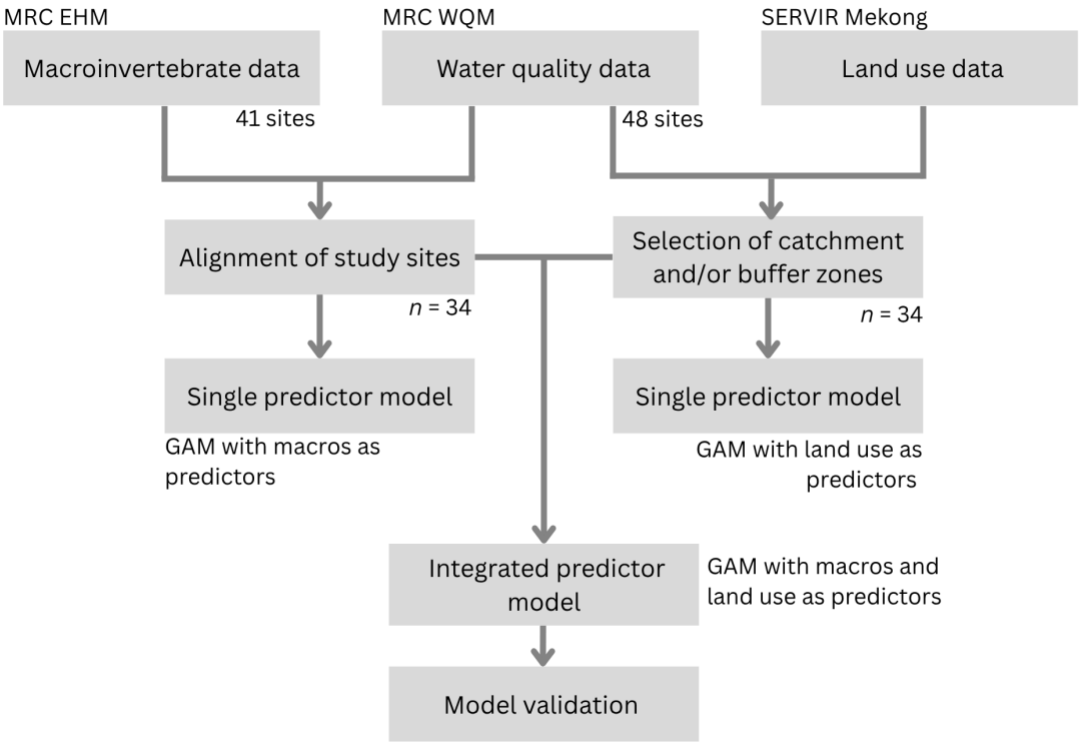

Figure 2: Schematic design of the methodology used to join data to run single and integrated predictor models.

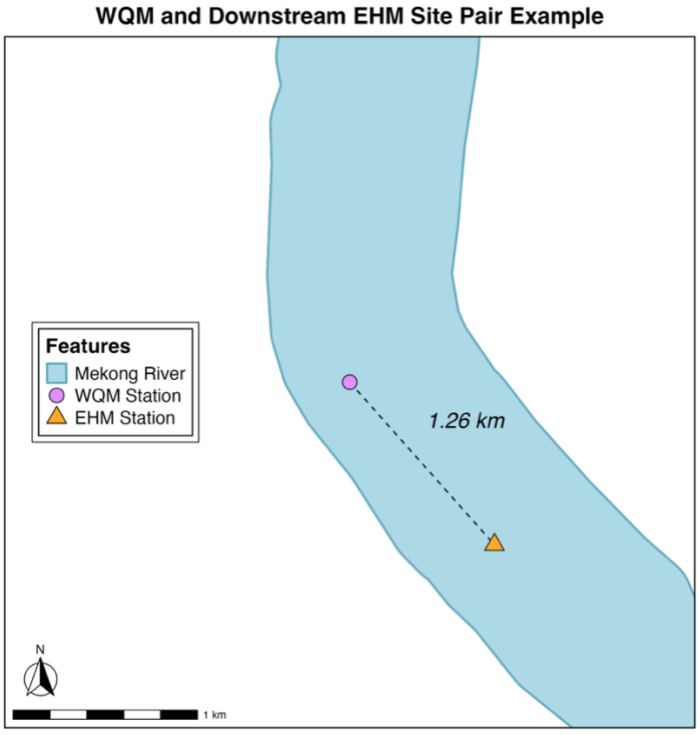

***Figure 3*: Example of a matched site pair along the Mekong River.** *The map shows a WQM station matched to its nearest upstream EHM station with a line representing the distance between sites.*

2.3 Water Quality Data

Water quality data was obtained from the MRC's monthly sampling dataset for temperature (°C), pH, total suspended sediment (mg/L), electrical conductivity (mS/m), total nitrogen (mg/L), total phosphorus (mg/L), dissolved oxygen (mg/L), and chemical oxygen demand (COD) (mg/L). Flow (m^3/s), biological oxygen demand (mg/L), and faecal coliforms (MPN/100mL) were excluded from statistical analysis due to missing data (>60% missing records).

The dataset was filtered by study year and limited to dry season months (December to March) to align with March macroinvertebrate sampling, ensuring macroinvertebrate responses reflect antecedent water quality conditions (Holguin-Gonzalez *et al*., 2013; Jerves-Cobo *et al*., 2020; Mekong River Commission, 2010).

2.4 Macroinvertebrate Data

Macroinvertebrate data was provided by the MRC for both benthic and littoral habitats. The following metrics were assessed:

1. Taxa Richness: Total number of macroinvertebrate taxa per sample
2. Abundance: Total count of macroinvertebrate individuals per sample
3. Average Tolerance Score per Taxon (ATSPT): Pollution tolerance index calculated as the average of individual taxa tolerance scores

Healthy ecosystems are indicated by high abundance, high richness and low ATSPT as organisms respond to the quality of the aquatic environment (Mekong River Commission, 2010).

To reflect comprehensive measures of macroinvertebrate conditions, benthic and littoral scores were aggregated. Richness and abundance were summed by simple addition:

$$Total\ Abundance = BM_{Abundance} + LM_{Abundance}$$

$$Total\ Richness = BM_{Richness} + LM_{Richness}$$

ATSPT scores, originally calculated as unweighted means, were combined using a richness-weighted formulation to account for differences in taxa representation across benthic and littoral zones:

$$Weighted\ ASPT = \frac{(BM_{ASPT} * BM_{Richness}) + (LM_{ASPT} * LM_{Richness})}{BM_{Richness} + LM_{Richness}}$$

The weighted approach ensured proportional influence from each habitat based on taxa presence, rather than sample volume (Iwasaki, Suemori, & Kobayashi, 2024; Vadeboncoeur, McIntyre & Vander Zanden, 2011).

2.5 Land Use Data

To evaluate upstream land use influences, WQM sites were categorized by river type. Tributary sites used hydrological catchment boundaries, while mainstem sites used buffer zones to determine land use area (Figure 4; Figure 5) (Zhu *et al*., 2024). The Mekong mainstem flows between catchments, making catchment boundaries less suitable for mainstem WQM sites. Instead, a 15 km buffer was applied to the eight geomorphologically distinct BioRA Zone segments based on the MRC's transboundary impact assessment (Figure 5) (Mekong River Commission, 2017). For additional details on zone delineation, see section 2.2.1 of the Technical Report.

Areas of land use for each site were reviewed in ArcGIS Online. Where major tributaries joined a primary catchment or buffer, their catchment land use was included in the total upstream land area to ensure comprehensive landscape representation for each site.

Land cover data was acquired from the SERVIR-SEA Regional Land Cover Monitoring System (RLCMS), which included 18 land cover classes at a spatial resolution of 30 x 30 meters. For each site's land use area, proportions of urban, agricultural, forest, and semi-natural areas were derived as indicators land use pressure on water quality (Corry-Roberts, 2025; Locke, 2024).

Land use analysis was performed in R version 4.4.2 (R Core Team 2024) using the 'sf' and 'terra' packages. These tools were used to extract land cover area, calculate percent cover, and analyze temporal change across the study period. These outputs formed the basis for modeling landscape-water quality relationships.

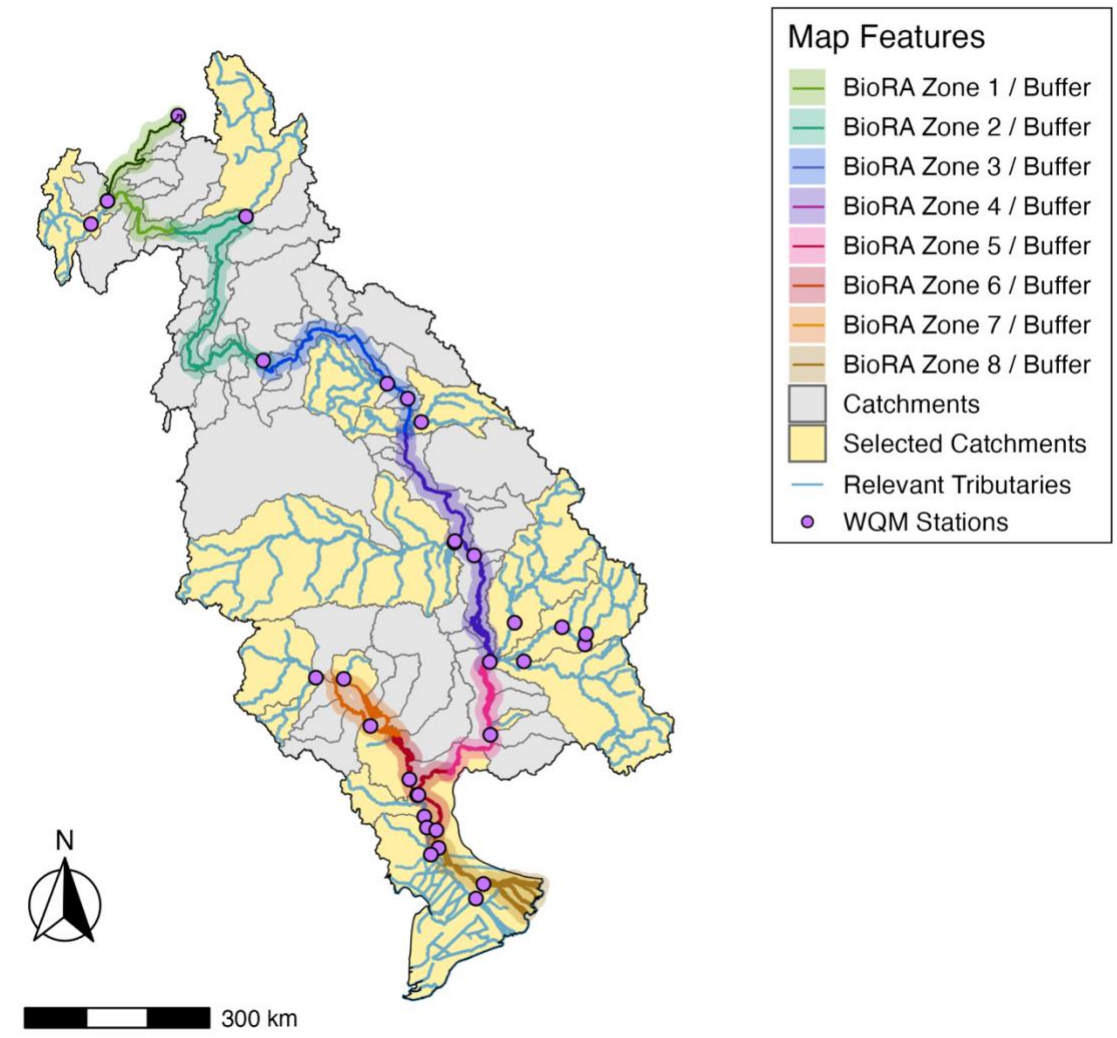

Figure 4: Selected catchments and buffered tributary zones used for land use analysis in the Lower Mekong Basin. *Tributary sites were delineated using hydrological watershed boundaries, while mainstem sites were defined by 15 km buffers around BioRA Zone segments. These spatial units formed the basis for quantifying upstream land cover influences on water quality.*

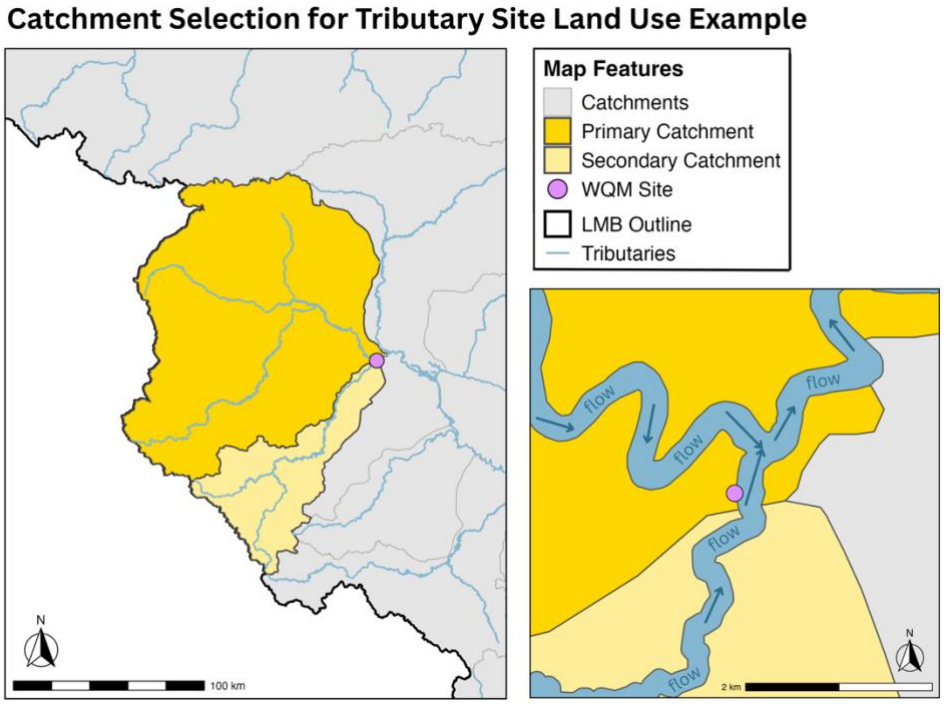

***Figure 5: Example of catchment selection for land use analysis at a WQM site**. Land use from a secondary catchment was included in the total upstream area due to a tributary confluence just upstream of the monitoring station. As a result, landscape conditions in both primary and secondary catchments may influence water quality at this site.*

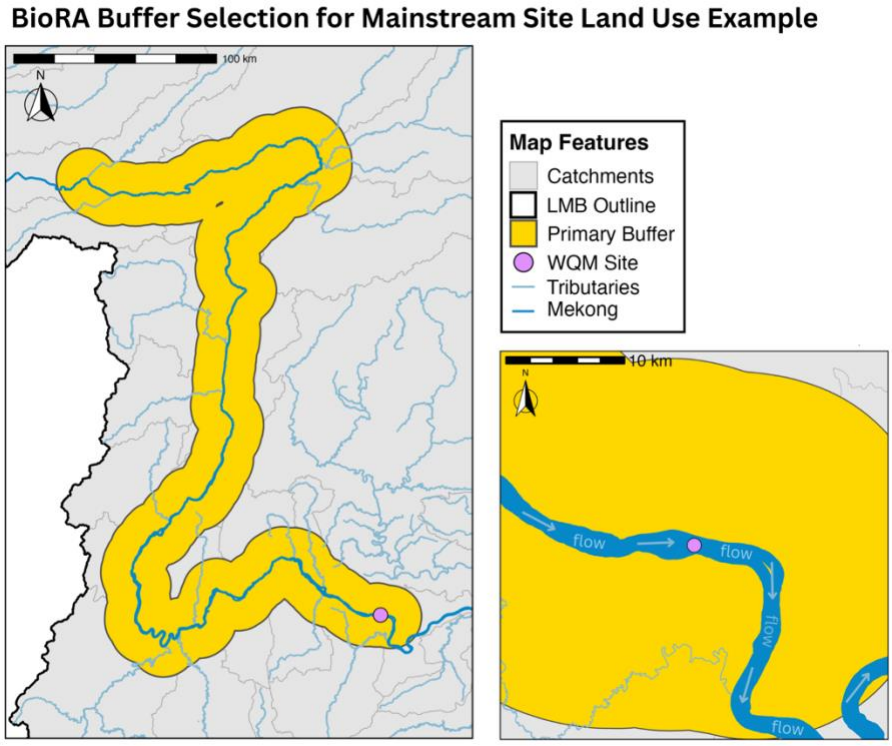

***Figure 6: Area selection for land use analysis at a Mekong WQM site.** No secondary catchments were included in this analysis, as tributaries joined the mainstem too far upstream to significantly influence water quality at the monitoring location.*

2.6 Statistical Analysis and Modeling Framework

All statistical analyses were performed in R (v 4.2.2) using 'vegan', 'mgcv', 'corr', and 'ggplot2' packages. A hierarchal analytical framework was used to identify spatial and temporal drivers of water quality and evaluate the potential of land cover and macroinvertebrate metrics as proxy indicators (Figure 2).

To account for environmental variability over time, both pooled and year-specific models were constructed. Pooled models evaluated consistency and average explanatory strength of predictors across the region, while annual models tested sensitivity to episodic changes, disturbance events, or sampling anomalies (Bignert *et al.*, 2014).

2.6.1 Single Predictor Proxy Modeling

To explore proxy potential for water quality variation, Redundancy Analysis (RDA) was conducted in two strands. First, macroinvertebrate metrics were analyzed alongside WQ parameters to characterize biological responses to water chemistry. Second, RDA was used to assess identify land use influences on water quality gradients (Cheng *et al*., 2022). These ordinations reduced dimensionality and provided a multivariate overview of ecological and spatial relationships to water quality.

Generalized Additive Models (GAMs) were then applied to examine nonlinear relationships in both directions: macroinvertebrate metrics predicting WQ, and land cover variables predicting WQ. Each model included a single predictor, allowing for targeted assessment of whether biological or spatial indicators alone could reliably serve as proxies for individual water quality metrics (Beck *et al*., 2022; Duque *et al*., 2022; Holguin-Gonzalez *et al*., 2013; Murphy *et al*., 2019).

2.6.3 Integrated Predictor Proxy Modeling

Full GAMs were built using both land cover and macroinvertebrate predictors to assess joint explanatory power. Variance partitioning quantified independent and shared contributions from land use and macroinvertebrate metrics to clarify each predictor's role in proxy modeling (Cheng *et al*., 2022; Peres-Neto *et al*., 2006).

Model performance was evaluated using the coefficient of determination (R^2) and root mean square error (RMSE). Four sites that were withheld from model fitting were used for model validation. Predicted values from the final model were compared to observed values to assess reliability.

3. Results

3.1 Spatial-Temporal Trends

From 2011 to 2021, land cover across the LMB shifted increasingly toward non-natural land use. Paddy rice and evergreen forest were dominant land types in 2011. Over the next decade, human-associated land covers expanded: rubber plantations increased by 92.2%, cropland by 70%, and crop plantations by 44.9% (Figure 7; Figure 8). Concurrently, natural covers declined,

with wetlands decreasing by 36.1%, deciduous forest by 18.6%, and evergreen forest by 13.5% (Figure 9). These patterns reflect sustained landscape conversion toward intensified agriculture and plantation production.

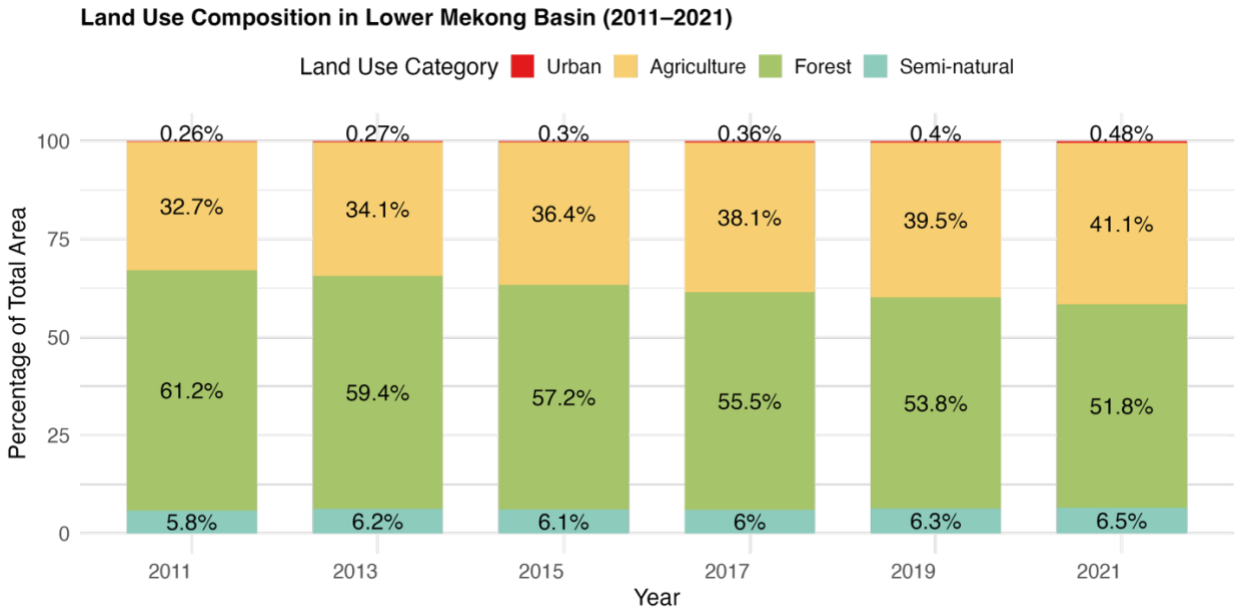

Figure 7: Temporal changes in land cover composition across the Lower Mekong Basin from 2011 to 2021. *Green segments represent natural land covers, which declined from 61.2% to 51.8%, while yellow segments (agriculture-associated land uses) increased from 32.7% to 41.1%. These trends reflect sustained landscape conversion toward intensified agriculture and plantation development. See Figure 8 for a spatial representation.*

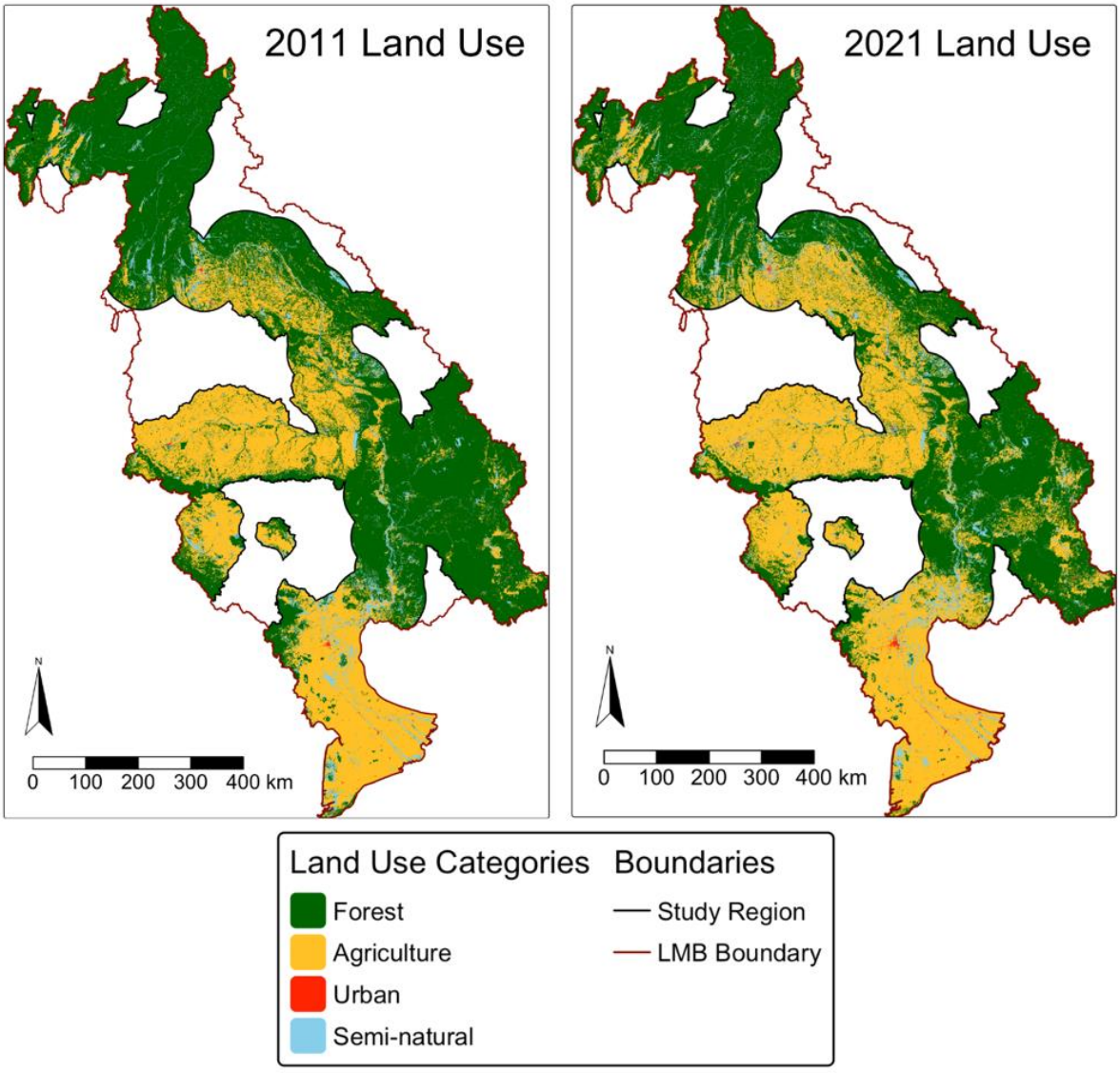

Figure 8: Land use patterns in the Lower Mekong Basin study region in 2011 (left) and 2021 (right). *Human-associated land covers, including rubber plantations, cropland, and crop plantations, expanded markedly over the decade, while natural land types such as evergreen forest, deciduous forest, and wetlands declined. These maps illustrate the spatial extent of landscape conversion underlying regional land cover trends.*

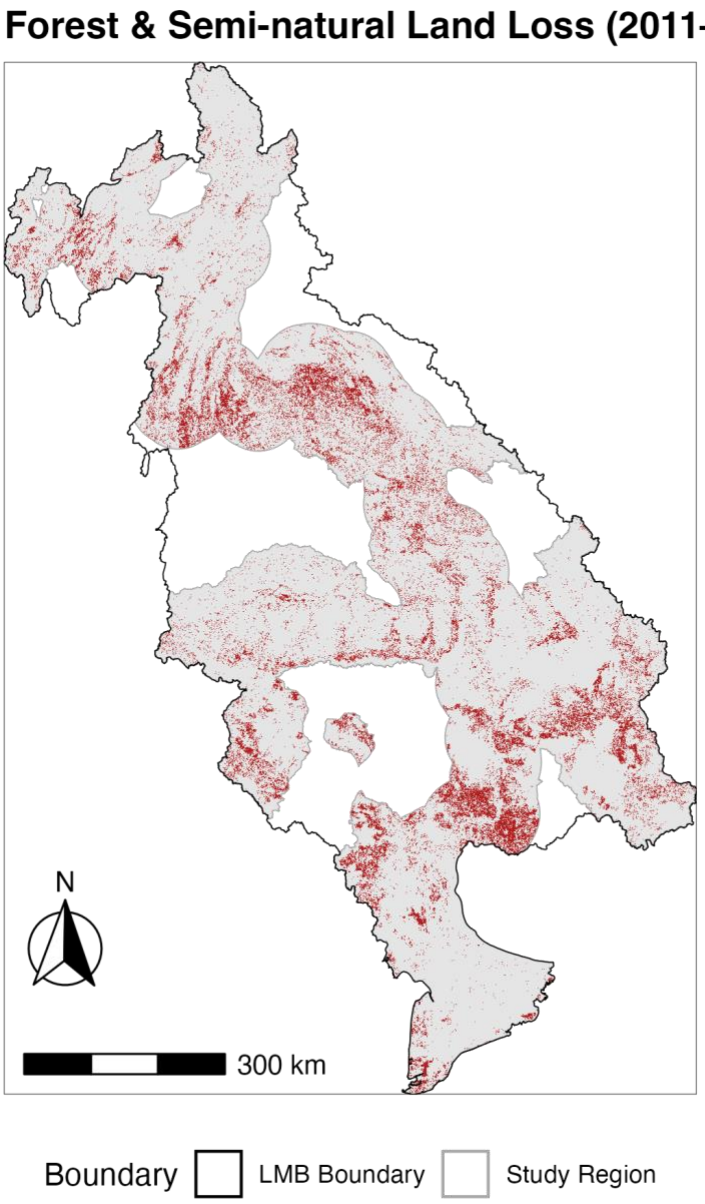

Figure 9: Spatial distribution of anthropogenic land cover expansion across the Lower Mekong Basin between 2011 and 2021. *Red markers indicate areas with pronounced increases in rubber plantations, cropland, and crop plantations. These changes reflect intensified agricultural and plantation development, contributing to regional landscape transformation.*

Surface water chemistry showed site-level variation but limited change over time. TP was generally low, although stations in Cambodia frequently exceeded ecological thresholds (Figure 10). TN averaged 0.4 mg/L, peaking at LHK and LPS. TSS varied widely, averaging 45 mg/L and peaking above 1,500 mg/L in CPK (Figure 11). Conductivity was mostly stable, with occasional spikes >200 mS/m in disturbed sites.

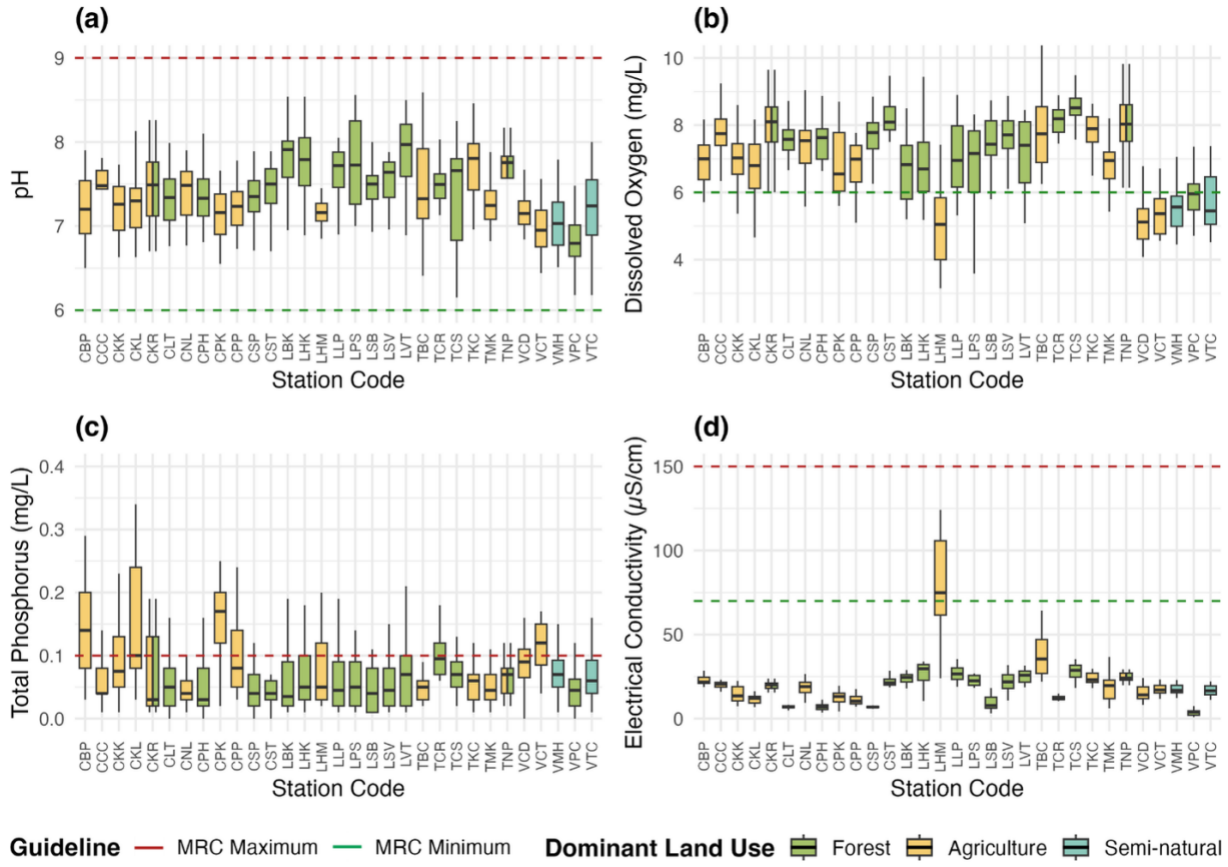

Figure 10: Box plots of water quality parameters across monitoring stations in the Lower Mekong Basin from 2011 to 2021: (a) pH, (b) dissolved oxygen (DO), (c) total phosphorus (TP), and (d) electrical conductivity (EC). *Dashed lines indicate ecological thresholds: green for acceptable ranges and red for exceedances. Box plot colours correspond to dominant land use types at each station, highlighting associations between landscape composition and water quality variation.*

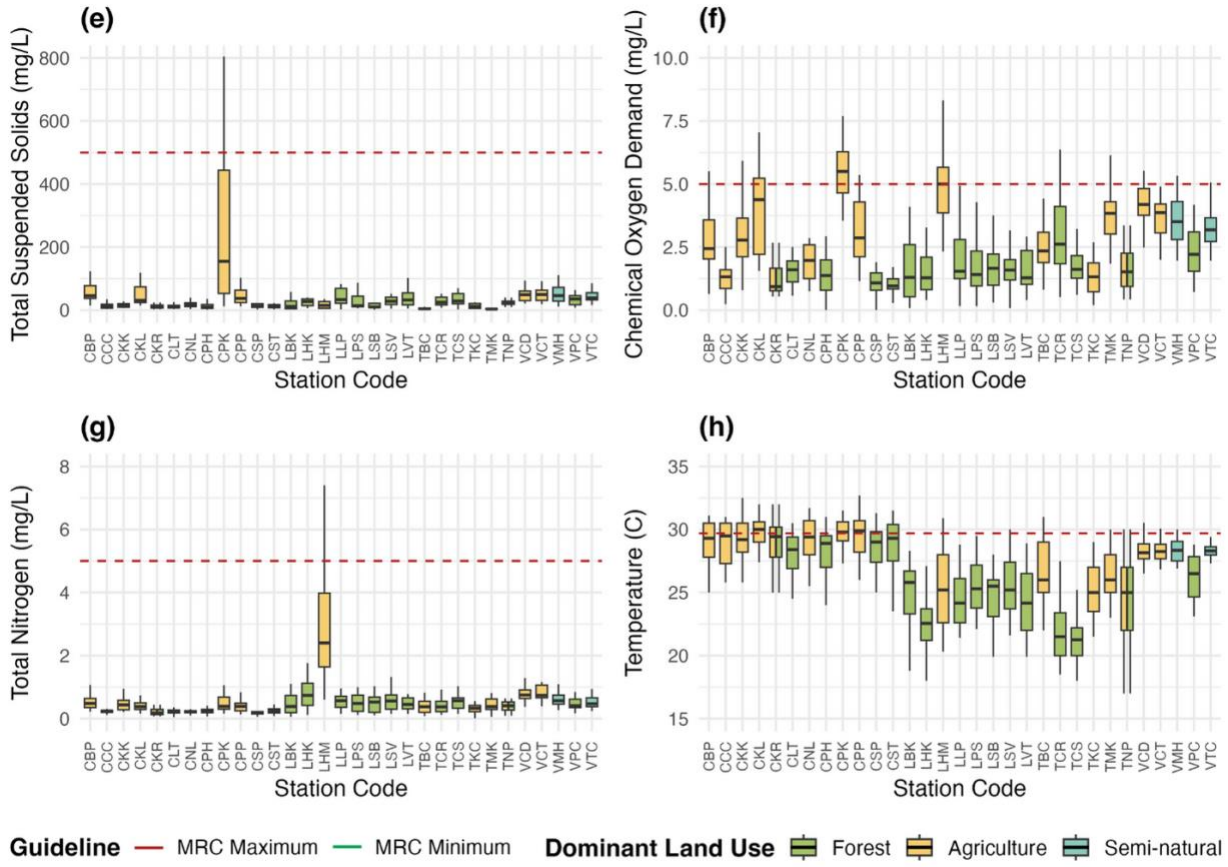

Figure 11: Box plots of water quality parameters across monitoring stations in the Lower Mekong Basin from 2011 to 2021: (e) total suspended solids (TSS), (f) chemical oxygen demand (COD), (g) total nitrogen (TN), and (h) temperature. *Dashed lines indicate ecological thresholds: green for acceptable ranges and red for exceedances. Box plot colours correspond to dominant land use types at each station, highlighting associations between landscape composition and water quality variation.*

Macroinvertebrate metrics also reflected diverse regional responses. Abundance increased in Cambodian sites, while decreasing in Lao PDR and Thailand. Richness improved basin-wide, although one site in Viet Nam remained below thresholds in 2021. ATSPT rose steadily in Cambodia, Lao PDR, and Thailand, indicating increased pollution tolerance.

3.2 Statistical Analysis Results

3.2.1 Land Use Water Quality Relationships

Redundancy analysis confirmed significant associations between upstream land cover and surface water chemistry (F = 12.29, p = 0.001) with urban land explaining the most variance, followed by agricultural, and forest cover (Figure 12). Collectively, these predictors explained 15.6% of total water quality variance.

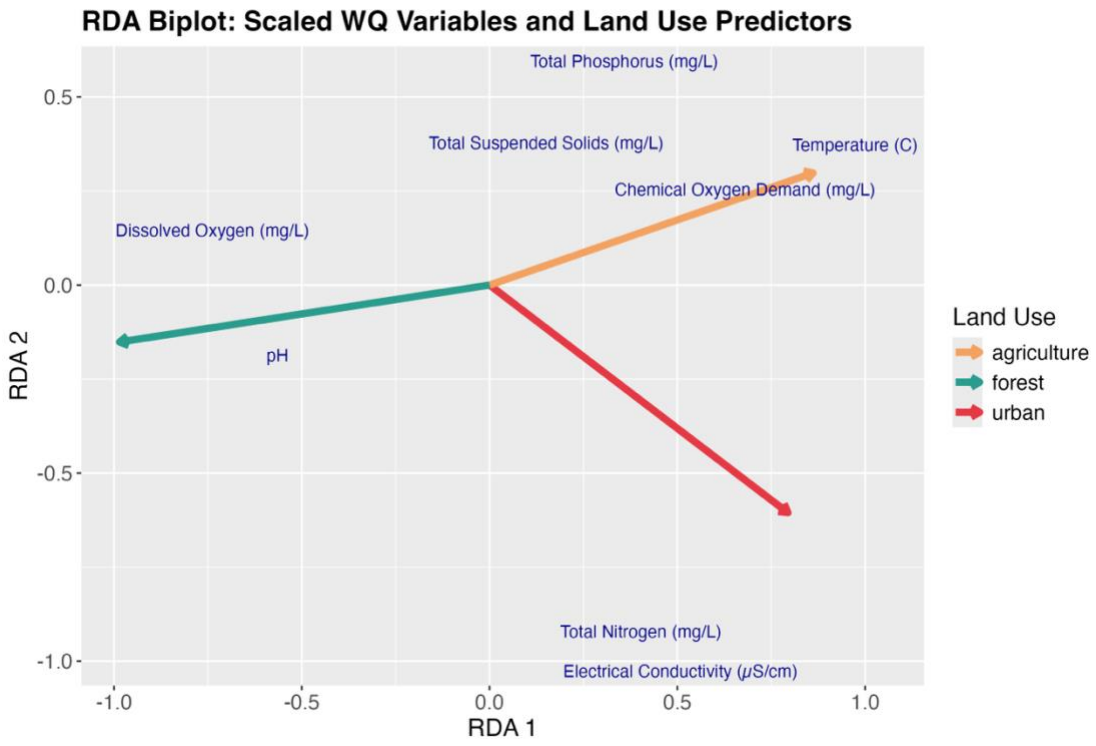

Figure 12: RDA biplot showing associations between land cover types and surface water chemistry in the Lower Mekong Basin. *Arrows represent explanatory variables; blue labels indicate water quality response variables. The direction and length of arrows reflect the strength and nature of each land cover type's influence on water chemistry.*

The single predictor pooled GAM supported land use as a meaningful proxy for water quality. Conductivity was predicted with high accuracy (*Adj.* R^2 = 0.52), with significant contributions from all land use types ($p < 0.001$). Moderate predictive power was observed for COD (R^2 = 0.37) and DO (R^2 = 0.36), with all land use variables showing significance (Figure 13).

Urban and agricultural land cover were the most consistently significant indicators of reduced water quality, particularly for DO, COD, TN, and conductivity. Lower predictive strength was observed for pH, TP, and TSS, though urban and agricultural cover still showed some influence on these parameters. Forest land cover showed a generally weak but statistically significant relationship with most water quality parameters, with particularly strong significance for DO, TP, and COD. However, its explanatory power was modest compared to urban and agricultural predictors, suggesting forest cover is not the dominant driver of water quality.

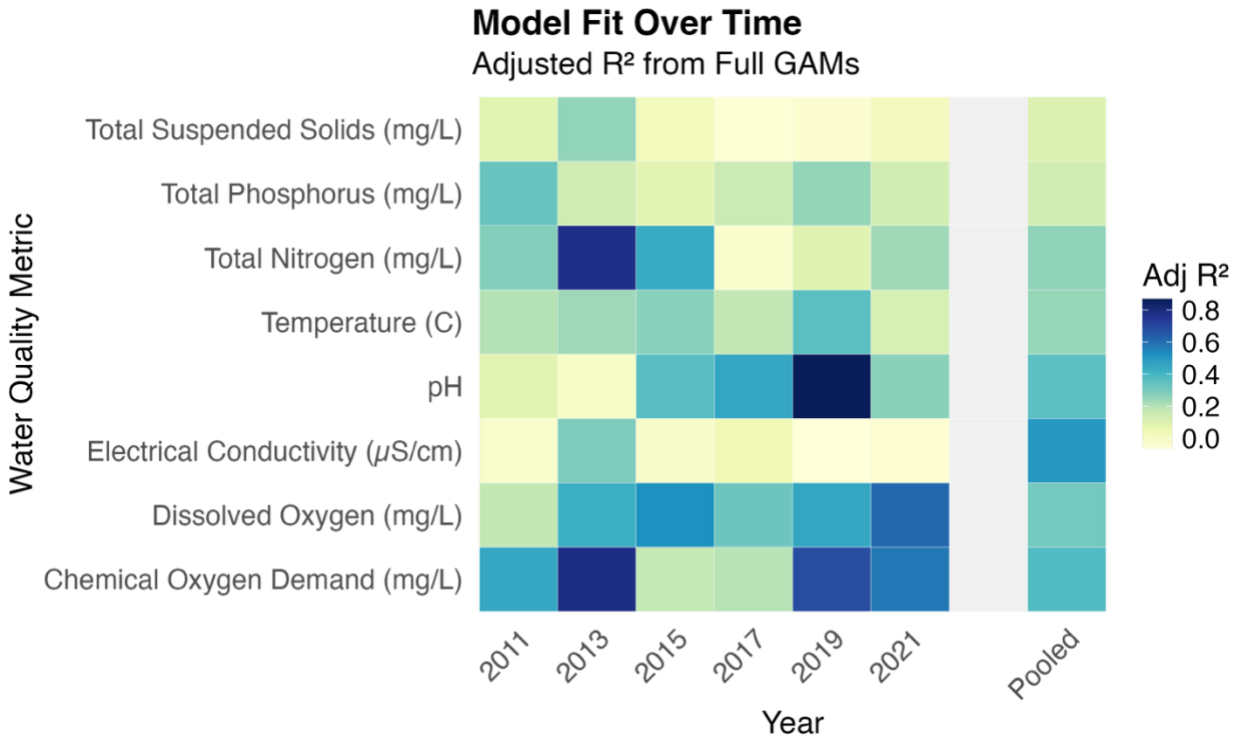

Figure 13: Adjusted R^2 values from yearly land use-only GAMs (left) and a pooled regional model (right) for water quality parameters in the Lower Mekong Basin, 2011–2021.

Year-specific GAMs revealed that the strength and significance of relationships varied across sample years (Figure 13). Models in 2013, 2019, and 2021 performed the strongest, particularly for TN, COD, and DO. Urban cover consistently predicted elevated TN and COD in both 2013 and 2021 ($p < 0.001$), highlighting its role in nutrient and organic pollution. Forest cover showed weaker but occasionally significant associations, with negative relationships to COD and DO degradation in 2013 and 2021. In contrast, models in 2015 and 2017 were weaker and less consistent, with lower explanatory power and several predictors failing to reach significance.

3.2.2 Macroinvertebrate Water Quality Relationships

Redundancy analysis revealed a significant relationship between macroinvertebrate assemblages and overall water quality predictors (Figure 14) ($F = 13.42$, $p = 0.002$). Among the variables, TSS exerted the strongest influence ($F = 85.16$, $p = 0.003$), followed by temperature ($F = 12.27$, $p = 0.001$) and pH ($F = 3.67$, $p = 0.037$). Other variables, including conductivity, TN, TP, DO, and COD, were not statistically significant contributors ($p > 0.05$).

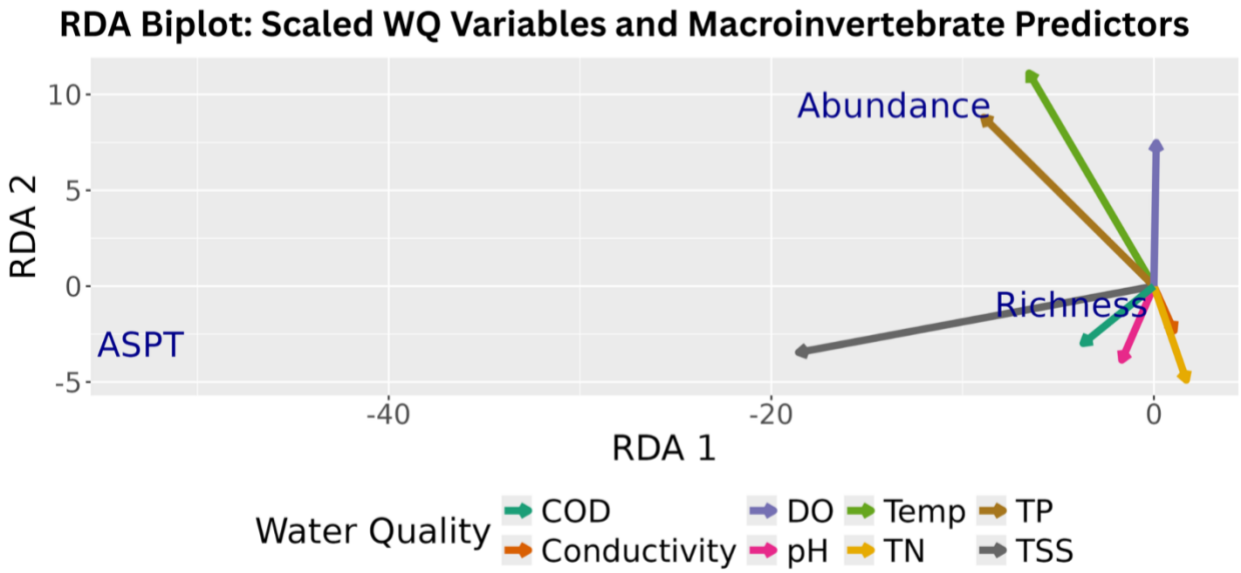

Figure 14: RDA biplot illustrating relationships between scaled water quality variables and macroinvertebrate community metrics across the LMB.

The single predictor pooled GAM enhanced ecological interpretability, revealing nonlinear and context-sensitive associations (Figure 15). TSS was predicted with high accuracy (*Adj.* $R^2 = 0.80$), driven primarily by ATSPT ($p < 0.001$). Moderate predictive power was observed for TP (*Adj.* $R^2 = 0.22$), with all three macroinvertebrate metrics showing statistical significance.

ATSPT was the most consistently significant indicator, particularly for TSS, TP, and temperature, while richness and abundance contributed more sporadically. Predictive strength was weak for pH, DO, TN, and COD, with adjusted R^2 values below 0.13 and limited deviance explained. Conductivity was not reliably predicted by any macroinvertebrate metric.

Year-specific GAMs identified strong biological signal in 2011 and 2015 (Figure 15). In 2011, temperature and pH were predicted with high accuracy (*Adj.* $R^2 = 0.86$ and 0.50). In 2015, TP and DO models exhibited improved fit and statistical relevance, suggesting periods of heightened macroinvertebrate sensitivity.

Richness gained predictive strength in later years. In 2019, it was significant for pH, DO, and COD ($p < 0.05$), suggesting its growing utility in biological proxy models. In 2021, richness contributed meaningfully to TP prediction ($p = 0.0008$), affirming its responsiveness to nutrient dynamics. Abundance was intermittently significant, especially in temperature and TP models, but remained the least predictive overall.

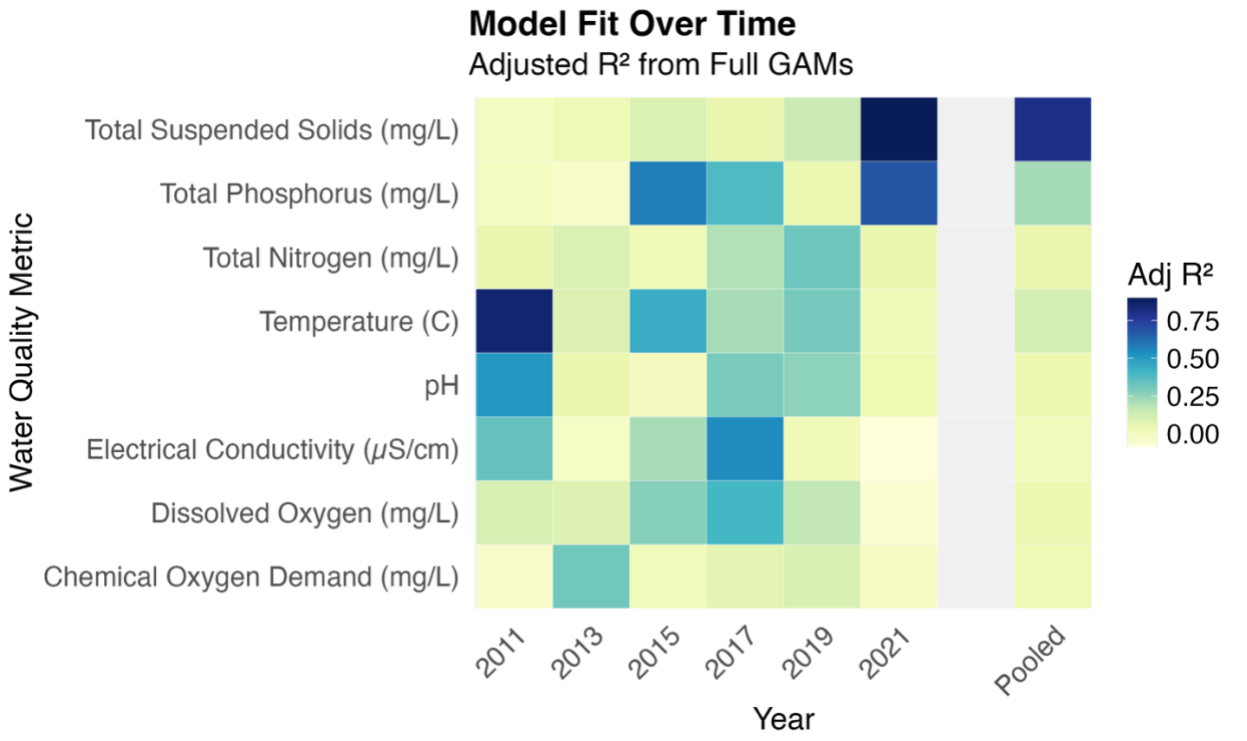

Figure 15: Adjusted R^2 values from yearly macroinvertebrate-only GAMs (left) and a pooled regional model (right) for water quality parameters in the Lower Mekong Basin, 2011–2021.

3.2.3 Integrated Models: Land Use and Macroinvertebrates

GAMs integrating both land use and macroinvertebrate predictors yielded the strongest predictive power across all assessed water quality parameters (Table 1). The pooled model for TSS accounted for 87.1% of deviance, with ATSPT and urban land cover emerging as dominant predictors (Figure 16). Conductivity and DO also exhibited moderate model fit, shaped by highly significant contributions urban and forest cover, alongside weaker but relevant biological metrics. COD was strongly associated with agricultural and forest cover, as well as abundance and richness, while ATSPY retained moderate influence.

***Table 1: GAM results integrating macroinvertebrate metrics and land use predictors for water quality parameters**. Bolded R^2 indicate strong to moderate model strength and p-values indicate significance at $p < 0.05$. P-values reported as 0 are below detection limits and represent highly significant relationships.*

Response	Adj R^2	Deviance Explained	ASPT p	Richness p	Abundance p	Urban p	Agriculture p	Forest p
Temp	**0.403**	43.677	**2.92E-05**	**0.0003**	**0.024**	0.082	0.074	0.216
pH	0.298	35.271	**0.003**	0.091	0.385	**0.0003**	**0.026**	**0.018**
TSS	**0.853**	87.067	**0**	0.817	0.411	**0.0008**	0.126	0.097
Cond	**0.515**	57.577	0.642	0.824	0.691	**0**	**2.04E-05**	**0.0001**
TN	0.315	37.001	0.840	0.908	**0.033**	**0.028**	**6.74E-06**	**0.006**
TP	0.260	30.858	**0.001**	0.068	**0.008**	0.091	0.757	0.223
DO	**0.414**	45.233	0.051	0.238	0.077	**0.0002**	**0**	**0**
COD	**0.396**	42.801	0.376	0.813	0.052	**6.14E-05**	**0.005**	**1.56E-06**

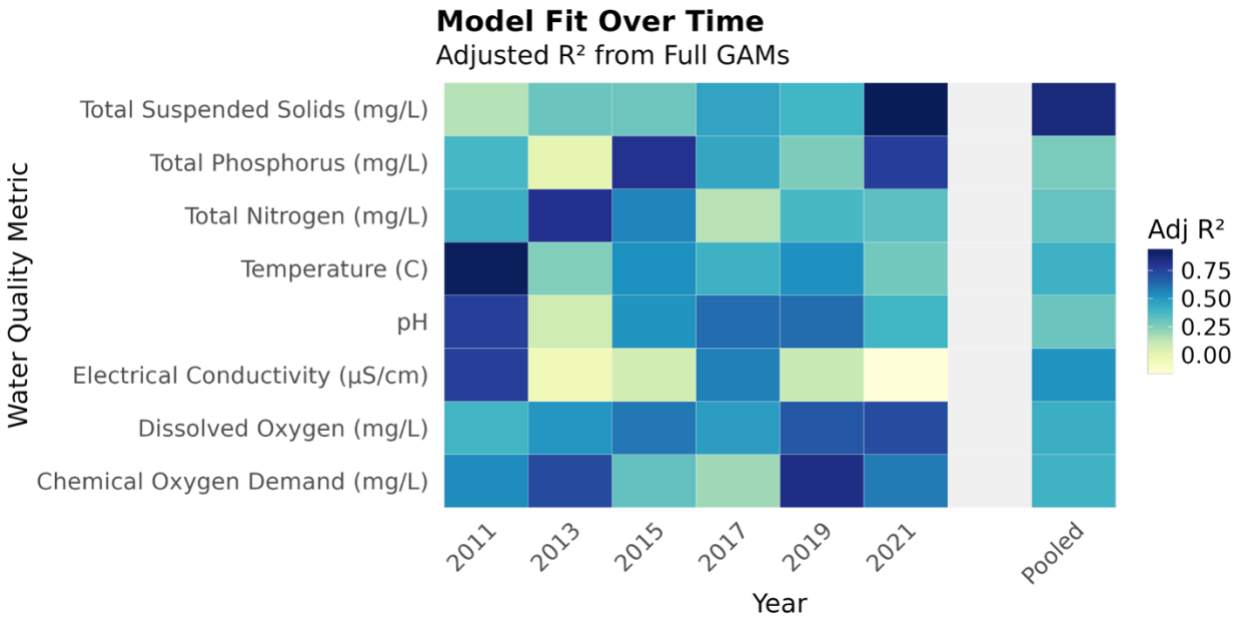

Figure 16: Heatmap showing year-wise correlations between water quality parameters and macroinvertebrate-land use predictors in the Lower Mekong Basin. *Darker shades exhibited heightened sensitivity to both biological and landscape variables, reinforcing the value of integrated proxy models.*

Although models for TP, TN, and pH explained less deviance overall (*Adi.* R^2 = 0.26-0.32), the combined influence of spatial and biological metrics revealed ecologically meaningful relationships. Among biotic predictors, ATSPT consistently demonstrated strong associations with water quality, significantly influencing TSS, TP, pH, and temperature. Richness showed moderate predictive strength, contributing meaningfully to temperature, TP, and COD, while abundance retained significance in models for temperature, TP, and TN, though its overall influence remained limited.

Urban land use emerged as a dominant spatial predictor, significantly associated with pH, TSS, conductivity, TN, DO, and COD ($p < 0.001$), reinforcing its role in driving chemical and physical degradation. Agricultural land use strongly influenced TP, TN, and COD, particularly in nutrient-enriched systems. Forest cover significantly predicted temperature, conductivity, TP, and COD, with effects likely reflective of buffering capacity and landscape integrity rather than direct pollutant inputs.

The pooled models captured robust, average effects across the study period, while the yearly models revealed important temporal variation in both the strength and significance of predictor variables. Year-specific full GAMs showed peak model strength in 2015 and 2021. DO models in 2021 explained 80.6% of deviance, with statistical significance across ATSPT, abundance, urban, agriculture, and forest ($p <$ 1e-05). Richness and forest showed increased influence in later years, which reinforced the value of integrated proxy approaches.

3.2.4 Variance Partitioning

Variance partitioning was conducted to quantify the unique and shared contributions of macroinvertebrates and land use to water quality variation. In the pooled model, macroinvertebrates uniquely explained 13.3% of variance ($F = 6.99$, $p = 0.001$), while land use accounted for 7.6% ($F = 11.45$, $p = 0.001$). Shared variance remained negligible (–0.02%), affirming the non-redundant nature of the two predictor sets. Residual variance remained high (79%), suggesting additional influences from hydrological, geomorphological, or unmonitored sources.

3.2.5 Model Validation

To assess the predictive performance of the pooled GAMs, four sites withheld from model fitting were used for validation. Model predictions closely matched observed values for pH, DO, COD, TN, TP, and temperature, with low RMSE across years and sites (Figure 17). These results indicate that the spline structures in the models were well calibrated to the underlying ecological and land use gradients.

In contrast, TSS and conductivity were less reliably predicted. Despite high adjusted R^2 and deviance explained for TSS in training models, TSS, predictions diverged sharply from observations in certain years. Predictions for conductivity were more erratic, suggesting that the model may not have fully captured its nonlinear dynamics.

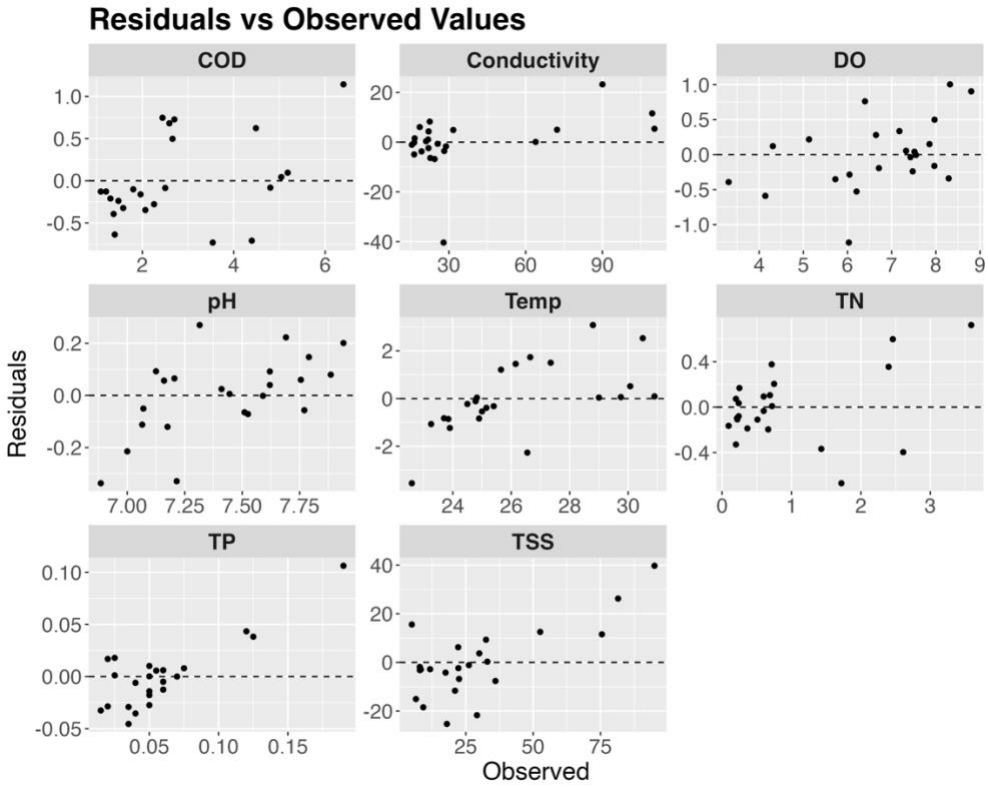

Figure 17: Observed vs residual plots for pooled GAM validation across water quality parameters.

4. Discussion

This study set out to evaluate whether upstream land use and macroinvertebrate community metrics can serve as reliable proxies for surface water quality across the LMB. The results support this hypothesis for select water quality variables. While only a subset of water quality variable showed strong model fits, the integration of predictor improved explanatory power and revealed ecologically interpretable relationships, and their integration yielded the most robust models. These findings support the viability of a proxy-based monitoring framework. With further model refinement and cautious interpretation, this approach has the potential to extend data coverage and improve efficiency in data-limited river systems.

4.1 Evaluating Proxy Utility

4.1.1 Strong Predictive Potential

The integrated model revealed varying degrees of predictability across water quality variables, underscoring both the use and limitations of a proxy-based monitoring framework in the LMB. TSS emerged as the most reliably predicted variable, reflecting strong associations with upstream land use and downstream macroinvertebrate community composition. Landscape-scale drivers are important in shaping sediment dynamics, directly influencing movement and transport processes (Nelson & Booth, 2002). TSS responded strongly to spatial proxies, particularly urban land use. Urban environments often contribute elevated sediment loads through impervious surfaces, channel alterations, and runoff pathways (Russel, Vietz, & Fletcher, 2017; Cheng *et al.*, 2022).

Similarly, literature consistently shows that macroinvertebrate diversity metrics decline in response to increased sedimentation (Fierro *et al*., 2017; Kemp *et al*., 2011). Excess sediment can smother benthic habitats, impair respiratory function, and reduce food availability, weakening community structure and reducing ecological integrity (Jones *et al*., 2012).

4.1.2 Moderate Predictive Potential

Moderate predictive performance was observed for conductivity, COD, DO, and temperature. For conductivity and COD, model strength was primarily driven by urban and agricultural land use, which are widely associated with elevated concentrations of ions, organic pollutants, and chemical additives (Cheng *et al*., 2022). Urban areas contribute to elevated conductivity and COD through impervious surface runoff and sewage discharge, while agricultural zones introduce nutrient and organic loads via fertilizer application, pesticide runoff, and livestock waste (Cheng *et al*., 2022; Camara, Jamil, & Abdullah, 2019).

While more dynamic over time, DO and temperature showed associations with spatial and biological predictors. DO was moderately predicted by urban and forest land cover, reflecting two contrasting influences: urbanization often reduces DO through nutrient enrichment and increased biochemical oxygen demand, while forested areas tend to promote higher oxygen levels through retaining natural channels and stabilizing temperatures (Ice *et al*., 2021). This relationship has been documented in tropical systems, where riparian deforestation is linked to hypoxic conditions (Holguin-Gonzalez *et al*., 2013; Camara, Jamil, & Abdullah, 2019).

Temperature patterns were strongly predicted by macroinvertebrate metrics, particularly ATSPT and richness. These biological indicators declined in warmer conditions, aligning with literature that shows reduced macroinvertebrate diversity and sensitive taxa abundance at elevated temperatures, likely due to increased metabolic stress, lower oxygen availability, and altered food resources (Bonancia *et al*., 2022). The predictive role of ATSPT in temperature models supports its value as an integrative ecological indicator.

4.1.3 Weak Predictive Potential

In contrast, pH, TN, and TP could not be confidently predicted. This outcome is surprising given the basin's substantial agricultural expansion and well-documented eutrophication, especially within the Tonle Sap system (Yoshimura, Khanal, & Sovannara, 2022; Sor *et al*., 2021). The weak nutrient signal may be due to seasonal fertilizer application, potential lags between land use change and aquatic expression, and spatial resolution (Shen *et al*., 2020; Van Meter & Basu, 2015). Coarse agricultural proxies may inadequately capture fertilizer application, as literature suggests nutrient dynamics may need finer spatial resolution than sediment or conductivity-related processes (Wu, He, & Lu, 2025; Shen *et al*., 2020). This suggests the need for finer-scale agricultural metrics and temporally sensitive variables to improve predictive power.

4.2 Evaluating Model Performance

Despite statistically significant relationships between several predictors and water quality variables, the models often yielded moderate to low adjusted R^2 values. This apparent discrepancy reflects the difference between statistical significance and explanatory strength. A low R^2 indicates that the model accounts for only a limited portion of the variability in the response variable; however, it does not imply the absence of a relationship. Rather, it suggests that the relationship may be subtle, nonlinear, or obscured by noise and unmeasured factors. In contrast, low p-values signal that the observed effect is unlikely to have occurred by chance, even if the overall model fit is weak.

This nuance is particularly relevant in tropical river catchments, where water quality dynamics are shaped by indirect influences such as mixed land use and variable hydrological regimes. These complexities reduce the model's ability to explain variance yet allow individual predictors to retain statistically detectable influence. Additionally, some spatial proxies, like generalized land cover metrics, may align with WQ patterns in aggregate but fail to capture the finer spatial or temporal resolution needed for robust prediction. As a result, models may identify reliable signal but lack sufficient granularity for strong predictive power.

4.2.1 Model Validation

Validation testing revealed generally strong predictive performance for pH, DO, COD, TN, TP, and temperature. These results confirm that the spline structures in pooled GAMs were well-calibrated to landscape and ecological gradients. However, TSS and conductivity were less reliably predicted across validation sites.

The discrepancy in TSS predictions suggests that sediment loading may be highly localized or episodic, influenced by geomorphology, flood dynamics, or unmeasured tributary

contributions. Incorporating spatial stratification (e.g., river segment classification) or hydrological metrics may improve sediment modeling (Trang *et al*., 2017).

Conductivity likely reflects salinity inputs, groundwater seepage, or point-source pollution that are not well-captured by land use proxies or biological indicators. Additional geochemical parameters and urban infrastructure data could strengthen future conductivity models.

4.3 Suggestions for Model Refinement and Limitations

While initial results are promising, the models require further refinement to improve overall fit and predictive reliability. With additional data and methodological adjustments, future iterations are expected to yield stronger explanatory power. First, water quality data were restricted to dry season sampling, limiting the detection of monsoon-driven pulses or episodic pollution events. Expanding temporal coverage to include wet-season conditions would enhance ecological realism and model sensitivity.

Uncertainty was introduced through catchment delineation, buffer assignment, and site-pairing methodology. Model performance could be reassessed using alternative delineation strategies, such as reducing buffer sizes, excluding secondary catchments, or focusing on riparian zones, to better target localized or point-source pollution. Applying upstream distance weighting when pairing macroinvertebrate sites could refine spatial sensitivity and ecological relevance.

Finally, collinearity between predictors posed analytical challenges. Forest cover correlated strongly with other land use types, leading to its exclusion from final multivariate GAMs, despite ecological relevance. Similarly, ATSPT and richness demonstrated shared community structure. Variance partitioning confirmed their additive roles, but future models may benefit from including species-level traits or functional diversity to refine signal differentiation.

4.4 Implications for Monitoring and Management

The results suggest that proxy-based monitoring is feasible and effective in the LMB context with further model refinement. Urban and agricultural land use offer scalable indicators of pollutant loading, while macroinvertebrate metrics reflect ecological condition and resilience. Used together, these proxies provide a comprehensive framework for understanding aquatic health, prioritizing regulatory interventions, and evaluating the ecological impact of land use change.

For the Mekong River Commission and regional planners, integrating proxy models into routine monitoring programs could expand spatial coverage, reduce reliance on intensive sampling, and facilitate transboundary collaboration. Macroinvertebrate assessments and satellite-derived land use metrics are already available for many parts of the basin, and this study confirms their quantitative value for informing policy and resource allocation.

5. Conclusion

This study set out to evaluate whether land use and macroinvertebrate metrics can serve as effective proxies for surface water quality across the LMB. By integrating multivariate and nonlinear modeling approaches, including RDA, GAMs, and variance partitioning, this study characterized the contributions of landscape composition and biological communities to key water quality parameters.

The results affirm that land use, particularly urban and agricultural cover, is a consistent driver of water quality gradients, with strong effects on TSS, conductivity, COD and DO. Urban land use explained the greatest proportion of variance in pooled and yearly analyses, aligning with known aquatic stressors such as impervious surface runoff and sediment pollution. Agricultural land use also contributed meaningfully, though with greater temporal fluctuation. Forested cover provided comparatively weaker but interpretable signals, suggesting a potential buffering role in mitigating pollutant loads rather than functioning as a source of pollutants.

Macroinvertebrate metrics, especially ATSPT and richness, emerged as valuable biological predictors for parameters linked to organic pollution and nutrient enrichment. While their performance varied across years, their inclusion consistently improved model fit, supporting their role as partial proxies for water quality. Importantly, combined GAMs incorporating both land use and macroinvertebrates produced the highest explanatory power which demonstrates the value of integrated ecological modeling.

Variance partitioning confirmed that land use and biological indicators offer distinct insights, with minimal shared variance and unique contributions to water quality variation. Changes in predictor dominance over time highlight the importance of continued monitoring that captures seasonal variability and episodic events in order to ensure that indicator frameworks are responsive to these dynamics.

These findings highlight the value of proxy-based approaches for water quality assessment, particularly in regions where direct water quality monitoring is difficult. Land use metrics provide scalable, spatially continuous estimates of anthropogenic pressure, while macroinvertebrate indicators reflect ecological condition and resilience. To strengthen model insights and ecological interpretation, future research should incorporate additional spatial layers such as riparian land use, varied buffer sizes, and additional hydrological datasets. In the Lower Mekong Basin, the protection of water resources would benefit from increased use of GIS technologies within proxy-based and integrated modeling frameworks, enhancing environmental regulation through more efficient and scalable monitoring systems.

References

Azrina, M.Z. *et al.* (2006) 'Anthropogenic Impacts on the Distribution and Biodiversity of Benthic Macroinvertebrates and Water Quality of the Langat River, Peninsular Malaysia', *Ecotoxicology and Environmental Safety*, 64(3), pp. 337–347. doi: 10.1016/j.ecoenv.2005.04.003.

Beck, M.W. *et al.* (2022) 'Multi-Scale Trend Analysis of Water Quality Using Error Propagation of Generalized Additive Models', *Science of The Total Environment*, 802, p. 149927. doi: 10.1016/j.scitotenv.2021.149927.

Bignert, A. *et al.* (2014) 'Consequences of Using Pooled Versus Individual Samples for Designing Environmental Monitoring Sampling Strategies', *Chemosphere*, 94, pp. 177–182. doi: 10.1016/j.chemosphere.2013.09.096.

Błachuta, J. *et al.* (2014) 'How Do Environmental Parameters Relate to Macroinvertebrate Metrics? — Prospects for River Water Quality Assessment', *Polish Journal of Ecology*, 62(1), pp. 111–122. doi: 10.3161/104.062.0111.

Bonacina, L. *et al.* (2023) 'Effects of Water Temperature on Freshwater Macroinvertebrates: A Systematic Review', *Biological Reviews of the Cambridge Philosophical Society*, 98(1), pp. 191–221. doi: 10.1111/brv.12903.

Camara, M., Jamil, N.R. and Abdullah, A.F.B. (2019) 'Impact of Land Uses on Water Quality in Malaysia: A Review', *Ecological Processes*, 8(1), p. 10. doi: 10.1186/s13717-019-0164-x.

Cheng, C. *et al.* (2022) 'What is the Relationship Between Land Use and Surface Water Quality? A Review and Prospects from Remote Sensing Perspective', *Environmental Science and Pollution Research*, 29(38), pp. 56887–56907. doi: 10.1007/s11356-022-21348-x.

Chiarelli, D.D. *et al.* (2020) 'Hydrological Consequences of Natural Rubber Plantations in Southeast Asia', *Land Degradation & Development*, 31(15), pp. 2060–2073. doi: 10.1002/ldr.3591.

Corry-Roberts, A (2025) 'Assessing the Predictive Power of Land Use and Macroinvertebrates for Water Quality Assessment in the LMB, Part II Technical Report', pp. 1-103.

De Pauw, N., Gabriels, W., and Goethals, P. L. M. (2006) 'River Monitoring and Assessment Methods Based on Macroinvertebrates' in P. Quevauviller (ed.) *Water Quality Measurements*. Wiley, pp. 111–134. doi: 10.1002/0470863781.ch7.

Dickens, C. *et al.* (2018) *State of Knowledge: Monitoring the Health of the Greater Mekong's Rivers*. 9. Vientiane, Lao: CGIR Research Program on Water, Land and Ecosystems. Available at: https://www.academia.edu/89966000/Monitoring_the_health_of_the_greater_Mekong_s_rivers (Assessed 4 April 2025).

Duque, G. *et al.* (2022) 'Influence of Water Quality on the Macroinvertebrate Community in a Tropical Estuary (Buenaventura Bay)', *Integrated Environmental Assessment and Management*, 18(3), pp. 796–812. doi: 10.1002/ieam.4521.

Fierro, P. *et al.* (2017) 'Effects of Local Land-Use on Riparian Vegetation, Water Quality, and the Functional Organization of Macroinvertebrate Assemblages', *Science of The Total Environment*, 609, pp. 724–734. doi: 10.1016/j.scitotenv.2017.07.197.

Holguin-Gonzalez, J.E. *et al.* (2013) 'Development and Application of an Integrated Ecological Modelling Framework to Analyze the Impact of Wastewater Discharges on the Ecological Water Quality of Rivers', *Environmental Modelling & Software*, 48, pp. 27–36. doi: 10.1016/j.envsoft.2013.06.004.

Ice, G.G. *et al.* (2021) 'Understanding Dissolved Oxygen Concentrations in a Discontinuously Perennial Stream Within a Managed Forest', *Forest Ecology and Management*, 479, p. 118531. doi: 10.1016/j.foreco.2020.118531.

Iwasaki, Y., Suemori, T. and Kobayashi, Y. (2024) 'Predicting Macroinvertebrate Average Score Per Taxon (ATSPT) at Water Quality Monitoring Sites in Japanese Rivers', *Environmental Science and Pollution Research*, 31(19), pp. 28538–28548. doi: 10.1007/s11356-024-33053-y.

Jerves-Cobo, R. *et al.* (2020) 'Biological Water Quality in Tropical Rivers During Dry and Rainy Seasons: A Model-Based Analysis', *Ecological Indicators*, 108, p. 105769. doi: 10.1016/j.ecolind.2019.105769.

Jones, J.I. *et al.* (2012) 'The Impact of Fine Sediment on Macro-Invertebrates', *River Research and Applications*, 28(8), pp. 1055–1071. doi: 10.1002/rra.1516.

Kemp, P. *et al.* (2011) 'The Impacts of Fine Sediment on Riverine Fish', *Hydrological Processes*, 25(11), pp. 1800–1821. doi: 10.1002/hyp.7940.

Locke, K.A. (2024) 'Modelling Relationships Between Land Use and Water Quality Using Statistical Methods: A Critical and Applied Review', *Journal of Environmental Management*, 362, p. 121290. doi: 10.1016/j.jenvman.2024.121290.

Mekong River Commission (2010) *Biomonitoring Methods for the Lower Mekong Basin*. Vientiane, Lao: Mekong River Commission Secretariat. doi: 10.52107/mrc.ajhypg.

Mekong River Commission (2017). *The Council Study: Study on the Sustainable Management and Development of the Mekong River, Including Impacts of Mainstream Hydropower Projects. Biological Resource Assessment Final Technical Report Series*. Vientiane, Lao: Mekong River Commission Secretariat.

Mekong River Commission (2019) *State of the Basin Report 2018*. 1728–3248. Vientiane, Lao: Mekong River Commission Secretariat. Available at: https://www.mrcmekong.org/wp-content/uploads/2024/08/State-of-the-Basin-Report-2018-1.pdf (Accessed 20 December 2024).

Mekong River Commission (2021) *2021 Lower Mekong Water Quality Monitoring Report*. Vientiane, Lao: Mekong River Commission Secretariat. doi: 10.52107/mrc.c2xmzn.

Mekong River Commission. (2024). *Lower Mekong River Basin Atlas 2023*. Vientiane: MRC Secretariat. DOI: 10.52107/mrc.bjk3zl.

Murphy, R.R. *et al.* (2019) 'A Generalized Additive Model Approach to Evaluating Water Quality: Chesapeake Bay Case Study', *Environmental Modelling & Software*, 118, pp. 1–13. doi: 10.1016/j.envsoft.2019.03.027.

Nelson, E.J. and Booth, D.B. (2002) 'Sediment Sources in an Urbanizing, Mixed Land-Use Watershed', *Journal of Hydrology*, 264(1), pp. 51–68. doi: 10.1016/S0022-1694(02)00059-8.

Ongley, E.D. (2009) 'Water Quality of the Lower Mekong River' in I. Campbell (ed.) *The Mekong*. Amsterdam: Elsevier, pp. 297–320. doi: 10.1016/b978-0-12-374026-7.00003-6.

Pakoksung, K. *et al.* (2025) 'Seasonal Dynamics of Water Quality in Response to Land Use Changes in the Chi and Mun River Basins Thailand', *Scientific Reports*, 15(1), p. 7101. doi: 10.1038/s41598-025-91820-4.

Peres-Neto, P.R. *et al.* (2006) 'Variation Partitioning of Species Data Matrices: Estimation and Comparison of Fractions', *Ecology*, 87(10), pp. 2614–2625. doi: 10.1890/0012-9658(2006)87[2614:VPOSDM]2.0.CO;2.

Russell, K.L., Vietz, G.J. and Fletcher, T.D. (2017) 'Global Sediment Yields from Urban and Urbanizing Watersheds', *Earth-Science Reviews*, 168, pp. 73–80. doi: 10.1016/j.earscirev.2017.04.001.

SERVIR-SEA (2023) *Regional Land Cover Monitoring System (RLCMS)*. Version 1.0. SERVIR-Mekong, Asian Disaster Preparedness Center. Available at: https://servir.adpc.net/tools/rlcms_detail.html (Accessed: 16 April 2025).

Shen, L.Q. *et al.* (2020) 'Estimating Nitrogen and Phosphorus Concentrations in Streams and Rivers, Within a Machine Learning Framework', *Scientific Data*, 7(1), p. 161. doi: 10.1038/s41597-020-0478-7.

Sor, R. *et al.* (2017) 'Spatial Organization of Macroinvertebrate Assemblages in the Lower Mekong Basin', *Limnologica*, 64, pp. 20–30. doi: 10.1016/j.limno.2017.04.001.

Sor, R. *et al.* (2021) 'Water Quality Degradation in the Lower Mekong Basin', *Water*, 13(11), p. 1555. doi: 10.3390/w13111555.

Sripanya, J. *et al.* (2023) 'Benthic Macroinvertebrate Communities in Wadeable Rivers and Streams of Lao PDR as a Useful Tool for Biomonitoring Water Quality: A Multimetric Index Approach', *Water*, 15(4), pp. 625-651. doi: 10.3390/w15040625.

Tampo, L. *et al.* (2021) 'Benthic Macroinvertebrates as Ecological Indicators: Their Sensitivity to the Water Quality and Human Disturbances in a Tropical River', *Frontiers in Water*, 3, p. 662765. doi: 10.3389/frwa.2021.662765.

Trang, N.T.T. *et al.* (2017) 'Evaluating the Impacts of Climate and Land-Use Change on the Hydrology and Nutrient Yield in a Transboundary River Basin: A Case Study in the 3s River Basin (Sekong, Sesan, And Srepok)', *Science of The Total Environment*, 576, pp. 586–598. doi: 10.1016/j.scitotenv.2016.10.138.

Tromboni, F. *et al.* (2021) 'Changing Land Use and Population Density Are Degrading Water Quality in the Lower Mekong Basin', *Water*, 13(14), p. 1948. doi: 10.3390/w13141948.

Vadeboncoeur, Y., McIntyre, P.B. and Vander Zanden, M.J. (2011) 'Borders of Biodiversity: Life at the Edge of the World's Large Lakes', *BioScience*, 61(7), pp. 526–537. doi: 10.1525/bio.2011.61.7.7.

Van Meter, K.J. and Basu, N.B. (2015) 'Catchment Legacies and Time Lags: A Parsimonious Watershed Model to Predict the Effects of Legacy Storage on Nitrogen Export', *PLOS ONE*, 10(5), p. e0125971. doi: 10.1371/journal.pone.0125971.

Whitehead, P.G. *et al.* (2019) 'Water Quality Modelling of the Mekong River Basin: Climate Change and Socioeconomics Drive Flow and Nutrient Flux Changes to the Mekong Delta', *Science of The Total Environment*, 673, pp. 218–229. doi: 10.1016/j.scitotenv.2019.03.315.

Wu, J., He, S. and Lu, J. (2025) 'Multi-Scale Effects of Topography and Landscape Pattern on Riverine Nitrogen and Phosphorus Nutrients in an Agricultural Watershed', *Landscape Ecology*, 40(6), p. 112. doi: 10.1007/s10980-025-02131-y.

Yao, S. *et al.* (2023) 'Land Use as an Important Indicator for Water Quality Prediction in a Region Under Rapid Urbanization', *Ecological Indicators*, 146, p. 109768. doi: 10.1016/j.ecolind.2022.109768.

Yoshimura, C., Khanal, R. and Sovannara, U. (eds.) (2022) *Water and Life in Tonle Sap Lake*. Singapore: Springer Nature Singapore. doi: 10.1007/978-981-16-6632-2.

Zhu, J. *et al.* (2024) 'Multiple Scale Impacts of Land Use Intensity on Water Quality in the Chishui River Source Area', *Ecological Indicators*, 166, p. 112396. doi: 10.1016/j.ecolind.2024.112396.

Part II: Technical Report

Table of Contents

Table of Figures

Table of Tables

Abbreviations

ATSPT	Average Tolerance Score Per Taxon
BioRA	Biodiversity and Ecological Risk Assessment
COD	Chemical Oxygen Demand
DO	Dissolved Oxygen
EHM	Ecological Health Monitoring
GAM	Generalized Additive Model
LMB	Lower Mekong Basin
MLR	Multiple Linear Regression
MRC	Mekong River Commission
PCA	Principal Component Analysis
RDA	Redundancy Analysis
SDS	Site Disturbance Score
TN	Total Nitrogen
TP	Total Phosphorus
TSS	Total Suspended Sediment
WQM	Water Quality Monitoring
WFS	Web Feature Service

1. Introduction

This technical report supports the dissertation research titled *"*Evaluating the Proxy Potential of Land Use and Macroinvertebrates for Water Quality Assessment in the LMB*".* While the research paper presents core findings and interpretations through a scientific journal structure, the technical report provides clear documentation of the background information, analytical framework, data integration workflows, methodological decisions, and challenges encountered throughout the study. It serves as a complementary resource for readers seeking to replicate, critique, or extend the analysis for proxy modeling. The technical report has several key objectives:

- To characterize the LMB, fully establishing the need and validity of proxy monitoring.
- To explore external environmental drivers potentially influencing unexplained variance, such as riparian degradation, hydrological connectivity, and unregulated point sources.
- To full establish current monitoring efforts and conditions within the basin,
- To detail the processing of land cover, macroinvertebrate, and water quality datasets, highlighting the combination of MRC monitoring data and land use data.
- To describe the rationale and structures behind statistical methods employed, including collinearity diagnostics, exploratory analysis, and modeling structure and evaluation.
- To examine the sensitivity of proxy strength to buffer size and station typology, with a specific comparison between BioRA derived buffer zones and catchments.
- To discuss key limitations and adaptations made during data processing, sampling alignment, and model selection.

Beyond its immediate use, this report aims to support future applications of proxy-based monitoring across data-sparse river basins and contribute to ongoing research concerning the use of scalable proxy methods. It is especially relevant for regional planners, hydrologists, conservation scientists, and technical teams designing scalable frameworks under increasing anthropogenic pressure.

2. Literature Review

2.1 Justification of Water Quality Proxies

Macroinvertebrates occupy key functional niches in freshwater ecosystems, acting as shredders, grazers, and predators that contribute to energy flow and nutrient cycling (Dickens *et al*., 2018; Fierro *et al*., 2017; Sripanya *et al*., 2023; Tampo *et al*., 2021). Their community structure is shaped by a range of physicochemical water quality parameters, including DO, turbidity, temperature, pH, and nutrient concentrations. For instance, eutrophication driven by elevated TP and TN can increase primary productivity and oxygen demand, often resulting in hypoxic or anoxic conditions that suppress sensitive taxa (Hu *et al*., 2022). Low pH can impair shell formation and exoskeletal integrity, while elevated temperatures may exceed tolerance thresholds for many species. In contrast, tolerant taxa often thrive under degraded conditions, indicating pollution and habitat disturbance (Azrina *et al*., 2006; Fierro *et al*., 2017; Lammert & Allan, 1999).

These relationships between water quality and macroinvertebrate assemblages suggest that biological responses can be used in reverse to infer environmental conditions. In other words, the presence, absence, or dominance of particular taxa can serve as a proxy for underlying water quality dynamics. Sensitive groups are strongly associated with high DO, low nutrient loads, and stable pH, making them reliable indicators of ecological integrity. These relationships are especially pronounced during the dry season, when reduced flow and stable hydrological conditions allow macroinvertebrate communities to more clearly reflect underlying water quality dynamics (Holguin-Gonzalez *et al*., 2013). Shifts toward tolerant assemblages reflect increased sedimentation, nutrient enrichment, and organic pollution, conditions often linked to urbanization and agricultural land use (Dickens *et al*., 2018; Fierro *et al*., 2017).

Biological indicators integrate the cumulative effects of environmental stressors over time and space, which offers a more complete view of ecosystem health than snapshot physiochemical measurements (Azrina *et al*., 2006; Locke, 2024; Sripanya *et al*., 2023). Macroinvertebrates, along with other biological proxies such as fish and diatoms, have been widely adopted in biomonitoring programs across diverse geographic regions. Their cost-effectiveness, ecological relevance, and responsiveness to land use and hydrological disturbance make them particularly valuable in data-sparse or transboundary systems like the LMB (Dickens *et al*., 2018; Sripanya *et al*., 2023; Tampo *et al*., 2021).

Upstream land cover is a primary stressor influencing downstream water quality, particularly in systems where hydrological connectivity strengthens the effects of anthropogenic activities (Dickens *et al*., 2018; Fierro *et al*., 2017; Sor *et al*., 2017). Urbanization and agricultural expansion alter runoff dynamics, increase sedimentation, and elevate nutrient and pollutant loads (Azrina *et al*., 2006; Ongley, 2009; Whitehead *et al*., 2019). Empirical studies consistently link urban land cover to increased conductivity and impervious surface area, while agricultural land use correlates with elevated concentrations of TN, TP, and other nutrient-related indicators (Dickens *et al*., 2018; Tampo *et al*., 2021; Yao *et al*., 2023). pH, TSS, and DO also reflect anthropogenic disturbance (Fierro *et al*., 2017; Sripanya *et al*., 2023).

While urbanization effects tend to be localized near cities and do not dominate basin-wide water quality, landscape structure plays an important role in determining impacts (Ongley, 2009). Natural land connectivity and the presence of riparian buffers can intercept pollutants and mitigate degradation (Dickens *et al*., 2018; Sor *et al*., 2021; Sripanya *et al*., 2023; Yao *et al*., 2023). Riparian buffers directly influence streamside water quality, while catchment-scale land use represents broader sources of influence. Multi-scale buffer analyses enhance understanding of spatial land use impacts, with wider vegetated buffers generally associated with improved water quality outcomes (Locke, 2024; Sor *et al*., 2017; Tampo *et al*., 2021).

Given the complexity and scale of environmental systems direct measurement of all relevant variables is often impractical. Spatial proxies, derived from GIS and remotely sensed data, offer a valuable alternative (Chiarelli *et al*., 2020; Fierro *et al*., 2017; Sripanya *et al*., 2023). GIS-based analysis enables tracking of changing landscapes, especially in urbanizing regions, and supports

the development of predictive models linking landscape dynamics to water quality trends (Dickens *et al*., 2018; Pakoksung *et al*., 2025; Sor *et al*., 2017; Yao *et al*., 2023).

Spatial proxies complement biological indicators, providing non-redundant information: landscape context explains potential sources and pathways of pollutants, while macroinvertebrates integrate effects over time and reveal ecological integrity (Dickens *et al*., 2018; Locke, 2024; Tampo *et al*., 2021).

2.2 Study Area

The LMB encompasses a complex network of mainstem and tributary rivers that have different ecological and morphological characteristics. Mainstem reaches are usually deeper, more stable in flow, and composed of coarser substrates, while tributaries tend to be shallower, have more variable discharge, and dominated by finer sediments (Dickens *et al*., 2018; Sripanya *et al*., 2023). These physical differences shape habitat conditions and influence the distribution and sensitivity macroinvertebrate assemblages.

Hydrologically, the basin is characterized by a monsoon-driven regime that generates high suspended sediment loads during the wet season, which limits light penetration and suppresses algal growth (Whitehead *et al*., 2019; Ongley, 2009). The Mekong Delta and its canal systems are particularly vulnerable to eutrophication due to nutrient accumulation, while sediment redistribution from Tonle Sap Lake plays a critical role in downstream nutrient cycling and floodplain fertility (Gupta, 2009; Mekong River Commission, 2024). The basin's four physiographic regions, the Northern Highlands, Khorat Plateau, Tonle Sap Basin, and Delta, exhibit distinct topographic and hydrological profiles, which complicates ecological modeling.

The LMB's tributary network further complicates modeling efforts. Left-bank tributaries such as the Nam Ta, Nam Ou, Se Kong, Se San, and Sre Pok drain high-rainfall uplands and contribute disproportionately to wet-season flows and sediment transport. Right-bank tributaries like the Mun and Songkhram rivers drain lower-relief plateaus with more seasonal runoff. The transboundary 3S system (Sekong, Srepok, Sesan) plays a critical role in sediment and biotic exchange, exerting upstream influences on mainstem water quality (Trang *et al*., 2017).

Land use intensity varies widely across the basin, with urbanization, agriculture, and natural landscapes creating different pollution gradients (Mekong River Commission, 2023; Chiarelli *et al*., 2020). Increasing rubber plantations, expansion of cropland, and construction of hydropower dams (58 operational, 101 planned) have added significant stressors, including sediment trapping, flow alteration, and water scarcity (Mekong River Commission, 2022; Fan *et al*., 2015; Chiarelli *et al*., 2018). These cumulative impacts underscore the need for spatially explicit modeling frameworks that account for upstream-downstream linkages and land use dynamics.

These characteristics contextualize the results of the research paper and highlight the importance of incorporating additional variables in future modeling efforts. The relationships identified between land use, macroinvertebrate metrics, and water quality are promising, but

they represent only a subset of the basin's ecological complexity. Expanding monitoring to include wet season dynamics, underrepresented habitat types, and finer-scale land use classifications will enhance the predictive power and ecological relevance of proxy-based models.

2.2.1 BioRA Zonation

BioRA zones further divide the Mekong mainstream based on substrate composition, floodplain connectivity, flow regime variability, and species assemblage patterns (Mekong River Commission, 2017). Zones 1 through 8 correspond to combinations of physical drivers (gradient, flow variability) and ecological attributes (habitat, macroinvertebrate community structure). This subdivision allows for targeted biodiversity surveys and risk analyses that reflect natural environmental gradients and anthropogenic pressures (Mekong River Commission, 2017).

3. Data Overview

See Appendix A for data storage schema and file indexing.

3.1 Water Quality Monitoring

A continuous water quality monitoring program is managed by the MRC in order to track and manage physicochemical conditions across the basin (Mekong River Commission, 2021). In recognition of the importance of sustainable water development in the region, the program was established in 1985; in 2006 the routine assessment of water quality metrics began.

Forty-eight stations are used for water quality monitoring among the four countries within the LMB: 11 in Lao PDR, 8 in Thailand, 19 in Cambodia, and 10 in Viet Nam. Seventeen stations lie on the Mekong mainstem, and 31 are located on tributaries: 6 on the Tonle Sap system, 5 on the Bassac River, 5 on the 3S rivers, and 2 on the Nam Mun (Mekong River Commission, 2019). Monthly sampling occurs between the 13th and 18th of each month, using a standardized surface grab technique at mid-channel depths of 30–50 cm. Collected metrics include pH, electrical conductivity, TSS, nitrate-nitrite, ammonium, TP, TN, DO, COD, and BOD, and are assessed based on guidelines for Aquatic Life, Human Health, and Agricultural Health (Mekong River Commission, 2021). The temporal extent of the dataset allows for comparison of seasonal hydrological events and long-term water quality trends.

3.2 Ecological Health Monitoring

The MRC began ecological health monitoring in 2003 to establish baseline criteria and the contemporary program was formally outlined and established in 2010, as outlined in the Biomonitoring Handbook (Mekong River Commission, 2010). The distinct zones of the LMB have distinct substrate characteristics and responses to changes in flow regime; ecological health depends on both the quality of water and available habitats (Gupta, 2009; Mekong River Commission, 2018). Biennial sampling occurs at 41 sites across the basin's mainstem and tributaries.

Ecological health assessments sample littoral macroinvertebrates, benthic macroinvertebrates, zooplankton, and benthic diatoms (Mekong River Commission, 2010). Littoral macroinvertebrates live in dense aquatic vegetation and respond sensitively to shoreline modifications and nutrient inputs; they are collected along depositional riverbanks. Benthic macroinvertebrates inhabit deeper channel substrates, reflecting changes in flow regime, sediment composition, and pollution stress (Mekong River Commission, 2010, Fierro et al., 2017; Sripanya et al., 2023; Tampo et al., 2021). Monitoring these organisms helps detect pollution and habitat loss, thereby informing management actions to protect or restore river ecosystems (Azrina et al., 2006; Chiarelli et al., 2020; Sripanya et al., 2023).

A selection of physical and chemical variables (transparency, turbidity, temperature, dissolved oxygen, pH, conductivity) are recorded at the time of sampling to contextualize biological responses and support multivariate analyses of environmental drivers (Mekong River Commission, 2010).

A number of metrics are calculated and analyzed based on samples. Abundance is calculated based on the number of individuals per unit of area, volume or sample, as it reflects the number of individuals that belong to a certain indicator group (littoral macroinvertebrates, zooplankton, etc.) in a sample. Richness is determined by the mean number of taxa in an indicator group in a sample (Mekong River Commission, 2010). A SDS combines habitat condition data into a quantitative index of anthropogenic impact. EHM field teams complete a Substrate Characteristics Scoring Sheet, rating 12 descriptors of bank stability, substrate composition, and riparian vegetation. Through group discussion, observers assign each descriptor a score of 1 (minimal disturbance) to 3 (high disturbance) by comparing site characteristics against reference habitat descriptions (Mekong River Commission, 2010). SDS values are then used in determining ATSPT. ATSPT is the average tolerance of all taxa in a sample, without considering abundance (Mekong River Commission, 2010).

3.3 SERVIR-Mekong

The SERVIR-Mekong initiative is a joint effort with USAID, NASA and ADPC to provide mapping information and satellite imagery for Myanmar, Thailand, Cambodia, Laos, and Vietnam to support climate resilience. This is an open and free data source that is updated regularly. Raster data with a spatial resolution of 30 x 30 metres is available on a yearly basis from 2000 to 2023. SERVIR-Mekong used Google Earth Engine and also relies on field observations and the interpretation of high-resolution imagery by stakeholders to identify 18 land use classes (Table 1) (SERVIR-SEA, 2023). The success of the SERVIR-Mekong program has been expanded to the southeast Asia region to establish SERVIR-SEA in 2023.

Table 1: SERVIR-Mekong land use classification system.

Land Cover Typology	Pixel Value
aquaculture	1
barren	2
cropland	3
cropPlantation	4
deciduous	5
evergreen	6
floodedForest	7
forestPlantation	8
grass	9
mangrove	10
otherForest	11
palm	12
rice	13
rubber	14
shrub	15
urban	16
water	17
wetland	18
snow	19

4. Data Processing and Pre-Analysis Workflows

All data was processed in R version 4.4.2 (R Core Team 2024).

4.1 Spatial Data Collection

To prepare spatial layers for geospatial analysis, R was used to automate the retrieval and formatting of key datasets provided via the Mekong River Commission's WFS, available at: [https://geo.mrcmekong.org/geo/mrc/wfs] (Appendix B). A base WFS endpoint was specified and targeted six key layers: EHM monitoring stations, Bio-RA zones, Mekong mainstem and tributary networks, catchment polygons, and the LMB boundary. For each layer, the script generated a full WFS request URL, read the spatial features as GeoJSON, and reprojected them to UTM Zone 48N (EPSG:32648) for analytical consistency. To facilitate file management and future integration, the script sanitized filenames, created subdirectories per layer, and saved each dataset as a shapefile in an organized folder structure. Since the "Country Boundary" layer uses a unique naming convention, it was extracted separately, reprojected it to WGS84 (EPSG:4326), and dropped any Z or M geometry dimensions before exporting. Shapefiles were processed further to only contain features relevant to the study (Appendix C).

4.2 Land Use Data Processing

4.2.1 SERVIR-Mekong Data

Land cover rasters for six study years (2011, 2013, 2015, 2017, 2019, and 2021) were obtained from SERVIR-Mekong via the Regional Land Cover Monitoring System (RLCMS) portal [https://servir.adpc.net/tools/rlcms_detail.html]. Data were filtered by area and selected using the temporal slider tool provided on the interface (Figure 1). All files were downloaded in .tif raster format for spatial analysis. The SERVIR land cover dataset uses a 18-class schema, which was aggregated into broader categories based on their degree of naturalness and disturbance type (Table 2). Land use categories were used for statistical modeling and to simplify comparisons. Raster extents were first verified using the 'tmap' package to ensure spatial alignment with basin boundaries.

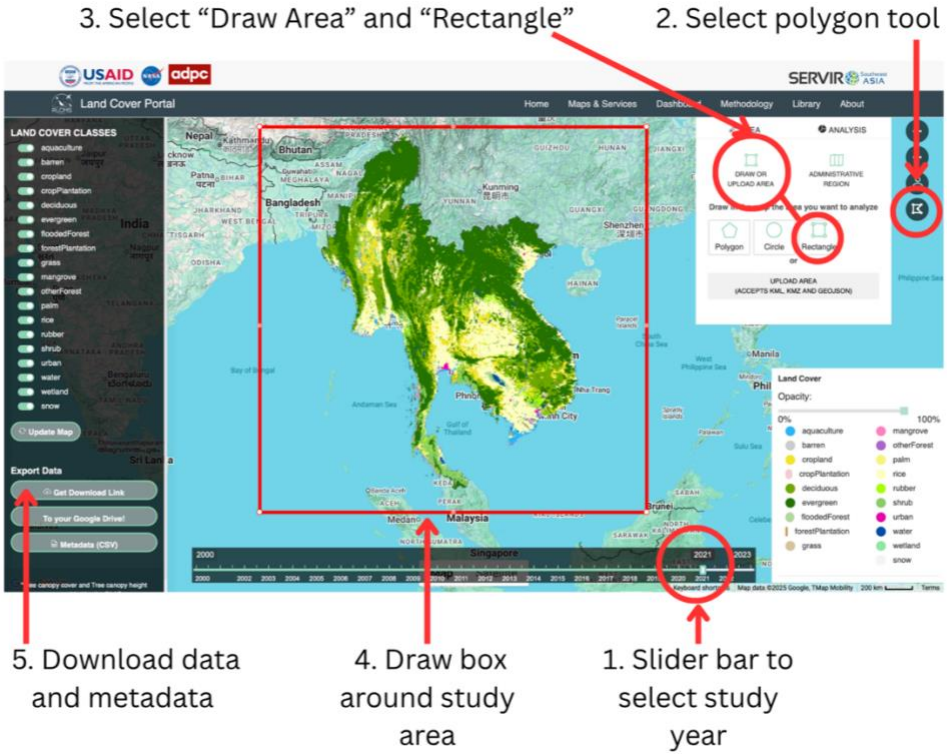

Figure 1: SERVIR-Mekong data download process.

Table 2: Land use classification groups with land cover types as defined by SERVIR-Mekong.

Category	Class Name	Class Number
Forest / Natural	Deciduous	5
	Evergreen	6
	Flooded forest	7
	Mangrove	10
	Other forest	11
	Wetland	18
Semi Natural	Forest plantation	8
	Grass	9
	Shrub	15
	Water	17
Agriculture	Aquaculture	1
	Cropland	3
	Crop plantation	4
	Palm	12
	Rice	13
	Rubber	14
Urban	Barren	2
	Urban	16

4.2.2 Land Use Area Selection

To account for upstream impacts of land use, land use in a boundary above each WQM site was defined. For tributary stations, the catchment the WQM station was located in was used as the land use boundary. If the catchment or bio-RA zone received flow from an upstream tributary in an additional catchment directly above the WQM site, land cover data from that connected sub-catchment were also included.

Because the Mekong mainstem flows along catchment boundaries rather than within catchments, catchments were not appropriate boundaries for mainstem stations. Instead, BioRA zones established by the MRC were used to segment the river into 8 sections, with a 15 km buffer applied on either side of each zone to reflect surrounding and upstream landscape influence (Appendix D). These zones provided ecologically meaningful divisions, reflecting transboundary influences, geomorphological boundaries, socio-economic gradients, and variation in aquatic habitats and biological assemblages (Mekong River Commission, 2017).

Final land use summaries were joined to their respective WQM station using site codes and station IDs to ensure alignment with other datasets (Appendix E). Both percent-based and absolute area land cover tables were generated for modeling.

Using both catchment-scale and buffer zone-based land use summaries allowed for a comparison of the impact of land use scale on proxy strength. Section 6.1 of this report examines whether smaller buffers yield more precise land use–water quality relationships or whether catchment-scale data are equally effective.

4.2.3 Land Use Data Processing

A loop using the 'terra', 'sf', and 'dplyr' packages reprojected each raster to EPSG:32648, maintaining the original 30 × 30 m cell resolution. For each year, the code cropped and masked the landcover raster to every land use (catchment or Bio-RA buffer) boundary, calculated a frequency distribution of pixel classes, and converted pixel counts into area using a consistent 30×30 meter resolution.

To validate results, total land cover area across all classes for select catchments was compared against the catchment boundary shapefile-derived catchment area and compared to confirm accuracy.

The resulting CSV files recorded pixel counts, area in square meters (calculated as count × cell size), WQM station identification, and year. Land use change was calculated for each monitoring station by year from 2011 to 2021 using results from both the catchment and BioRA zone processing workflows. Aggregated summaries were pivoted to wide format for percent based and absolute comparisons across time, and additional grouping by land cover type (natural, semi natural, agriculture, urban) supported analysis.

4.3 Water Quality Data Processing

Water quality data for this study was obtained directly from contacts at the MRC. Although the MRC has a public data portal for requesting available datasets available at [https://portal.mrcmekong.org/time-series/dashboard], this request was made through direct communication with Dr. Phan Nam Long, specifying the desired water quality metrics, study sites, and years. Data for WQM sites was requested only if they had an EHM site located downstream; stations without a corresponding EHM site were excluded from the request. Data from two requested stations, Ubon and Nam Kae, were not provided.

The water quality data was provided in CSV format, containing columns with study site descriptor and metric data (Table 3). The dataset was then reviewed and cleaned in R (Appendix F). Cleaning steps included filtering out records from non-study years, removing irrelevant stations, converting fields to proper numeric formats, and correcting station identifiers. A separate CSV file containing basic station metadata, including three-digit station codes, was joined to the water quality data.

Table 3: Column structure of WQM data provided by the MRC.

Column Name	Description of Data
StatID	Station Id beginning with H followed by 6 digits
Name	Name of the station
Country_Code	Either CA, LA, TH, or VN
Wtr_Body_Type	Either MS for mainstream of TB for tributary
River	Denotes the river the site monitors
Wtr_Body_Name	Denotes the name of the water body the site is on (generally the same as "River")
Sample_Date	Day of the week, Month Date, Year
Year_Collected	Four digit year
Month_Collected	Numeric month
TIDEHL	Tide height
FLOW_m3s	Discharge
TEMP_C	Temperature Celsius
pH	pH
TSS_mgl	Total suspended sediment
COND_mSm	Conductivity
TOTN_mgL	Total nitrogen
TOTP_mgL	Total phosphorus
DO_mgL	Dissolved oxygen
CODMN_mgL	Chemical oxygen demand
FC_MPN_mgL	Faecal coliform
BOD_mgL	Biological oxygen demand

Finally, the dataset was filtered by study year and restricted to dry season months. Dry season months (December, January, February, and March) were selected because ecological health monitoring in the LMB occurs during March. Filtering for months preceding ecological sampling ensures that macroinvertebrate responses reflect the water quality conditions they were exposed to in the lead-up to collection. Median values were calculated for each station-year to represent central tendency for water quality parameters, including temperature, pH, TSS, conductivity, TN, TP, DO, and COD. Water quality variables were visually inspected for outliers using boxplots and summary statistics. Extremely high readings for TSS and EC at select stations were flagged but retained, given their contextual relevance to land use impacts. Variables with high NA counts (above 60%) were excluded from the analysis (tide, flow, faecal coliform, and BOD).

4.4 Ecological Health Data Processing

EHM data was also obtained directly from a contact at the MRC, Dr. Kongmeng Ly. As with water quality data, ecological sites were selected only if they were located downstream of WQM stations; when EHM sites were situated upstream, data from those locations were not requested. The dataset was provided in Excel format, comprising multiple sheets, including SDS scores by station and separate sheets for abundance, richness, and ASPT values for littoral macroinvertebrates, benthic macroinvertebrates, zooplankton, and benthic diatoms. This study focused exclusively on littoral and benthic macroinvertebrates, as these groups are most

commonly used as bioindicators of freshwater condition (Tampo *et al*., 2021). Ecological health metrics were extracted and cleaned, where littoral and benthic macroinvertebrate records were combined for each site-year pair. Raw abundance and richness scores were summed, and ASPT values were calculated on a weighted average. Combined littoral and benthic macroinvertebrate metrics for richness, abundance, and ASPT were saved as separate CSV files.

Raw taxonomic-level data could not be obtained. Had such data been available, additional sensitivity metrics, such as EPT, could have been calculated and integrated into the analysis. Although EHM monitoring began in 2011, no data was collected for Cambodia during that first year. EHM data was available for 2011, 2013, 2015, 2017, 2019, and 2021. This determined the selected study years and aligned the selection of water quality and land use datasets accordingly. EHM data were reshaped from wide to long format to extract annual values for macroinvertebrate abundance, richness, and ASPT scores.

4.5 Joining Datasets

To prepare the final analysis dataset, water quality, ecological health, and land use data were merged and cleaned using 'tidyverse' packages (Appendix G). Site correspondence tables were used to pair EHM stations to upstream WQM sites and WQM sites to corresponding BioRA zone and or catchment(s). Records with complete data across all selected variables were retained for statistical modeling, and no imputation was applied.

5. Statistical and Modeling Methodologies

All data analysis was conducted in R version 4.4.2 (R Core Team 2024), with workflows built in various packages.

All tests were run at both pooled and yearly levels to enable comparison of long-term patterns with annual variability. This approach helped evaluate proxy reliability under changing landscape conditions and helped determine which ecological or spatial metrics consistently performed across time and at scale.

5.1 Exploratory Analysis

Exploratory analyses were conducted to examine overall trends and evaluate whether relationships among metrics aligned with those reported in existing literature.

5.1.1 Collinearity Diagnostics

To ensure robustness in the statistical modeling framework, land use groups and macroinvertebrate metrics variables were initially screened for collinearity (Appendix H). Highly correlated variables ($r > 0.7$) were flagged and reviewed for potential exclusion to prevent redundancy in models but were ultimately retained. Pairwise correlation matrices (Pearson's *r*) flagged relationships between agriculture and forest, urban and forest, and ATSPT and richness (Figure 2).

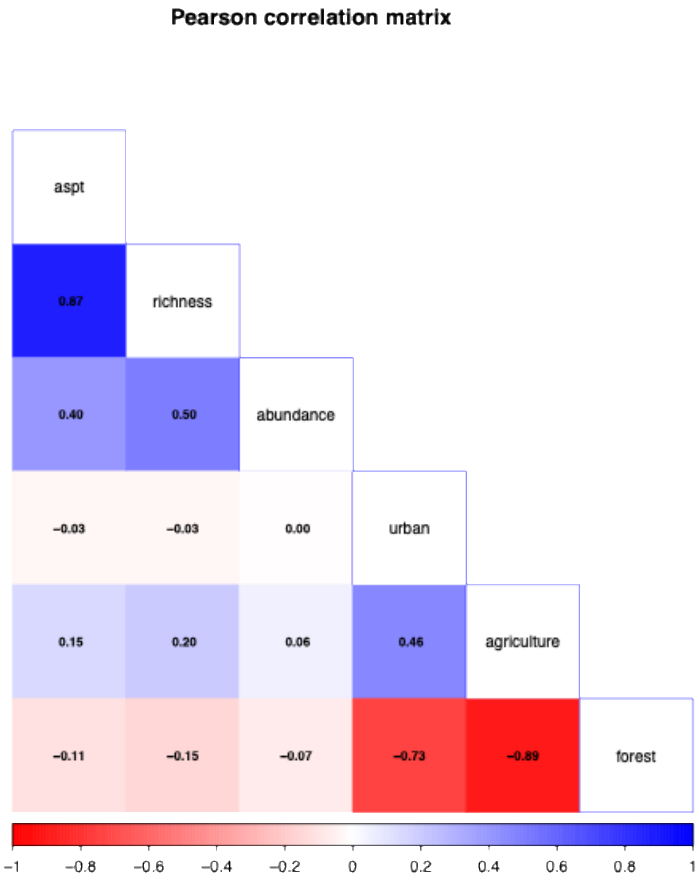

Figure 2: Pearson's correlation matrix of all predictor variables.

Richness displayed moderate correlation with ASPT. Year-specific GAMs revealed context-sensitive associations (e.g., significant temperature effects in 2011, 2015, and 2019) with richness as a predictor which reinforcing its proxy value. Excluding richness diminished the sensitivity of temporal models to ecological change.

Forest cover showed intermittent statistical significance across year-specific models and moderate correlation with both richness and urban. Two key factors supported its inclusion in models. In yearly GAMs, forest cover contributed significantly to DO and pH models in later years (2019, 2021). Additionally, exclusion of forest cover weakened variance explained in RDA models. Forest cover captures landscape resilience not reflected by urban or agricultural metrics, making it a critical component for characterizing ecological heterogeneity. Despite collinearity, its mechanistic role in nutrient cycling and habitat complexity justified its retention.

5.1.2 Exploratory Analysis

RDA conducted using the 'vegan' package to assess how much variation in water quality could be explained by land use and macroinvertebrates independently (Appendix I) (Locke, 2024; Sor *et al*., 2017; Tampo *et al*., 2021). Water quality variables were scaled using z-score

normalization (mean-centered and standardized) to allow variables with different units of measurement to be compared. Land use metrics expressed as percent cover per station-year for urban, agriculture, and forest/natural were used as predictors. RDA models were fitted, and adjusted R^2 values, F-statistics, and permutation-based p-values were extracted to assess overall model significance. Permutation tests (999 iterations) were used to assess statistical significance for the overall model and individual predictors, and results were exported to support spatial and temporal comparisons (Peres-Neto et al., 2006).

5.2 Modeling Methodologies

5.2.1 Multiple Linear Regression

MLR models were run to evaluate the strength of association between macroinvertebrate metrics and water quality parameters (Appendix J). Water quality values were not standardized, as water quality was the response variable, prior trials indicated that standardization did not materially affect regression outcomes, and interpretation was clearer using original units. Significance testing for annual models was conducted only when data availability met a minimum threshold necessary for statistical reliability; specifically, models were run only when at least 10 complete records were available per year to ensure stable estimates and interpretable results (Peres-Neto *et al*., 2006). Results from MLR models were not used in the results set in the Research Paper. While several predictors reached statistical significance, the overall explanatory power of the models was modest, and in many cases, weaker than GAM models. These findings underscore the importance of considering nonlinear modeling approaches when evaluating ecological responses, as macroinvertebrate metrics may exhibit threshold effects, saturation, or other nonlinear dynamics that are not well captured by linear regression.

5.2.2 Generalized Additive Models

GAMs were fitted using the 'mgcv' package to explore nonlinear relationships between predictor sets and water quality responses (Dickens *et al*., 2018; Locke, 2024; Sor *et al*., 2017; Tampo *et al*., 2021). GAMs were run on pooled data, by year, and for disaggregated station groups of mainstem and tributary segments using unscaled water quality values. Models were constructed for macroinvertebrates and water quality, land use and water quality, and both groups combined, allowing for the identification of independent effects and potential synergy between biological and spatial indicators (Appendix K). Pooled models tested macroinvertebrate metrics, land use categories, and combined predictors against individual water quality responses. Year-specific GAMs were run to evaluate temporal variability, using the same structure and minimum data thresholds to ensure model reliability.

Additional GAMs were developed for mainstem and tributary stations separately to compare landscape influence across river types and buffer size (Appendix K). Yearly mainstem and tributary models, were not run due to inadequate sample size. Fewer than 10 complete observations across water quality variables and predictors could result in overfitting and poor reliability.

For each model, adjusted R^2 values, deviance explained, and p-values for smoothed predictor terms were extracted to evaluate explanatory strength and nonlinear significance, and outputs were exported for comparative analysis. This workflow supported identification of strong, context-sensitive proxy relationships and helped clarify the distinct and complementary influences of biological and landscape variables on water quality.

Rather than plotting predictor metrics against water quality metrics to address relationships directly, GAMs were selected to model complexity overlooked by direct relationships. Direct relationships overlook whether trends are driven by predictors or if relationships are influenced by other predictors. In contrast, GAMs isolate the unique effect on water quality while controlling for other variables. This allows the analysis to delve further than establishing whether land use correlated to water quality and can determine how, where, and under what conditions effects emerge.

5.2.3 Variance Partitioning

Variance partitioning was conducted using the varpart() function in 'vegan' to estimate the unique and shared contributions of macroinvertebrate metrics and land use to variation in water quality (Appendix L) (Locke, 2024; Tampo *et al*., 2021). In pooled models, water quality data were standardized prior to partitioning to allow for unbiased comparison across variables of differing magnitude. Two predictor blocks were defined: one for land use and one for macroinvertebrates. Partial RDAs were used to isolate the independent contribution of each predictor set, and significance was tested using permutation-based ANOVA (999 iterations). This process was repeated for each study year from 2011 to 2021 using a custom R function that filtered complete cases and validated block structure. For each year, separate matrices were constructed, variance fractions were calculated, and significance testing was performed for both predictor sets. Outputs, including adjusted R^2 values for land use, macroinvertebrate metrics, shared variance, residual variance, and total explained variance, were saved as CSV files. This approach revealed temporal shifts in proxy strength and helped assess the independence and complementarity of biological and landscape indicators across dynamic environmental contexts.

5.2.4 Model Refinement and Validation

To further examine model strength in order to assess proxy potential, outlier sites that commonly fell outside the 95% confidence interval were assessed for each predictor GAM model (Appendix K). Outlier sites from the full model were further analyzed to determine the cause of sites not fitting in the model.

While integrated GAMs improved prediction accuracy at several stations, a subset of locations remained consistently outside modeled confidence intervals. Stations like LMX, CSP, TCS, VKB, and LPB still fell outside the confidence interval upwards for 70% of the time across water quality variables. This suggests that significant local drivers are still unaccounted for; unmeasured environmental drivers, misaligned land use summaries, or ecological thresholds are poorly reflected by proxy indicators. Full models struggled to account for local divergence in nutrient or oxygen-related metrics specifically.

To explore the model fit further, the outlier sites were stratified by zone, river system, and coordinates. These results showed that river systems with small catchments, high fragmentation, or intensive land use show the worst model performance; proxy relationships may be too coarse to capture this (Appendix M). Tributaries showed slightly higher residuals when compared to mainstem sites. High residuals were observed in latitude–longitude bins near 13.8–14.2°N, 107–107.8°E, with average outlier percentages reaching 75–85%. These areas correspond to eastern Cambodia and southern Laos, where tributaries such as the Sesan and Srepok flow through landscapes dominated by rubber plantations, agricultural areas, and increasing development.

Similarly elevated residuals (>70%) were detected in northern bins between 20–21.5°N, including sites along the Laos–Thailand border. These mountainous or forested zones often exhibit strong macroinvertebrate signals, yet land use predictors failed to account for water quality variation. Conversely, bins located around 10.8–10.9°N and 105.2–105.5°E, within the Mekong Delta, yielded lower residuals (27–46.9%), which in this context indicates the model's predictions are more closely aligned with observed water quality values in the bins.

Although broad generalizations, these findings highlight meaningful spatial patterns and suggest strong potential for deeper investigation into region-specific drivers of proxy performance.

6. Complete Results Set

Full results sets for results discussed in the Research Paper can be found in Appendix M – Q.

6.1 Proxy Strength Across Spatial Scales: Tributaries vs. Mainstem Stations

To evaluate how land use influences water quality at different spatial scales, separate GAMs were conducted for tributary and mainstem stations. Tributary WQM stations derived land use boundaries derived from their full upstream catchment area, while mainstem WQM stations derived land use boundaries constrained to a 15 km buffer around upstream BioRA zones. This distinction allowed for a direct comparison of proxy strength between broad-scale and more localized land use.

6.1.1 Mainstem Station GAM Results

Pooled models, aggregating data across all years, demonstrated variable explanatory power, with adjusted R^2 values ranging from 0.03 (TP) to 0.64 (temperature) (Appendix R). Temperature exhibited the strongest fit, with deviance explained reaching 69.2%. Urban land cover was the most consistent predictor, significantly associated with temperature ($p < 0.001$), dissolved oxygen ($p \approx 0.03$), and pH ($p < 0.001$). Conductivity and COD also yielded moderate fits (adjusted $R^2 = 0.53–0.55$), with both urban and forest cover showing statistically significant associations. In contrast, nutrient variables such as TN and TP were poorly explained by the models. Among biological indicators, macroinvertebrate metrics were inconsistently associated with water quality. ASPT was significantly linked to pH ($p = 0.001$), while richness and abundance exhibited limited significance across most pooled models.

6.1.2 Tributary Station GAM Results

The tributary model results underscore notable differences in the explanatory power of predictors when compared to mainstem systems, particularly in their utility as proxies for water quality (Appendix R). Overall, tributary responses appear more sensitive to both land cover composition and biological metrics, although predictor performance varies across parameters and years.

In the pooled tributary model, TSS stood out with the strongest model performance (Adj. R^2 = 0.89), driven predominantly by urban cover (p = 1e-5), agriculture (p = 0.001), and forest (p = 0.01). These land use predictors offer robust proxy potential for sediment loads. Conductivity also yielded a strong fit (Adj. R^2 = 0.71), and although the urban effect was significant (p < 0.001), forest cover and macroinvertebrate abundance were also notable. Temperature (Adj. R^2 = 0.49) was strongly predicted by all three land use types (all p < 0.001) and biological indices, especially ASPT and richness. In contrast, nutrient indicators such as TN and TP showed modest explanatory power (Adj. R^2 ≈ 0.34–0.48) and weak, inconsistent associations with land use, suggesting limited proxy utility in pooled models. Biological metrics were more relevant for TP, particularly richness (p = 0.03), highlighting a modest biological contribution to nutrient prediction. DO and COD models showed limited explanatory power (<0.45 Adj. R^2), and biological variables held minimal predictive significance in these contexts.

6.1.3 Comparison of Mainstem and Tributary Models

The comparison between the mainstem and tributary GAM results reveals distinct ecological dynamics and differing proxy reliability across catchment contexts.

Mainstem models generally exhibited robust fits for physical parameters such as temperature and dissolved oxygen, with adjusted R^2 values up to 0.64 in pooled analyses. Urban land cover emerged as the most consistent predictor across these responses, particularly for temperature, pH, and conductivity. However, biological indices, like ASPT and macroinvertebrate richness, played a more modest role, often lacking significance except in select cases. Nutrient parameters (TN and TP) were poorly explained in mainstem settings, which is unexpected given the smaller land use area; this contradicts literature suggesting that point source pollution is a key driver of agricultural nutrient loads, and that smaller buffers should therefore yield stronger explanatory power (Camara, Jamil, & Abdullah, 2019; Allafta & Opp, 2022).

Tributary models, in contrast, demonstrated stronger associations between both biological and land cover predictors and response variables. Pooled TSS and conductivity models achieved notably high adjusted R^2 values (0.89 and 0.71, respectively), and predictors from all three land cover categories were statistically significant. Biological metrics also showed enhanced explanatory power, especially ASPT and richness in temperature and TP models.

In the integrated model from the Research Paper, model fit was generally lower across most response variables compared to the tributary-only model and slightly reduced relative to the mainstem-only model (Figure 3). For example, temperature in the integrated model had an adjusted R^2 of 0.36, explaining ~39.6% of deviance; substantially lower than both tributary (Adj.

R^2 = 0.49) and mainstem (Adj. R^2 = 0.64) fits. Despite this, ASPT and richness were highly significant in the integrated model ($p < 1e-4$), indicating continued proxy potential from biological indicators.

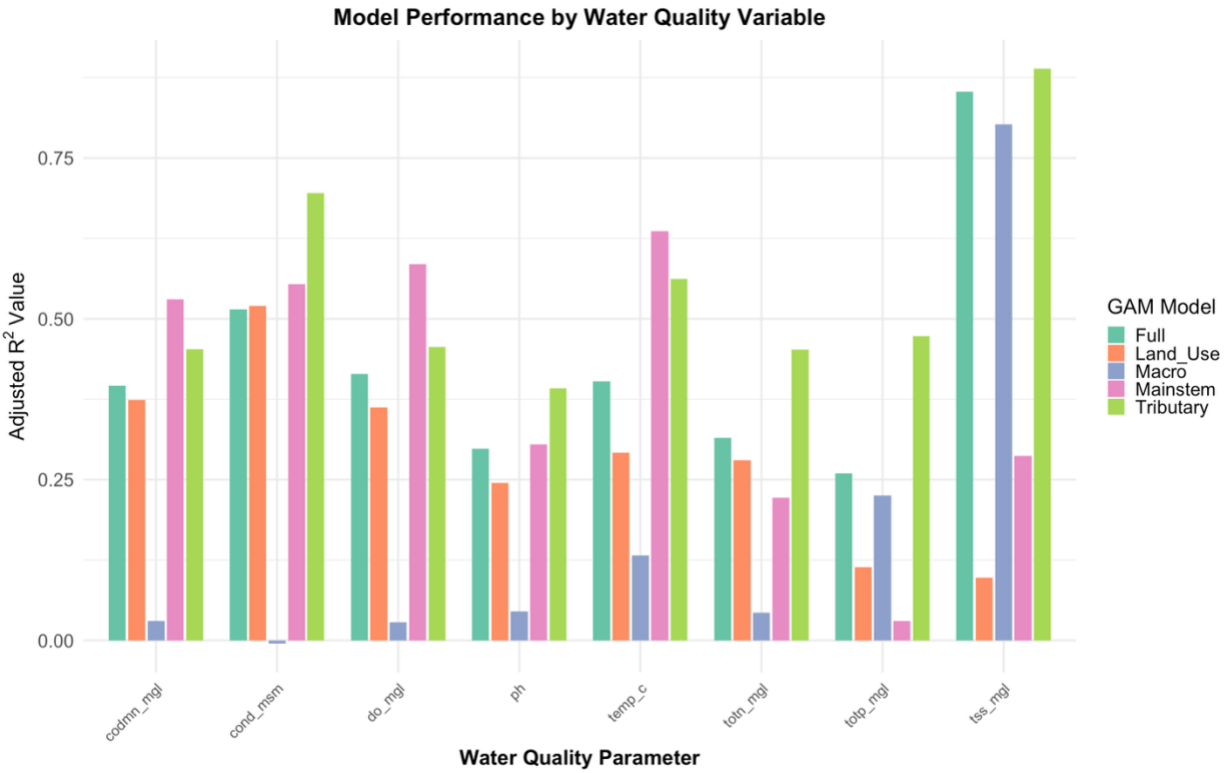

Figure 3: Adjusted R^2 values for five GAMs across eight water quality parameters, highlighting variation in explanatory power by ecological and spatial inputs.

TSS showed strong proxy relationships in the integrated model (Adj. R^2 = 0.85), consistent with tributary results and far stronger than the mainstem model (Adj. R^2 = 0.29). This suggests that sediment-associated responses retain influence across spatial scales.

In summary, stratified modeling yields richer ecological insights and stronger proxy reliability. Tributary models consistently outperformed the integrated model and, in many cases, the mainstem model, especially for biological and sediment-related parameters. However, it was surprising that the mainstem model did not exhibit stronger performance, particularly for nutrient parameters, given the smaller land use areas delineated by the BioRA Zones. Literature suggests that point source pollution and riparian zone dynamics should enhance model fit in such contexts, especially for agricultural and urban pollutants (Ongley, 2009; Allafta & Opp, 2022). Nutrient parameters remained weakly predicted across all models, suggesting that these variables are subject to effects not well captured by current land use areas and that the current buffer scale is still too coarse to capture localized impacts. A more refined approach that uses smaller buffers or focuses specifically on land use immediately adjacent to monitoring sites may improve sensitivity to point source influences and better reflect fine-scale ecological pressures.

6.2 Comparative Model Diagnostics

To validate model structure and support reproducibility, a full suite of diagnostic checks was conducted across full GAM framework (Appendix K).

6.2.1 Residual Analysis

Residual plots were inspected for homoscedasticity, normality, and spatial independence. GAM residuals showed acceptable fit across most parameters, with slight clustering in TSS and pH models (Figure 4). Despite its strong adjusted R^2, the TSS model exhibited signs of heteroscedasticity, with increasing residual spread at higher fitted values, suggesting variance inflation or influential events. The pH model showed layered residual patterns, hinting at underlying grouping effects or unmodeled spatial structure.

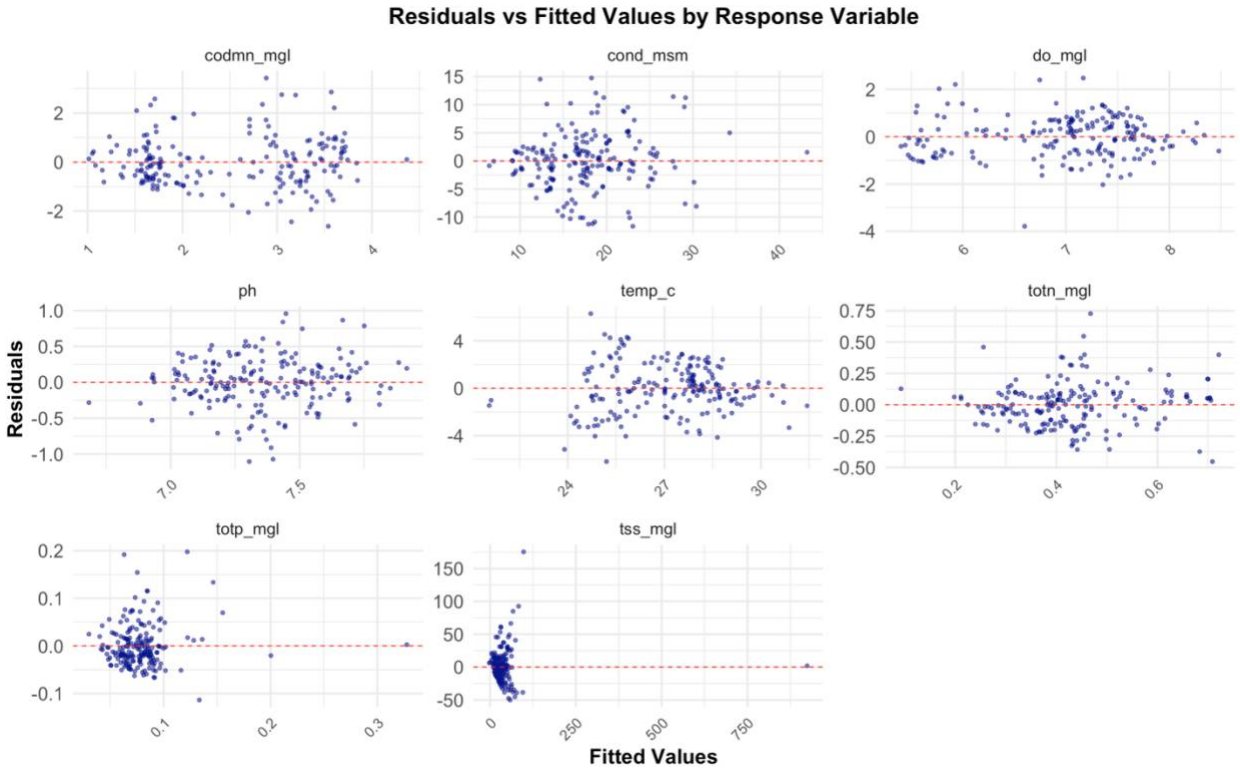

Figure 4: Residuals versus fitted values for the full GAM across eight water quality parameters.

6.3 Temporal Dynamics and Ecological Lag

The strength of proxy relationships fluctuated across sample years, reinforcing the importance of verifying proxy relationships across multi-year datasets. These temporal trends likely reflect both landscape transition and ecological lag effects. For instance, periods of rapid agricultural expansion (2011–2015) may have introduced nutrient loads that manifested biologically in subsequent years. Macroinvertebrate richness and ASPT rose in later years, suggesting community shifts toward disturbance-tolerant taxa, even as taxonomic diversity increased.

These lagged responses mirror findings in other tropical catchments, where biological indicators respond cumulatively to land use stressors (Jerves-Cobo *et al*., 2020; Yao *et al*., 2023; Dickens *et al*., 2018).

Year-specific variability also demonstrates the ecological complexity of proxy relationships. In 2015, TP and DO models performed strongly, supported by elevated biological signal. In 2019 and 2021, richness gained relevance as a predictor, indicating its evolving value as a proxy for nutrient dynamics and oxygen availability.

7. Technical Challenges and Adaptations

7.1 MRC Data Challenges

Numerous technical challenges emerged during the integration and analysis of ecological health, land use, and water quality datasets for the LMB. These difficulties stemmed from both limitations in data structure and resolution, as well as broader issues of standardization across spatial datasets and monitoring protocols. While these challenges presented barriers to analysis, adaptations were implemented to improve data consistency, model reliability, and reproducibility across workflows.

The primary challenge during initial data screening and analysis was the inconsistency of station identifiers provided by the MRC. The water quality metric dataset included a mix of six-digit station IDs and station names, while the site metadata (provided in a separate CSV) used three-letter identifiers and station names but lacked station IDs.

Compounding this issue, some stations listed in the water quality dataset were missing from the metadata file, which was intended to represent a complete inventory of WQM sites. Additionally, several station IDs in the metric dataset were incorrectly entered and required manual review and correction (see Appendix B).

Naming conventions across datasets were inconsistent due to misspellings, input errors, and language differences, making it difficult to reliably match metadata to water quality and macroinvertebrate metrics. Without manual verification, these discrepancies could have led to misalignment across merged datasets and introduced errors in linking upstream land use to water chemistry and biological indicators.

To resolve these issues, metadata from MRC documents was cross-referenced with field reports and shapefile coordinates. Name mismatches were manually corrected where necessary, and each WQM site included in the final analysis was mapped against its downstream EHM counterpart to confirm hydrological connectivity. In cases where naming conflicts persisted, primary identifiers (such as StatID and geographic coordinates) were prioritized over descriptive labels. This approach minimized misidentification and improved confidence in the integrity of model inputs.

Incompleteness across key datasets also presented constraints. Despite specific data requests to the MRC, water quality records for several desired stations, including Ubon and Nam Kae, were not provided. Similarly, ecological monitoring data were absent for Cambodia in 2011, and raw taxonomic-level data for macroinvertebrates were unavailable, precluding the calculation of sensitivity metrics such as EPT richness or functional trait indices. This limitation narrowed the analytic scope, as available biological metrics were restricted to abundance, richness, and ASPT scores.

Several water quality metrics (e.g., faecal coliform, flow, and tide height) contained extensive missing values (>60%) and were excluded from model inputs to avoid bias and reduce noise. Rather than attempting to impute missing values and potentially introduce artificial structure, statistical models were used only complete data. Minimum row count thresholds ($n \geq 10$) were enforced for year-specific GAMs to ensure stability of smoothed terms and avoid overfitting. When sample sizes fell below this threshold, models were not run.

7.2 Land Use Data Challenges

The integration of spatial raster data posed another set of difficulties. SERVIR-Mekong's Regional Land Cover Monitoring System provides valuable land use data, but full-basin coverage for six study years generated large raster files. Processing these layers required substantial computational resources and led to repeated memory allocation failures in R when attempting to convert raster data to vector format and clip data to catchment or BioRA extents. To address this, data was retained in raster format and subsequent workflows were run using the University of Edinburgh's high-performance computing (HPC) cluster, which allowed efficient reprojection, buffering, and summary extraction without bottlenecks. Intermediary outputs from raster clipping and masking were exported to reduce processing power in subsequent analysis.

7.3 Modeling Challenges

The selection and structure of statistical models also involved key trade-offs. GAMs were selected over multiple linear regression and other parametric approaches due to their flexibility in capturing nonlinear relationships typical of ecological datasets. To preserve interpretability and minimize overparameterization, models were structured with smoothed terms for macroinvertebrate and land use predictors, while keeping response water quality variables untransformed.

Finally, confidence interval analysis showed that several stations consistently fell outside the predicted range, even in the full model. Residual mapping identified these outliers as important flags for future refinement. Their misfit was interpreted not as model failure, but as evidence of missing environmental drivers that fall outside the scope of land use and macroinvertebrate proxies. These stations may require alternative buffer designs, additional covariates, or inclusion of qualitative disturbance scores in future iterations.

Taken together, these challenges highlight the complexities of transboundary ecological modeling using publicly available datasets. While limitations in data structure, completeness, and computational scale were substantial, the workflow developed here provides a resilient and reproducible approach to bringing together different datasets. The technical adaptations described enabled water quality proxy modeling and helped ensure that findings reflect environmental processes rather than methodological biases. Therefore, they support both the integrity of the current analysis, and the scalability of proxy-based frameworks for regional freshwater monitoring.

8. Summary and Cross Reference to Research Paper

This technical report serves as a companion to the dissertation "Evaluating the Proxy Potential of Land Use and Macroinvertebrates for Water Quality Assessment in the LMB". It expands upon the methodological foundations, data integration strategies, and analytical decisions that underpin the primary research findings. Specifically, it documents the spatial workflows, statistical frameworks, and scale-dependent proxy evaluations used to assess the reliability of land use and biological indicators in predicting water quality across the LMB.

The report provides a detailed examination of how land cover characteristics and macroinvertebrate metrics predict key water quality parameters. By analyzing the full statistical methodology, it reinforces the central finding that proxy strength is both context-specific and scale-dependent. Notably, catchment-derived land use summaries around tributary stations produced significantly stronger predictive signals than buffer-based summaries used for mainstem stations, highlighting the importance of spatial boundary selection in proxy modeling.

Extensive data processing steps are outlined, including land use raster extraction via SERVIR-Mekong and the integration of ecological and water quality datasets sourced from the MRC. Technical challenges such as inconsistent station naming, missing values, and technical demands associated with raster manipulation are transparently addressed, supporting reproducibility and methodological rigor.

These technical results validate the integrated modeling framework and provide a grounded rationale for the analytical choices made throughout the broader study. By complementing the core research paper with deeper methodological transparency, this report enhances the reproducibility, interpretability, and applicability of proxy-based water quality assessments across diverse spatial and temporal contexts.

References

Allafta, H. and Opp, C. (2022) 'Understanding the Combined Effects of Land Cover, Precipitation and Catchment Size on Nitrogen and Discharge—A Case Study of the Mississippi River Basin', *Water*, 14(6), pp. 865-881. doi: 10.3390/w14060865.

Azrina, M.Z. *et al.* (2006) 'Anthropogenic Impacts on the Distribution and Biodiversity of Benthic Macroinvertebrates and Water Quality of the Langat River, Peninsular Malaysia', *Ecotoxicology and Environmental Safety*, 64(3), pp. 337–347. doi: 10.1016/j.ecoenv.2005.04.003.

Camara, M., Jamil, N.R. and Abdullah, A.F.B. (2019) 'Impact of Land Uses on Water Quality in Malaysia: A Review', *Ecological Processes*, 8(1), p. 10. doi: 10.1186/s13717-019-0164-x.

Chiarelli, D.D. *et al.* (2018) 'The Water-Land-Food Nexus of Natural Rubber Production', *Journal of Cleaner Production*, 172, pp. 1739–1747. doi: 10.1016/j.jclepro.2017.12.021.

Chiarelli, D.D. *et al.* (2020) 'Hydrological Consequences of Natural Rubber Plantations in Southeast Asia', *Land Degradation & Development*, 31(15), pp. 2060–2073. doi: 10.1002/ldr.3591.

Dickens, C. *et al.* (2018) *State of Knowledge: Monitoring the Health of the Greater Mekong's Rivers*. 9. Vientiane, Lao: CGIR Research Program on Water, Land and Ecosystems. Available at: https://www.academia.edu/89966000/Monitoring_the_health_of_the_greater_Mekong_s_rivers (Assessed 4 April 2025).

Fan, H., He, D. and Wang, H. (2015) 'Environmental Consequences of Damming the Mainstream Lancang-Mekong River: A Review', *Earth-Science Reviews*, 146, pp. 77–91. doi: 10.1016/j.earscirev.2015.03.007.

Fierro, P. *et al.* (2017) 'Effects of Local Land-Use on Riparian Vegetation, Water Quality, and the Functional Organization of Macroinvertebrate Assemblages', *Science of The Total Environment*, 609, pp. 724–734. doi: 10.1016/j.scitotenv.2017.07.197.

Gupta, A. (2009) 'Geology and Landforms of the Mekong Basin' in I. Campbell (ed.) *The Mekong*. Amsterdam: Elsevier, pp. 29–51. doi: 10.1016/b978-0-12-374026-7.00003-6.

Holguin-Gonzalez, J.E. *et al.* (2013) 'Development and Application of an Integrated Ecological Modelling Framework to Analyze the Impact of Wastewater Discharges on the Ecological Water Quality of Rivers', *Environmental Modelling & Software*, 48, pp. 27–36. doi: 10.1016/j.envsoft.2013.06.004.

Hu, X. *et al.* (2022) 'Response of Macroinvertebrate Community to Water Quality Factors and Aquatic Ecosystem Health Assessment in a Typical River in Beijing, China', *Environmental Research*, 212, p. 113474. doi: 10.1016/j.envres.2022.113474.

Jerves-Cobo, R. *et al.* (2020) 'Biological Water Quality in Tropical Rivers During Dry and Rainy Seasons: A Model-Based Analysis', *Ecological Indicators*, 108, p. 105769. doi: 10.1016/j.ecolind.2019.105769.

Lammert, M. and Allan, J.D. (1999) 'Assessing Biotic Integrity of Streams: Effects of Scale in Measuring the Influence of Land Use/Cover and Habitat Structure on Fish and Macroinvertebrates', *Environmental Management*, 23(2), pp. 257–270. doi: 10.1007/s002679900184.

Locke, K.A. (2024) 'Modelling Relationships Between Land Use and Water Quality Using Statistical Methods: A Critical and Applied Review', *Journal of Environmental Management*, 362, p. 121290. doi: 10.1016/j.jenvman.2024.121290.

Mekong River Commission (2010) *Biomonitoring Methods for the Lower Mekong Basin*. Vientiane, Lao: Mekong River Commission Secretariat. doi: 10.52107/mrc.ajhypg.

Mekong River Commission (2017) *The Council Study: Study on the Sustainable Management and Development of the Mekong River, Including Impacts of Mainstream Hydropower Projects. Biological Resource Assessment Final Technical Report Series*. Vientiane, Lao: Mekong River Commission Secretariat.

Mekong River Comission. (2018) *Report on the 2017 Biomonitorng Survey of the Lower Mekong River and Selected Tributaries*. Vientienne, Lao: MRC Secretariat. doi: 10.52107/mrc.ajg4wd.

Mekong River Commission (2019) *State of the Basin Report 2018*. 1728–3248. Vientiane, Lao: Mekong River Commission Secretariat. Available at: https://www.mrcmekong.org/wp-content/uploads/2024/08/State-of-the-Basin-Report-2018-1.pdf (Accessed 20 December 2024).

Mekong River Commission (2021) *2021 Lower Mekong Water Quality Monitoring Report*. Vientiane, Lao: Mekong River Commission Secretariat. doi: 10.52107/mrc.c2xmzn.

Mekong River Comission. (2022) *Joint Environmental Monitoring Programme at Two Mekong Mainstream Dams: The Don Sahong and Xayaburi Hydropower Projects.* Vientienne: MRC Secretariat. doi: 10.52107/mrc.aqrs7o.

Mekong River Commission (2023) *Enhancing the MRC Land Use and Land Cover 2020 Mapping Products.* Vientaine, Lao: Mekong River Commission Secretariat. doi: 10.52107/mrc.aqrsbr.

Ongley, E.D. (2009) 'Water Quality of the Lower Mekong River' in I. Campbell (ed.) *The Mekong*. Amsterdam: Elsevier, pp. 297–320. doi: 10.1016/b978-0-12-374026-7.00003-6.

Pakoksung, K. *et al.* (2025) 'Seasonal Dynamics of Water Quality in Response to Land Use Changes in the Chi and Mun River Basins Thailand', *Scientific Reports*, 15(1), p. 7101. doi: 10.1038/s41598-025-91820-4.

Peres-Neto, P.R. *et al.* (2006) 'Variation Partitioning of Species Data Matrices: Estimation and Comparison of Fractions', *Ecology*, 87(10), pp. 2614–2625. doi: 10.1890/0012-9658(2006)87[2614:VPOSDM]2.0.CO;2.

SERVIR-SEA (2023) *Regional Land Cover Monitoring System (RLCMS)*. Version 1.0. SERVIR-Mekong, Asian Disaster Preparedness Center. Available at: https://servir.adpc.net/tools/rlcms_detail.html (Accessed: 16 April 2025).

Sor, R. *et al.* (2017) 'Spatial Organization of Macroinvertebrate Assemblages in the Lower Mekong Basin', *Limnologica*, 64, pp. 20–30. doi: 10.1016/j.limno.2017.04.001.

Sor, R. *et al.* (2021) 'Water Quality Degradation in the Lower Mekong Basin', *Water*, 13(11), p. 1555. doi: 10.3390/w13111555.

Sripanya, J. *et al.* (2023) 'Benthic Macroinvertebrate Communities in Wadeable Rivers and Streams of Lao PDR as a Useful Tool for Biomonitoring Water Quality: A Multimetric Index Approach', *Water*, 15(4), pp. 625-651. doi: 10.3390/w15040625.

Tampo, L. *et al.* (2021) 'Benthic Macroinvertebrates as Ecological Indicators: Their Sensitivity to the Water Quality and Human Disturbances in a Tropical River', *Frontiers in Water*, 3, p. 662765. doi: 10.3389/frwa.2021.662765.

Trang, N.T.T. *et al.* (2017) 'Evaluating the Impacts of Climate and Land-Use Change on the Hydrology and Nutrient Yield in a Transboundary River Basin: A Case Study in the 3s River Basin (Sekong, Sesan, And Srepok)', *Science of The Total Environment*, 576, pp. 586–598. doi: 10.1016/j.scitotenv.2016.10.138.

Whitehead, P.G. *et al.* (2019) 'Water Quality Modelling of the Mekong River Basin: Climate Change and Socioeconomics Drive Flow and Nutrient Flux Changes to the Mekong Delta', *Science of The Total Environment*, 673, pp. 218–229. doi: 10.1016/j.scitotenv.2019.03.315.

Yao, S. *et al.* (2023) 'Land Use as an Important Indicator for Water Quality Prediction in a Region Under Rapid Urbanization', *Ecological Indicators*, 146, p. 109768. doi: 10.1016/j.ecolind.2022.109768.

Appendices

Appendix A. File Index

Numbers denote file paths; letters denote file names.

1. dissfinal/code
    a. appendix f wq clean.R
        i. Cleans the water quality data provided by the MRC
        ii. 3 KB
    b. appendix b wfs extract.R
        i. Extracts shapefiles from the MRC's WFS
        ii. 4 KB
    c. appendix c select shp.R
        i. Selects relevant features from shapefiles and exports new .shp
        ii. 6 KB
    d. appendix d lu processing.R
        i. Determines land cover in catchments/buffers from SERVIR rasters
        ii. 7 KB
    e. appendix e lu metrics.R
        i. Calculates land use metrics in relevant areas for each WQM site
        ii. 12 KB
    f. appendix g stats.R
        i. Set up for subsequent statistical analysis code; cleans and joins datasets
        ii. 6 KB
    g. appendix h pearsons.R
        i. Runs Pearsons Correlation and makes figure
        ii. 2 KB
    h. appendix j mlr.R
        i. Runs MLRs with macro metrics predicting WQ
        ii. 3 KB
    i. appendix i rda.R
        i. Runs RDAs for macro and land use predictors
        ii. 9 KB
    j. appendix k macro gam.R
        i. Runs GAMs with macro metrics predicting WQ
        ii. 13 KB
    k. appendix k lu gam.R
        i. Runs GAMs with land use predicting WQ
        ii. 13 KB
    l. appendix k full gam.R
        i. Runs GAMs with macro metrics and land use predicting WQ
        ii. 12 KB
    m. appendix l vp.R
        i. Variance partitioning on full GAM

        ii. 6 KB
    n. appendix k mvt gam.R
        i. Runs full GAMs on mainstem and tributary sites separately
        ii. 7 KB
    o. figures.R
        i. Makes figures that are not in statistical analysis codes
        ii. 40 KB
2. dissfinal/ecological_health_data
    a. abundance.CSV, aspt.CSV, richness.CSV
        i. "code"- three letter station identifiers
        ii. "sf_name" – name of the site that matches the shapefile station name; added manually for joining
        iii. "*year*" – combined lm and bm abundance values per year
        iv. 3 KB
        v. Adapted from MRC EH dataset
    b. eh_sds.CSV
        i. "code"- three letter station identifiers
        ii. "sf_name" – name of the site that matches the shapefile station name; added manually for joining
        iii. "*year*" – site disturbance score per year
        iv. 4 KB
        v. MRC EH dataset
    c. eh_sites.CSV
        i. "no" – arbitrary station number
        ii. "code" – three letter station identifier
        iii. "sf_name" – name of the site that matches the shapefile station name; added manually for joining
        iv. "monitoring_site" – original site name in metadata
        v. "river" – river name of site
        vi. "countries" – country site is in
        vii. "latitiude_ESPG_3148" – latitude
        viii. "longitude_ESPG_3148" – longitude
        ix. 3 KB
        x. MRC EH metadata
    d. MRC DATA.XLSX
        i. EHM dataset with separate sheets for sites, SDS, and abundance, richness and ATSPT for benthic macros, littoral macros, zooplankton, and benthic diatoms
        ii. 69 KB
        iii. Original MRC EHM dataset
2. dissfinal/figures
    a. Multiple PNG/PDF files
        i. 884 KB – 8 KB

ii. Various figures created in R; figures relating to statistical analysis were created in the same file as their respective test/model, while maps and general figures were created in “appendix o figures.R”

3. dissfinal/join_metadata
    a. wq_catchments_for_landcover.CSV
        i. “code” - three letter station identifiers
        ii. “statid” – H+7 digit identifier
        iii. “station_name” – name of the WQM
        iv. “river_name_og” – name of the river as in WQ dataset
        v. “river_name_sf” – name of the river as in WQ shapefile
        vi. “country” – country WQM falls in
        vii. “catchment” – manually derived from examining which catchment the WQM falls in or if mainstem station may be a secondary catchment
        viii. “biora zone” – BioRA zone mainstem wqm falls in
        ix. “catchment_2”, “catchment_3” – additional catchment name(s) if site has inflow from additional major tributaries
        x. “notes” – occasional notes on site matching
        xi. 4 KB
        xii. MRC metadata / spatial data
    b. wq_catchments_for_landcover_final.CSV
        i. Same as “wq_catchments_for_landcover” without sites withheld for model fitting
        ii. 3 KB
        iii. MRC metadata / spatial data
    c. wq_corresponding_eh.CSV
        i. “wq_code” – three letter WQM identifier
        ii. “wq_statid” - H+7 digit identifier
        iii. “wq_station_name” – name of the WQM
        iv. “wq_river_name_og” – name of the river as in WQ dataset
        v. “wq_river_name_sf” – name of the river as in WQ shapefile
        vi. “wq_country” – country WQM falls in
        vii. “eh_code” – three letter EHM identifier
        viii. “eh_sf_name” – name of the EHM as per EHM shapefile
        ix. “eh_station_name” – name of the EHM as per EHM metadata
        x. “eh_river” – river name of EHM
        xi. “eh_country” – name EHM falls in
        xii. “notes” – any notes on site matching
        xiii. “type” – t or m denotes tributary or mainstem river
        xiv. 6 KB
        xv. MRC metadata / spatial data
    d. wq_corresponding_eh_final.CSV
        i. Same as “wq_corresponding_eh_final” without sites withheld for model fitting
        ii. 6 KB

iii. MRC metadata / spatial data

4. dissfinal/server_landcover_output
    a. lulc_percent_summary.CSV
        i. 1 KB
        ii. “Year” – land use year
        iii. “Category” – land use group
        iv. “Percent” – percent of land use in category
        v. Source: “appendix e lu metrics.R”
    b. wqm_landcover_area_by_year.CSV
        i. 43 KB
        ii. “code” – three letter WQM identifier
        iii. “Year”
        iv. “Class_*#*” – area (m2) for each land use class
        v. “Source” – denotes land use boundary type
        vi. Source: “appendix e lu metrics.R”
    c. wqm_landcover_change_by_class.CSV
        i. 3 KB
        ii. “land_cover_class” – one class per row
        iii. “Year_*year*” – percent area by year
        iv. “releative_percent_change_2011_2021” – land use change from 2011 to 2021
        v. Source: “appendix e lu metrics.R”
    d. wqm_landcover_change_by_group.CSV
        i. 1 KB
        ii. “land_use” – land use group
        iii. “Year_*year*” – percent area by year
        iv. “releative_percent_change_2011_2021” – land use change from 2011 to 2021
        v. Source: “appendix e lu metrics.R”
    e. wqm_landcover_percent_by_year.CSV
        i. 84 KB
        ii. “code” – three letter WQM identifier
        iii. “Year”
        iv. “Class_*#*” – area (m2) for each land use class
        v. “Source” – denotes land use boundary type
        vi. “statid” - H+7 digit WQM identifier
        vii. “station_name” – WQM station name
        viii. “river_name_sf” – WQM river
        ix. Source: “appendix e lu metrics.R”
5. dissfinal/server_landcover_output/biora_landcover
    a. biora_landcover_all_yearsCSV
        i. 7KB
        ii. Contains all data from individual year CSVs
        iii. “Zone” – denotes BioRA zone

iv. "Year" – land use year
v. "Class_*#*" – number of pixels per class
vi. Source: "appendix e lu metrics.R"

b. landcover_*year*_biora_summary.CSV
i. 6 KB
ii. "layer" – arbitrary column
iii. "land_cover_class" – land use class number
iv. "count" – number of pixels
v. "area_m2" – calculated area (cell size x count)
vi. "Zone" – denotes BioRA zone
vii. "Year" – land use year
viii. Source: "appendix e lu metrics.R"

6. dissfinal/server_landcover_output/catchment_landcover
   a. catchment_landcover_all_years.CSV
      i. 73 KB
      ii. "Catchment" – denotes catchment
      iii. "Year" – land use year
      iv. "Class_*#*" - number of pixels per class
      v. Source: "appendix e lu metrics.R"
   b. landcover_*year*_catchment_summary.CSV
      i. 64 KB
      ii. "layer" – arbitrary column
      iii. "land_cover_class" – land use class number
      iv. "count" – number of pixels
      v. "area_m2" – calculated area (cell size x count)
      vi. "Catchment" – denotes catchment
      vii. "Year" – land use year
      viii. Source: "appendix e lu metrics.R"
7. dissfinal/spatial_data/
   a. BioRA_Zone, Catchment, countries, EHM_station_CRMN, LMB, River_maintributary, River_Mekong, wqm folders with corresponding PRJ, SHP, SHX
      i. 1521 KB – 10965 KB
      ii. These shapefiles and their accompanying files consist of some of the relevant spatial data for the project, mostly used for mapping
      iii. Extracted from the MRC WFS in "appendix b wfs extract.R"
   b. lu_study_area, major_tribs, select_catchments, select_ehm, select_tribs, select_wqm folders with corresponding PRJ, SHP, SHX
      i. 1 KB – 5482 KB
      ii. These shapefiles and their accompanying files contain spatial features relevant to the study; ie features were selected from the MRC shapefiles and saved as new output, "appendix c select shp.R"
8. dissfinal/spatial_data/servir_landcover
   a. Server_*year*.TIF

        i. ~ 1500 KB
        ii. One raster file of land use for the whole LMB per year
        iii. Downloaded from SERVIR-Mekong
9. dissfinal/statistical_analysis
    a. pearson_correlations.CSV
        i. 12 KB
        ii. Results of Pearsons correlation by year
        iii. Source: “appendix h pearsons.R”
10. dissfinal/statistical_analysis/gams
    a. gam_full_pooled.CSV
        i. 2 KB
        ii. Results of the pooled GAM with macros and land use as predictors
        iii. Source: “appendix k full gam.R”
    b. gam_full_yearly.CSV
        i. 8 KB
        ii. Results of the yearly GAM with macros and land use as predictors
        iii. Source: “appendix k full gam.R”
    c. gam_lu_pool.CSV
        i. 1 KB
        ii. Results of the pooled GAM with land use as predictor
        iii. Source: “appendix k lu gam.R”
    d. gam_lu_yearly.CSV
        i. 6 KB
        ii. Results of the yearly GAM with land use as predictor
        iii. Source: “appendix k lu gam.R”
    e. gam_macro_no_richness.CSV
        i. 1 KB
        ii. Results of the pooled macro GAM without richness as a predictor
        iii. Source: “appendix k macro gam.R”
    f. gam_macro_pool.CSV
        i. 1 KB
        ii. Results of the pooled GAM with macros as predictors
        iii. Source: “appendix k macro gam.R”
    g. gam_macro_yearly.CSV
        i. 5 KB
        ii. Results of the yearly GAM with macros as predictors
        iii. Source: “appendix k macro gam.R”
    h. gam_pooled_mainstem.CSV
        i. 2 KB
        ii. Results of the full pooled GAM with mainstem sites only
        iii. Source: “appendix k mvt gam.R”
    i. gam_pooled_tributary.CSV
        i. 2 KB
        ii. Results of the full pooled GAM with tributary sites only

iii. Source: “appendix k mvt gam.R”

j. gam_yearly_mainstem.CSV
   i. 8 KB, Results of the full yearly GAM with mainstem sites only
   ii. Source: “appendix k mvt gam.R”

k. gam_yearly_tributary.CSV
   i. 8 KB
   ii. Results of the full yearly GAM with tributary sites only
   iii. Source: “appendix k mvt gam.R”

l. gam_full_predictions_rmse.CSV
   i. 1 KB
   ii. RMSE and residual values of each WQ metric for the predicted sites
   iii. Source: “appendix k full gam.R”

m. gam_full_predictions.CSV
   i. 9 KB
   ii. Contains the actual and predicted values for each WQ metric per year
   iii. Source: “appendix k full gam.R”

n. gam_full_outliers.CSV
   i. 1 KB
   ii. Percent of times a site is an outlier
   iii. Source: “appendix k full gam.R”

o. gam_full_outliers_river.CSV
   i. 1 KB
   ii. Percent of times a site is an outlier for major rivers
   iii. Source: “appendix k full gam.R”

p. gam_full_outliers_bins.CSV
   i. 1 KB
   ii. Percent of times a site is an outlier for spatial bins
   iii. Source: “appendix k full gam.R”

q. gam_full_outliers_zone.CSV
   i. 1 KB
   ii. Percent of times a site is an outlier for tributary or mainstem rivers
   iii. Source: “appendix k full gam.R”

11. dissfinal/statistical_analysis/mlr_macro
    a. mlr_macro_pooled.CSV
       i. 4 KB
       ii. Results of the pooled MLR with macros as predictors
       iii. Source: “appendix j mlr.R”
    b. mlr_macro_yearly.CSV
       i. 20 KB
       ii. Results of the yearly MLR with macros as predictors
       iii. Source: “appendix j mlr.R”
12. dissfinal/statistical_analysis/rda_results
    a. rda_lu_pooled_overall.CSV
       i. 1 KB

ii. Results of the pooled RDA with land use as predictor
iii. Source: “appendix i rda.R”

b. rda_lu_pooled_terms.CSV
i. 1 KB
ii. Results of the pooled RDA with land use as predictor
iii. Source: “appendix i rda.R”

c. rda_lu_yearly_overall*year*.CSV
i. 1 KB
ii. Results of the yearly RDA with land use as predictor; one file per year
iii. Source: “appendix i rda.R”

d. rda_lu_yearly_summary.CSV
i. 1 KB
ii. Summary file of yearly RDA with land use as predictor
iii. Source: “appendix i rda.R”

e. rda_lu_yearly_terms*year*.CSV
i. 1 KB
ii. Results of the yearly RDA with land use as predictor; one file per year
iii. Source: “appendix i rda.R”

f. rda_macro_pooled_overall.CSV
i. 1 KB
ii. Results of the pooled RDA with macro metrics as predictors
iii. Source: “appendix i rda.R”

g. rda_macro_pooled_terms.CSV
i. 1 KB
ii. Results of the pooled RDA with macro metrics as predictors
iii. Source: “appendix i rda.R”

13. dissfinal/water_quality_data

a. wq_data_cleaned.CSV
i. 3 KB
ii. Original dataset from MRC and cleaned in “appendix f wq clean.R”

b. wq_data_relevant_year_month.CSV
i. Same as wq_data_cleaned but for dry season months and study years
ii. Source: “appendix f wq clean.R”

c. selected_wq_sites.CSV
i. 114 KB
ii. Three sites missing from original file but included in WQ data were added
iii. “code” – three letter WQM identifier
iv. “statid” - H+7 digit WQM identifier
v. “station_name”
vi. “River_Name” – name of river as per metadata
vii. “river_name_sf” – name of river as per tributaries shapefile
viii. “Country”
ix. “Latitiude”
x. “Longitude”

xi. Source: MRC WQ metadata

d. WQMN_data_for_Ava.CSV
    i. 1021 KB
    ii. See Table 3 for full descriptions of column structure and data
    iii. Source: MRC WQM dataset

Appendix B. Data Extraction from MRC WFS Code

In order to successfully run the code detailed throughout the appendices, the working directory should be set prior to running any code chunk; all other file paths in the code will be correct as long at the working directory ends with “dissfinal” and the necessary baseline data is in the correct location. Code can be run in order of appendices.

This code is from “dissfinal/code/appendix b wfs extract.R”

```
# ---------------------------------- DATA EXTRACTION FROM MRC WFS ----------------------------------

# Set relevant working directory
setwd("/home/s2502571/dissertation/dissfinal")

# Install and load libraries
install.packages("sf")
install.packages("httr")
library(sf) # For working with spatial data using simple features
library(httr) # For handling web APIs

# -------------- MRC WFS Extraction -------------

# Define the WMS URL and layer
wfs_base <- "https://geo.mrcmekong.org/geo/mrc/ows?"

# Target layers
layers <- c("EHM_station_CRMN", "BioRA_Zone", "River_Mekong", "River_maintributary",
"Catchment", "LMB")

# Specify output directory
base_output_dir <- "spatial_data"

# Loop through layers
for (layer_name in layers) {
  message("Downloading: ", layer_name)

  # Construct full WFS URL for this layer
  wfs_url <- paste0(
    wfs_base,
    "service=WFS&version=1.1.0&request=GetFeature&typeName=", layer_name,
"&outputFormat=application/json"
  )
```

```
# Read as GeoJSON from WFS
  vec <- st_read(wfs_url, quiet = TRUE)

  # Reproject to ESPG 32648
  vec_espg <- st_transform(vec, 32648)

  # Clean layer name for file paths
  safe_layer_name <- gsub("[:% ]", "_", layer_name)  # Replace colons, spaces, and percent signs
with underscores
 # Create a subfolder for each layer
  layer_dir <- file.path(base_output_dir, layer_name)
  if (!dir.exists(layer_dir)) {
    dir.create(layer_dir, recursive = TRUE)
  }

  # Save as shapefile inside its respective folder
  out_path <- file.path(layer_dir, paste0(layer_name, ".shp"))
  st_write(vec_espg, out_path, delete_layer = TRUE, quiet = TRUE)
  message("Saved: ", out_path)
}

# ---------- Extract Country Boundaries Separately ----------
# (would not work in the loop due to naming conventions)

wfs_base <- "https://geo.mrcmekong.org/geo/mrc/ows?"

# Target layer for country boundaries
layer_name <- URLencode("mrc:Country Boundary", reserved = TRUE)

# Construct full WFS request
wfs_url <- paste0(
  wfs_base,
  "service=WFS&version=1.1.0&request=GetFeature",
  "&typeName=", layer_name,
  "&outputFormat=application/json"
)

# Read the layer
country_boundary <- st_read(wfs_url)

# Transform to ESPG 32648
country_boundary_proj <- st_transform(country_boundary, 4326)
```

```
# Drop Z (and M if present) dimensions
country_boundary_2d <- st_zm(country_boundary_proj)

# Create the output folder and save the shapefile
dir.create("spatial_data/countries", recursive = TRUE)
st_write(country_boundary_2d, "spatial_data/countries/countries.shp")

# ---------- Extract WQM separately ----------
wfs_base <- "https://geo.mrcmekong.org/geo/mrc/ows?"

# Helper function to create output directories
create_output_dir <- function(path) {
  if (!dir.exists(path)) {
    dir.create(path, recursive = TRUE)
  }
}

# Save Water Quality Stations
extract_wqm <- function(output_dir) {
  create_output_dir(output_dir)

  layer_name <- URLencode("Water Quality Stations", reserved = TRUE)

  wfs_url <- paste0(
    wfs_base,
    "service=WFS&version=1.1.0&request=GetFeature",
    "&typeName=", layer_name,
    "&outputFormat=application/json"
  )

  message("Fetching WQM data...")
  wqm <- st_read(wfs_url)

  st_write(wqm, file.path(output_dir, "wqm_stations.shp"), delete_dsn = TRUE)

  return(wqm)
}
# Prepare and save country boundaries
save_wqm <- function(wqm, output_dir, crs = 32648) {
  create_output_dir(output_dir)

  wqm_proj <- st_transform(wqm, crs)
  wqm_2d <- st_zm(wqm_proj)
```

```
  st_write(wqm_2d, file.path(output_dir, "wqm.shp"), delete_dsn = TRUE)
}

# Run the full pipeline
output_dir <- "spatial_data/wqm"
wqm <- extract_wqm(output_dir)
save_wqm(wqm, output_dir)
```

Appendix C. Shapefile Processing Code

Code from “dissfinal/code/appendix c select shp.R”

```
# ---------------------------------------- PROCESS SHAPEFILES ----------------------------------------

# Install new libraries and load all libraries
install.packages("dplyr")
install.packages("tidyr")
library(dplyr) # Data manipulation and transformation
library(sf)
library(tidyr) # Data tidying and reshaping

# Load shapefiles
lmb <- st_read("spatial_data/LMB/LMB.shp") %>%
  st_transform("EPSG:32648")

catchment <- st_read("spatial_data/Catchment/Catchment.shp") %>%
  st_transform("EPSG:32648")

country <- st_read("spatial_data/countries/countries.shp") %>%
  st_transform("EPSG:32648")

mekong <- st_read("spatial_data/River_Mekong/River_Mekong.shp") %>%
  st_transform("EPSG:32648")

biora <- st_read("spatial_data/BioRA_Zone/BioRA_Zone.shp") %>%
  st_transform("EPSG:32648")

tributary <- st_read("spatial_data/River_maintributary/River_maintributary.shp") %>%
  st_transform("EPSG:32648")

ehm <- st_read("spatial_data/EHM_station_CRMN/EHM_station_CRMN.shp") %>%
  st_transform("EPSG:32648")

wqm <- st_read("spatial_data/wqm/wqm.shp") %>%
  st_transform("EPSG:32648")

# Load data
eh_data <- read.csv("ecological_health_data/eh_siteranking.csv")
wq_data <- read.csv("water_quality_data/wq_data_cleaned.csv")
# Load metadata for joins
select_stations_md <- read.csv("join_metadata/wq_corresponding_eh_final.csv")
select_catchments_md <- read.csv("join_metadata/wq_catchments_for_landcover_final.csv")
```

```
# ---------- EXTRACT RELEVANT EHM ----------

# Extract selected EHM
sf_select_ehm <- ehm[ehm$Site_name %in% select_stations_md$eh_sf_name, ]

# Check match
length(unique(select_stations_md$eh_sf_name))
length(unique(sf_select_ehm$Site_name))

# Create a new folder (if it doesn't already exist)
dir.create("spatial_data/select_ehm", showWarnings = FALSE)

# Save the shapefile and read for later
st_write(sf_select_ehm, "spatial_data/select_ehm/select_ehm.shp", delete_layer = TRUE)
select_ehm <- st_read("spatial_data/select_ehm/select_ehm.shp")

# ---------- EXTRACT RELEVANT WQM ----------

# Extract selected WQM
sf_select_wqm <- wqm[wqm$Sttn_nm %in% select_stations_md$wq_station_name, ]

# Check match
length(unique(select_stations_md$wq_station_name))
length(unique(sf_select_wqm$Sttn_nm))

# Create a new folder (if it doesn't already exist)
dir.create("spatial_data/select_wqm", showWarnings = FALSE)

# Save the shapefile and read for later
st_write(sf_select_wqm, "spatial_data/select_wqm/select_wqm.shp", delete_layer = TRUE)
select_wqm <- st_read("spatial_data/select_wqm/select_wqm.shp")

# ---------- EXTRACT RELEVANT CATCHMENTS ----------
# Gather all catchment names into one column
catchment_names <- select_catchments_md %>%
  select(catchment, catchment_2, catchment_3) %>%
  pivot_longer(everything(), names_to = "source", values_to = "catchment_name") %>%
  filter(!is.na(catchment_name)) %>%
  distinct(catchment_name) %>%
  pull(catchment_name)

# Extract selected catchments
sf_select_catchments <- catchment %>%
  filter(Ctchmn_ %in% catchment_names)
```

```
# Create a new folder (if it doesn't already exist)
dir.create("spatial_data/select_catchments", showWarnings = FALSE)

# Save the shapefile and read for later
st_write(sf_select_catchments, "spatial_data/select_catchments/select_catchments.shp",
delete_layer = TRUE)
select_catchments <- st_read("spatial_data/select_catchments/select_catchments.shp")

# ---------- EXTRACT RELEVANT TRIBUTARIES ----------

# Extract tributaries in selected catchments
sf_select_tribs <- tributary %>%
  st_filter(sf_select_catchments, .predicate = st_intersects)

## Create a new folder (if it doesn't already exist)
dir.create("spatial_data/select_tribs", showWarnings = FALSE)

## Save the shapefile and read for later
st_write(sf_select_tribs, "spatial_data/select_tribs/select_tribs.shp", delete_layer = TRUE)
select_tribs <- st_read("spatial_data/select_tribs/select_tribs.shp")

# Extract major LMB tributaries
## Create a vector of target tributary names
major_trib_names <- c("Bassac", "Nam Khan", "Nam Mae Ing", "Nam Mae Kok", "Nam Mun",
                      "Nam Ngum", "Nam Ou", "Nam Song", "Nam Tha", "Se Bang Fai",
                      "Se Bang Hieng", "Se Kong", "Se San", "Sre Pok", "Tonle Sap",
                      "Nam Ka Dinh", "Nam Songkhram")

## Check column names in tributary layer
sort(unique(tributary$River_Name))
names(tributary)

## Filter based on correct column
tribs_major <- tributary %>%
  filter(River_Name %in% major_trib_names)

## Create a new folder (if it doesn't already exist)
dir.create("spatial_data/major_tribs", showWarnings = FALSE)

## Save the shapefile and read for later
st_write(tribs_major, "spatial_data/major_tribs/tribs_major.shp", delete_layer = TRUE)
major_tribs <- st_read("spatial_data/major_tribs/tribs_major.shp")
```

```
# ---------- CREATE LU STUDY AREA BOUNDARY ----------

# Buffer Mekong
biora_buffered <- biora %>%
  summarise(geometry = st_union(geometry)) %>%
  st_buffer(dist = 15000)

# Combine & simplify study area
combined_area <- rbind(
  select_catchments %>% select(geometry),
  biora_buffered %>% select(geometry)
)
study_region <- st_union(combined_area)  # merge catchments and buffer into single shape
study_region_clipped <- st_intersection(study_region, lmb)  # clip to LMB

# Create a new folder (if it doesn't already exist)
dir.create("spatial_data/lu_study_area", showWarnings = FALSE)

# Save the shapefile and load for later
st_write(study_region_clipped, "spatial_data/lu_study_area/lu_study_area.shp", delete_layer =
TRUE)
lu_study_area <- st_read("spatial_data/lu_study_area/lu_study_area.shp")
```

Appendix D. Landcover Data Processing Code

Code from “dissfinal/code/appendix d lu processing.R”

```
# ------------------------------------ LANDCOVER DATA PROCESSING ------------------------------------

# Note: landcover rasters for 2011, 2013, 2015, 2017, 2019, and 2021 were obtained from SERVIR-Mekong and saved in a folder called "servir_landcover"

# Install new libraries and load all libraries
install.packages("terra")
install.packages("tmap")
library(dplyr)
library(sf)
library(terra) # For spatial analysis
library(tmap) # Mapping and spatial visualization

# ----- Check the extent of LMB and landcover raster ------

# Load example landcover raster for plotting
landcover_21 <- rast("spatial_data/servir_landcover/servir_2021.tif")

# Load shapefiles as vectors
mekong <- vect("spatial_data/River_Mekong/River_Mekong.shp")
lmb <- vect("spatial_data/LMB/LMB.shp")

# Check crs and extent
crs(landcover_21)
crs(mekong)
crs(lmb)
ext(landcover_21)
ext(mekong)
ext(lmb)

# Plot using the extent of the LMB shapefile
plot(landcover_21, main = "Full LMB with Raster Overlay" )
lines(mekong, col = "blue")
lines(lmb, col = "orange")

# ---------- Determine Catchment Landcover -----------

# Load catchment shapefile
catchments <- st_read("spatial_data/Catchment/Catchment.shp")
```

```
# Define years and corresponding raster file paths
years <- c(2011, 2013, 2015, 2017, 2019, 2021)
raster_paths <- paste0("spatial_data/servir_landcover/servir_", years, ".tif")

# Output folders
output_folder <- "servir_landcover_output/catchment_landcover/"
catchment_raster_folder <- paste0(output_folder, "catchment_rasters/")
dir.create(catchment_raster_folder, recursive = TRUE, showWarnings = FALSE)

# Loop through each year and process land use data
for (j in seq_along(years)) {
  year <- years[j]
  raster_path <- raster_paths[j]

  # Load and reproject raster
  lu_raster <- rast(raster_path)
  lu_raster <- project(lu_raster, "EPSG:32648", res=c(30, 30))

  # Get cell area
  cell_area <- prod(res(lu_raster))
  print(res(lu_raster))  # Ensure it's still (30, 30)

  zone_summaries <- list()

  # Iterate over catchments
  for (i in seq_len(nrow(catchments))) {
    zone_geom <- catchments[i, ]  # Select catchment geometry
    zone_name <- zone_geom$Ctchmn_
    zone_safe <- gsub("[^a-zA-Z0-9_]", "_", zone_name)

    # Convert sf object to SpatVector for terra functions
    zone_geom_vect <- vect(zone_geom)

    # Check if extents overlap before cropping
    if (!relate(ext(lu_raster), ext(zone_geom_vect), "intersects")) {
      message("Skipping ", zone_name, " for year ", year, " - No overlap with raster")
      next
    }

    # Mask landcover raster by catchment
    lc_zone <- mask(crop(lu_raster, zone_geom_vect), zone_geom_vect)

    # Frequency table
    freq_df <- as.data.frame(freq(lc_zone)) %>%
```

```
    filter(!is.na(value))

  if (nrow(freq_df) == 0) next

  freq_df <- freq_df %>%
    mutate(
      area_m2 = count * cell_area,
      Catchment = zone_name,
      Year = year
    ) %>%
    rename(land_cover_class = value)

  zone_summaries[[i]] <- freq_df
 }

 # Combine summaries for the year
 catchment_landcover_summary <- bind_rows(zone_summaries)

 # Save as CSV file
 csv_path <- paste0(output_folder, "landcover_", year, "_catchment_summary.csv")
 write.csv(catchment_landcover_summary, csv_path, row.names = FALSE)

 message("Saved land cover summary for ", year, " at ", csv_path)
}

message("All years processed successfully!")

# Ensure calculated and actual catchment area are similar / the same
## Filter for Nam Ou catchment
nam_ou <- catchments %>% filter(Ctchmn_ == "NAM OU")

## Calculate area in square meters
nam_ou_area_m2 <- st_area(nam_ou)

## Print results
print(paste("Nam Ou Catchment Area:", round(nam_ou_area_m2, 2), "m²")) # Compare result to
that of csv

# ------ Determine Mekong / BioRA Landcover --------

# Load BioRA river zones and apply buffer once (15km)
biora <- st_read("spatial_data/BioRA_Zone/BioRA_Zone.shp") %>%
 st_transform(32648)
```

```
biora <- biora %>%
  group_by(Zone) %>%
  summarise(geometry = st_union(geometry), .groups = "drop")

biora_buffered <- st_buffer(biora, dist = 15000)  # Apply buffer ONCE before loop
biora_buf_vect <- vect(biora_buffered)  # Convert to SpatVector for terra functions

# Define years and raster file paths
years <- c(2011, 2013, 2015, 2017, 2019, 2021)
raster_paths <- paste0("spatial_data/servir_landcover/servir_", years, ".tif")

# Output folder
output_folder <- "servir_landcover_output/biora_landcover/"
biora_raster_folder <- paste0(output_folder, "biora_rasters/")
dir.create(biora_raster_folder, recursive = TRUE, showWarnings = FALSE)

# Loop through each year
for (j in seq_along(years)) {
  year <- years[j]
  raster_path <- raster_paths[j]

  # Load and reproject raster
  lu_raster <- rast(raster_path)
  lu_raster <- project(lu_raster, "EPSG:32648", res = c(30, 30))

  # Get cell area
  cell_area <- prod(res(lu_raster))

  zone_summaries <- list()

  # Iterate over buffered BioRA zones
  for (i in seq_len(nrow(biora_buffered))) {
    zone_geom <- biora_buf_vect[i]  # Select buffered river geometry
    zone_name <- biora$Zone[i]  # Use original zone name
    zone_safe <- gsub("[^a-zA-Z0-9_]", "_", zone_name)

    # Check if extents overlap before cropping
    if (!relate(ext(lu_raster), ext(zone_geom), "intersects")) {
      message("Skipping ", zone_name, " for year ", year, " - No overlap with raster")
      next
    }

    # Mask landcover raster by buffered BioRA zone
    lc_zone <- mask(crop(lu_raster, zone_geom), zone_geom)
```

```
  # Frequency table
  freq_df <- as.data.frame(freq(lc_zone)) %>%
    filter(!is.na(value))

  if (nrow(freq_df) == 0) next

  freq_df <- freq_df %>%
    mutate(
      area_m2 = count * cell_area,
      Zone = zone_name,
      Year = year
    ) %>%
    rename(land_cover_class = value)

  zone_summaries[[i]] <- freq_df
 }

 # Combine summaries for the year
 landcover_summary <- bind_rows(zone_summaries)

 # Save as CSV file
 csv_path <- paste0(output_folder, "landcover_", year, "_biora_summary.csv")
 write.csv(landcover_summary, csv_path, row.names = FALSE)

 message("Saved land cover summary for ", year, " at ", csv_path)
}

message("All years processed successfully!")
```

Appendix E. Land Use Metrics Calculation Code

Code from “dissfinal/code/appendix e lu metrics.R”

```
# ------------------------------------ CALCULATE LAND USE METRICS ------------------------------------

# Install new libraries and load all libraries
install.packages("purr")
install.packages("readr")
install.packages("stringr")
library(dplyr)
library(purrr)  # Map functions
library(readr)  # Reading text files
library(readxl)
library(stringr)  # String manipulation
library(terra)
library(tidyr)

# Load polygon
study_area <- vect("spatial_data/lu_study_area/lu_study_area.shp")

# Load rasters
years <- c(2011, 2013, 2015, 2017, 2019, 2021)
lulc_files <- paste0("spatial_data/servir_landcover/servir_", years, ".tif")
lulc_rasters <- lapply(lulc_files, rast)
names(lulc_rasters) <- years

# Define land use class groupings
class_groups <- list(
  forest = c(5, 6, 7, 10, 11, 18),
  agriculture = c(1, 3, 4, 12, 13, 14),
  urban = c(2, 16),
  semi_natural = c(8, 9, 15, 17)
)

# ------------- CALCULATE LU PER CATEGORY WITHIN STUDY AREA -------------

# Function to extract and process values
extract_grouped_percents <- function(raster, year) {
  # Mask the raster using the polygon
  masked <- mask(crop(raster, study_area), study_area)

  # Get all pixel values as vector
  values <- values(masked, mat = FALSE)
```

```
  values <- na.omit(values)  # Remove NA values

  # Tabulate counts
  freq_table <- as.data.frame(table(values))
  colnames(freq_table) <- c("Class", "Count")
  freq_table$Class <- as.integer(as.character(freq_table$Class))

  # Group into categories and compute percent
  freq_table <- freq_table %>%
    mutate(Category = case_when(
      Class %in% class_groups$forest ~ "forest",
      Class %in% class_groups$agriculture ~ "agriculture",
      Class %in% class_groups$urban ~ "urban",
      Class %in% class_groups$semi_natural ~ "semi_natural",
      TRUE ~ "other"
    )) %>%
    group_by(Category) %>%
    summarise(Total = sum(Count), .groups = "drop") %>%
    mutate(Year = year, Percent = 100 * Total / sum(Total)) %>%
    select(Year, Category, Percent)
}

# Apply to all years
lulc_summary <- lapply(seq_along(lulc_rasters), function(i) {
  extract_grouped_percents(lulc_rasters[[i]], years[i])
}) %>% bind_rows()

# Save
write.csv(lulc_summary, "servir_landcover_output/lulc_percent_summary.csv", row.names =
FALSE)

# ---------- REFORMAT DATA ----------

# Define years to process
years <- c(2011, 2013, 2015, 2017, 2019, 2021)

# Initialize empty list to store ALL catchment data
catchment_list <- list()
```

```
# Loop through each year and read data
for (year in years) {
  file_path <- paste0("servir_landcover_output/catchment_landcover/landcover_", year,
"_catchment_summary.csv")

  if (file.exists(file_path)) {
    df_catchment <- read_csv(file_path) %>%
      mutate(Year = year)  # Add Year column for tracking

    # Pivot to restructure data
    df_catchment_transformed <- df_catchment %>%
      select(Catchment, land_cover_class, count, Year) %>%
      pivot_wider(names_from = land_cover_class, values_from = count, names_prefix = "Class_")
%>%
      replace(is.na(.), 0)  # Fill missing classes with zeros

    catchment_list[[as.character(year)]] <- df_catchment_transformed
  } else {
    print(paste("File not found:", file_path))
  }
}

# Combine all years into one dataset
catchment_combined <- bind_rows(catchment_list)

# Save to CSV
write_csv(catchment_combined,
"servir_landcover_output/catchment_landcover/catchment_landcover_all_years.csv")

# Initialize empty list to store ALL BioRA data
biora_list <- list()

# Loop through each year and read data
for (year in years) {
  file_path <- paste0("servir_landcover_output/biora_landcover/landcover_", year,
"_biora_summary.csv")

  if (file.exists(file_path)) {
    df_biora <- read_csv(file_path) %>%
      mutate(Year = year)  # Add Year column for tracking
```

```
 # Pivot to restructure data
   df_biora_transformed <- df_biora %>%
     select(Zone, land_cover_class, count, Year) %>%
     pivot_wider(names_from = land_cover_class, values_from = count, names_prefix = "Class_")
%>%
     replace(is.na(.), 0)  # Fill missing classes with zeros

   biora_list[[as.character(year)]] <- df_biora_transformed
  } else {
   print(paste("File not found:", file_path))
  }
}

# Combine all years into one dataset
biora_combined <- bind_rows(biora_list)

# Save to CSV
write_csv(biora_combined,
"servir_landcover_output/biora_landcover/biora_landcover_all_years.csv")

# ---------- SUMMARIZE LU FOR EACH WQ SITE ----------

# Load station metadata and clean
station_meta <- read.csv("join_metadata/wq_catchments_for_landcover.csv") %>%
  mutate(across(everything(), ~str_trim(as.character(.))))
wq_sites <- read_csv("water_quality_data/selected_wq_sites.csv") %>%
  mutate(code = str_trim(as.character(code)))

# Define years
years <- c(2011, 2013, 2015, 2017, 2019, 2021)

# Load land use files
load_landuse <- function(years, folder_path, file_suffix, level) {
  map_dfr(years, function(yr) {
    file <- paste0(folder_path, "landcover_", yr, file_suffix)
    read_csv(file) %>%
      mutate(Year = yr, Level = level)
  })
}

catchment_lu <- load_landuse(
  years,
  "servir_landcover_output/catchment_landcover/",
  "_catchment_summary.csv",
```

```
  "Catchment"
)

biora_lu <- load_landuse(
  years,
  "servir_landcover_output/biora_landcover/",
  "_biora_summary.csv",
  "Biora"
)

# Summarize land use by unit
summarize_area <- function(df, zone_col) {
  df %>%
    group_by(across(all_of(c(zone_col, "Year", "land_cover_class")))) %>%
    summarise(area_m2 = sum(area_m2), .groups = "drop")
}

catchment_area <- summarize_area(catchment_lu, "Catchment")
biora_area     <- summarize_area(biora_lu, "Zone")

# Expand station metadata across all years and normalize
station_years <- expand_grid(station_meta, Year = years) %>%
  mutate(across(everything(), ~str_trim(as.character(.)))) %>%
  mutate(
    biora.zone = str_to_upper(na_if(biora.zone, "")),
    catchment  = na_if(catchment, ""),
    Year = as.numeric(Year)
  )

biora_area <- biora_area %>%
  mutate(Zone = str_to_upper(str_trim(Zone))) # clean and standardize zone

# ---------- Aggregate raw areas by station-year

# Tributary-only stations
trib_stations <- station_years %>%
  filter(is.na(biora.zone)) %>%
  pivot_longer(cols = starts_with("catchment"), names_to = NULL, values_to = "Catchment_ID")
%>%
  filter(!is.na(Catchment_ID)) %>%
  left_join(catchment_area, by = c("Catchment_ID" = "Catchment", "Year")) %>%
  group_by(code, Year, land_cover_class) %>%
  summarise(area_m2 = sum(area_m2), .groups = "drop")
```

```
# Biora-only stations
biora_only <- station_years %>%
  filter(!is.na(biora.zone) & is.na(catchment)) %>%
  left_join(biora_area, by = c("biora.zone" = "Zone", "Year")) %>%
  group_by(code, Year, land_cover_class) %>%
  summarise(area_m2 = sum(area_m2), .groups = "drop")

# Mixed stations (zone + catchment)
biora_zone_data <- station_years %>%
  filter(!is.na(biora.zone) & !is.na(catchment)) %>%
  select(code, Year, biora.zone) %>%
  distinct() %>%
  left_join(biora_area, by = c("biora.zone" = "Zone", "Year"))

catchment_data <- station_years %>%
  filter(!is.na(biora.zone) & !is.na(catchment)) %>%
  pivot_longer(cols = starts_with("catchment"), names_to = NULL, values_to = "Catchment_ID")
%>%
  filter(!is.na(Catchment_ID)) %>%
  left_join(catchment_area, by = c("Catchment_ID" = "Catchment", "Year"))

biora_catch_combined <- bind_rows(
  biora_zone_data %>% select(code, Year, land_cover_class, area_m2),
  catchment_data  %>% select(code, Year, land_cover_class, area_m2)
) %>%
  filter(!is.na(area_m2)) %>%
  group_by(code, Year, land_cover_class) %>%
  summarise(area_m2 = sum(area_m2), .groups = "drop")

# ---------- LU PERCENT CALCULATION ----------

# Convert areas to percent cover
convert_to_percent <- function(df) {
  df %>%
    filter(!is.na(land_cover_class)) %>%
    group_by(code, Year) %>%
    mutate(Total = sum(area_m2, na.rm = TRUE)) %>%
    ungroup() %>%
    mutate(Percent = (area_m2 / Total) * 100) %>%
    select(-area_m2, -Total) %>%
    pivot_wider(
      names_from = land_cover_class,
      values_from = Percent,
      names_prefix = "Class_",
```

```
    values_fill = list(Percent = 0)
  )
}

trib_pct   <- convert_to_percent(trib_stations)
biora_pct2 <- convert_to_percent(biora_only)
mixed_pct  <- convert_to_percent(biora_catch_combined)

# Merge all station-year cover summaries
station_lu_final <- bind_rows(
  trib_pct   %>% mutate(Source = "Catchment"),
  biora_pct2 %>% mutate(Source = "Biora"),
  mixed_pct  %>% mutate(Source = "Zone + Catchment")
)

# Add station ids
station_lu_final <- station_lu_final %>%
  left_join(wq_sites, by = "code")
# Remove unnecessay columns from wq_sites join
station_lu_final <- station_lu_final %>%
  select(-Longitude, -Latitude, -River_Names_og, -Country)

# Export
write_csv(station_lu_final, "servir_landcover_output/wqm_landcover_percent_by_year.csv")

# ---------- LU AREA M2 CALCULATION ----------

# Pivot area data to wide format by land cover class (same structure, just with area_m2)
pivot_area_data <- function(df) {
  df %>%
    filter(!is.na(land_cover_class)) %>%
    pivot_wider(
      names_from = land_cover_class,
      values_from = area_m2,
      names_prefix = "Class_",
      values_fill = list(area_m2 = 0)
    )
}

trib_area   <- pivot_area_data(trib_stations)
biora_area2 <- pivot_area_data(biora_only)
mixed_area  <- pivot_area_data(biora_catch_combined)
```

```
# Add Source label and combine
station_area_final <- bind_rows(
  trib_area   %>% mutate(Source = "Catchment"),
  biora_area2 %>% mutate(Source = "Biora"),
  mixed_area  %>% mutate(Source = "Zone + Catchment")
)

# Export to CSV
write_csv(station_area_final, "servir_landcover_output/wqm_landcover_area_by_year.csv")

# ---------- LU AREA CHANGE ----------

# Pivot and calculate percent of each class per year
regional_percents <- station_area_final %>%
  filter(Year %in% c(2011, 2013, 2015, 2017, 2019, 2021)) %>%
  pivot_longer(cols = starts_with("Class_"), names_to = "land_cover_class", values_to =
"area_m2") %>%
  group_by(Year, land_cover_class) %>%
  summarise(total_area_m2 = sum(area_m2, na.rm = TRUE), .groups = "drop") %>%
  group_by(Year) %>%
  mutate(percent_of_total = (total_area_m2 / sum(total_area_m2, na.rm = TRUE)) * 100) %>%
  ungroup() %>%
  select(-total_area_m2) %>%
  pivot_wider(names_from = Year, values_from = percent_of_total, names_prefix = "Year_") %>%
  mutate(relative_percent_change_2011_2021 = ((Year_2021 - Year_2011) / Year_2011) * 100)

# Export
write_csv(regional_percents, "servir_landcover_output/wqm_landcover_change_by_class.csv")

# land use change for the whole region by aggregated type
## define class groups
class_grouping <- station_area_final %>%
  mutate(
    forest = Class_5 + Class_6 + Class_7 + Class_10 + Class_11,
    agriculture = Class_3 + Class_4 + Class_12 + Class_13 + Class_14 + Class_1,
    urban = Class_16 + Class_2,
    semi_natural = Class_8 + Class_9 + Class_15 + Class_17
  ) %>%
  select(code, Year, forest, agriculture, urban, semi_natural) %>%
  rename(year_collected = Year)
# Calculate total area by land use category and year
regional_group_percents <- class_grouping %>%
  filter(year_collected %in% c(2011, 2013, 2015, 2017, 2019, 2021)) %>%
  pivot_longer(cols = c(forest, agriculture, urban, semi_natural),
```

```
            names_to = "land_use", values_to = "area_m2") %>%
  group_by(year_collected, land_use) %>%
  summarise(total_area_m2 = sum(area_m2, na.rm = TRUE), .groups = "drop") %>%
  group_by(year_collected) %>%
  mutate(percent_of_total = (total_area_m2 / sum(total_area_m2)) * 100) %>%
  ungroup() %>%
  select(-total_area_m2) %>%
  pivot_wider(names_from = year_collected, values_from = percent_of_total, names_prefix =
"Year_") %>%
  mutate(relative_percent_change_2011_2021 = ((Year_2021 - Year_2011) / Year_2011) * 100)

# Export
write_csv(regional_group_percents,
"servir_landcover_output/wqm_landcover_change_by_group.csv")
```

Appendix F. Water Quality Data Cleaning Code

Code from “dissfinal/code/appendix f wq clean.R”

```
# ---------------------------------- CLEAN WATER QUALITY DATA -----------------------------------

# Install and load packages
install.packages("lubridate")
install.packages("readxl")
install.packages("tidyverse")
library(lubridate) # For formatting
library(readxl) # For reading Excel files
library(tidyverse) # Loads full tidyverse collection

# Read in the data shared by the MRC
data <- read.csv("water_quality_data/WQMN_data_for_Ava.csv")

# View basic structure
glimpse(data)
summary(data)

# Clean column names
data <- data %>%
  rename_with(tolower)

colnames(data) # check new column names

# Remove 2024 data (it’s in a different format, ugly, and unnecessary)
data <- data %>%
  filter(year_collected != 2024)

# Remove data for three unnecessary stations
data <- data %>%
  filter(!(name %in% c("Tan Thanh")))

# Check class type and convert
str(data)
```

```
## See nonnumeric data in each column that will cause conversion issues
lapply(data[, c("tidehl", "flow_m3s", "temp_c", "ph", "tss_mgl", "cond_msm", "totn_mgl",
   "totp_mgl", "do_mgl", "codmn_mgl", "fc_mpn_100ml", "bod_mgl")], function(x)
   unique(x[!grepl("^[-0-9\\.]+$", x)]))

## Convert and remove problematic data
data <- data %>%
  mutate(across(c(tidehl, flow_m3s, temp_c, ph, tss_mgl, cond_msm, totn_mgl, totp_mgl,
do_mgl, codmn_mgl, fc_mpn_100ml, bod_mgl),
              ~ifelse(grepl("^[-0-9\\.]+$", .), as.numeric(.), NA)))

## Verify conversion
sapply(data, class)

# Look at unique station name combinations
unique_combos <- data %>%
  distinct(name, statid) %>%
  arrange(statid, name)

print(unique_combos)

# Change wrong station ids
data <- data %>%
  mutate(statid = case_when(
    name == "Backprea" ~ "H020107",
    name == "Chrouy Changvar" ~ "H019801",
    name == "Kampong Loung" ~ "H020106",
    name == "Koh Thom" ~ "H033403",
    name == "Kaorm Samnor" ~ "H019807",
    name == "Prek Kdam" ~ "H020102",
    name == "Angdoung Meas" ~ "H0440103",
    TRUE ~ statid  # Keep original values for others
  ))

# Add three digit station code
selected_sites <- read.csv("water_quality_data/selected_wq_sites.csv")

## rename column from wq data
data <- data %>% rename(station_name = name)

## add relevant columns
data <- data %>%
  left_join(selected_sites %>% select(statid, station_name, code, river_name_sf),
            by = c("statid", "station_name"))
```

```
## rename code column
data <- data %>% rename(station_code = code)
glimpse(data)

# Save as new clean file to use
write_csv(data, "water_quality_data/wq_data_cleaned.csv")

# Filter for relevant years and dry season months
dry_season_years_data <- data %>%
  filter(year_collected %in% c(2011, 2013, 2015, 2017, 2019, 2021)) %>%
  filter(month_collected %in% c(12, 1, 2, 3))

# Save as relevant time series file to use
write_csv(dry_season_years_data, "water_quality_data/wq_data_relevant_year_month.csv")
```

Appendix G. Statistical Analysis Code

This code sets up the statistical analysis by merging and joining relevant layers; thus, it MUST be run before running any of the statistical analysis codes.
The code is from “dissfinal/code/appendix g stats.R”

```
# ---------------------------------------- STATISTICAL ANALYSIS ----------------------------------------------

# Install new libraries and load all libraries needed for statistical analysis and related plots
install.packages("car")
install.packages("corrplot")
install.packages("broom")
install.packages("ggplot2")
install.packages("ggrepel")
install.packages("gratia")
install.packages("mgcv")
install.packages("patchwork")
install.packages("RColorBrewer")
install.packages("vegan")
library(car) # Regression tools
library(corrplot) # Visualizes correlation matrices
library(broom) # Tidies model outputs into data frames
library(ggplot2) # Data visualization
library(ggrepel) # Label placement in ggplot2 plots
library(gratia) # Visualization and diagnostics for GAMs
library(mgcv) # Fits GAMs
library(patchwork) # Combines multiple plots into a layout
library(RColorBrewer) # Color palettes for plots
library(tidyverse)
library(vegan) # Multivariate analysis

# Load datasets
wq_data <- read_csv("water_quality_data/wq_data_cleaned.csv")
landcover_data <- read.csv("servir_landcover_output/wqm_landcover_percent_by_year.csv")
abundance <- read.csv("ecological_health_data/abundance.csv")
richness <- read.csv("ecological_health_data/richness.csv")
aspt <- read.csv("ecological_health_data/aspt.csv")
site_match <- read.csv("join_metadata/wq_corresponding_eh_final.csv")
predictor_site_match <- read.csv("join_metadata/wq_corresponding_eh.csv")

# Create folder for stats outputs
dir.create("statistical_analysis")
dir.create("figures")
dir.create("statistical_analysis/gams")
```

```
# ------------------------- PREP WQ DATA -------------------------
# Define numeric WQ variables (categorical/predictor variables are sites, land use type, and year)
water_quality_vars <- c(
  "temp_c", "ph", "tss_mgl", "cond_msm", "totn_mgl",
  "totp_mgl", "do_mgl", "codmn_mgl", "fc_mpn_100ml", "bod_mgl"
)

# Filter for dry season months (December, January, February & March) and take median value
dry_median <- wq_data %>%
  filter(month_collected %in% c(12, 1, 2, 3)) %>%
  filter(year_collected %in% c(2011, 2013, 2015, 2017, 2019, 2021)) %>%
  group_by(statid, year_collected, country_code) %>%
  summarise(across(all_of(water_quality_vars), median, na.rm = TRUE), .groups = "drop")

# Check missingness
missing_summary <- dry_median %>%
  summarise(across(all_of(water_quality_vars), ~mean(is.na(.)) * 100))

print(missing_summary)

# Drop heavily missing variables
reduced_vars <- setdiff(water_quality_vars, c("bod_mgl", "fc_mpn_100ml"))

# ------------------------- PREP EH DATA -------------------------

# Pivot eh data to long format
abund_long <- abundance %>%
  pivot_longer(cols = starts_with("X"), names_to = "year_collected", values_to = "abundance")
%>%
  mutate(
    year_collected = as.numeric(sub("X", "", year_collected)),
  ) %>%
  select(code, year_collected, abundance)

rich_long <- richness %>%
  pivot_longer(cols = starts_with("X"), names_to = "year_collected", values_to = "richness") %>%
  mutate(
    year_collected = as.numeric(sub("X", "", year_collected)),
  ) %>%
  select(code, year_collected, richness)

aspt_long <- aspt %>%
```

```
  pivot_longer(cols = starts_with("X"), names_to = "year_collected", values_to = "aspt") %>%
  mutate(
    year_collected = as.numeric(sub("X", "", year_collected)),
  ) %>%
  select(code, year_collected, aspt)

# Join eh metrics together
macro_metrics <- abund_long %>%
  left_join(rich_long, by = c("code", "year_collected")) %>%
  left_join(aspt_long, by = c("code", "year_collected"))

# -------------------------- JOIN WQ AND EH --------------------------

# Drop NA's and join data
wq_eh_merge <- site_match %>%
  left_join(select(dry_median, statid, year_collected),
            by = c("wq_statid" = "statid")) %>%
  left_join(macro_metrics, by = c("eh_code" = "code", "year_collected")) %>%
  left_join(select(dry_median, statid, year_collected, all_of(reduced_vars)),
            by = c("wq_statid" = "statid", "year_collected")) %>%
  filter(!is.na(aspt))

# Ensure joins worked cleanly
glimpse(wq_eh_merge)

# -------------------------- PREP LAND USE DATA --------------------------

# Aggregate land use categories
land_use <- landcover_data %>%
  mutate(
    forest = Class_5 + Class_6 + Class_7 + Class_10 + Class_11 + Class_18,
    agriculture = Class_3 + Class_4 + Class_12 + Class_13 + Class_14 + Class_1,
    urban = Class_16 + Class_2,
    semi_natural = Class_8 + Class_9 + Class_15 + Class_17
  ) %>%
  select(statid, Year, forest, agriculture, urban, semi_natural) %>%
  rename(year_collected = Year)

# -------------------------- JOIN WQ AND LU --------------------------

wq_lu_merge <- dry_median %>%
  left_join(land_use, by = c("statid", "year_collected")) %>%
  drop_na(forest, agriculture, urban, semi_natural)
```

```
# -------------------------- JOIN WQ, EH, LU --------------------------
wq_eh_lu_merge <- wq_eh_merge %>%
  left_join(select(land_use, statid, year_collected, urban, agriculture, forest),
            by = c("wq_statid" = "statid", "year_collected")) %>%
  select(year_collected, all_of(reduced_vars), aspt, richness, abundance, urban, agriculture,
forest, type) %>%
  drop_na()
```

Appendix H. Pearson's Correlation Code

Code from "dissfinal/code/appendix h pearsons.R"

```
# ------------------------------------- PEARSON'S CORRELATION ----------------------------------------

# Select numeric variables except 'year_collected'
numeric_vars <- wq_eh_lu_merge %>%
  select(where(is.numeric)) %>%
  select(-year_collected)

# Run Pearson's correlation
cor_matrix <- cor(numeric_vars, method = "pearson")

# Convert matrix to data frame
cor_df <- as.data.frame(cor_matrix)

# Write to CSV
write_csv(cor_df, "statistical_analysis/pearson_correlations.csv")

# Select only the predictor variables
selected_vars <- wq_eh_lu_merge %>%
  select(aspt, richness, abundance, urban, agriculture, forest)

# Run Pearson's correlation
cor_matrix <- cor(selected_vars, method = "pearson")

# Plot heatmap and save to pdf
pdf("figures/pearson_heatmap.pdf", width = 8, height = 8)

corrplot(cor_matrix, method = "color", type = "lower",
         tl.pos = "d",
         tl.cex = 0.8, tl.col = "black", number.cex = 0.7,
         addCoef.col = "black", col = colorRampPalette(c("red", "white", "blue"))(200),
         title = "Pearson correlation matrix", mar = c(0, 0, 2, 0))

dev.off()
```

Appendix I. Redundancy Analysis Code

Code from “dissfinal/code/appendix i rda.R”

```
# ---------------------------------------------------- LU RDA ----------------------------------------------------

# ---------------- LU RDA POOLED -----------------

# Select reduced water quality variables
wq_matrix <- wq_lu_merge %>%
  select(all_of(reduced_vars)) %>%
  drop_na()

# Remove na's
wq_lu_merge <- wq_lu_merge %>%
  drop_na(all_of(c(reduced_vars, "forest", "agriculture", "urban")))

# Select land use predictors
lu_vars <- c("urban", "agriculture", "forest")

# Select WQ and LU matrices from the same rows
wq_matrix <- wq_lu_merge %>% select(all_of(reduced_vars))
lu_matrix <- wq_lu_merge %>% select(all_of(lu_vars)) %>% drop_na()

# Run the RDA: WQ as response, LU as predictor with wq data normalized
wq_matrix_scaled <- scale(wq_matrix)
rda_model_scaled <- rda(wq_matrix_scaled ~ ., data = lu_matrix)

# Significance tests
anova(rda_model_scaled, permutations = 999)
anova(rda_model_scaled, by = "term", permutations = 999)  # Tests individual predictors

# Export the results
## Overall RDA test
rda_overall <- anova(rda_model_scaled, permutations = 999)
dir.create("statistical_analysis/rda_results")
write.csv(as.data.frame(rda_overall),
"statistical_analysis/rda_results/rda_lu_pooled_overall.csv")

## Term-specific RDA test
rda_terms <- anova(rda_model_scaled, by = "term", permutations = 999)
write.csv(as.data.frame(rda_terms), "statistical_analysis/rda_results/rda_lu_pooled_terms.csv")
```

```
# ---------------- LU RDA PLOT ----------------

# Extract scores for plotting
scores_rda <- scores(rda_model_scaled, display = c("species", "sites", "bp"), scaling = 2)

# Convert to data frames
species_scores <- as.data.frame(scores_rda$species) %>% rownames_to_column("WQ_Var")
biplot_scores <- as.data.frame(scores_rda$biplot) %>% rownames_to_column("Land_Use")

# Define metric labels
metric_labels <- c(
  "codmn_mgl" = "Chemical Oxygen Demand (mg/L)",
  "cond_msm" = "Electrical Conductivity (µS/cm)",
  "do_mgl" = "Dissolved Oxygen (mg/L)",
  "ph" = "pH",
  "temp_c" = "Temperature (C)",
  "totn_mgl" = "Total Nitrogen (mg/L)",
  "totp_mgl" = "Total Phosphorus (mg/L)",
  "tss_mgl" = "Total Suspended Solids (mg/L)"
)

# Apply metric_labels to species_scores
species_scores <- species_scores %>%
  mutate(Label = metric_labels[WQ_Var])

# Update ggplot text aesthetic
rda_biplot <- ggplot() +
  geom_segment(
    data = biplot_scores,
    aes(x = 0, y = 0, xend = RDA1, yend = RDA2, color = Land_Use),
    arrow = arrow(length = unit(0.2, "cm")), size = 2
  ) +
  geom_text_repel(
    data = species_scores,
    aes(x = RDA1, y = RDA2, label = Label),
    color = "darkblue", size = 3.5,
    max.overlaps = Inf
  ) +
  scale_color_manual(values = c(
    "urban" = "#E63946",
    "agriculture" = "#F4A261",
    "forest" = "#2A9D8F"
  )) +
  coord_equal() +
```

```
  labs(
    title = "RDA Biplot: Scaled WQ Variables and Land Use Predictors",
    x = "RDA 1",
    y = "RDA 2",
    color = "Land Use"
  ) +
  theme(
    plot.title = element_text(size = 16, face = "bold"),
    axis.title = element_text(size = 14),
    axis.text = element_text(size = 12),
    legend.title = element_text(size = 14),
    legend.text = element_text(size = 12)
  )

# Save the plot
ggsave("figures/rda_lu_pooled.png", rda_biplot, width = 10, height = 6, dpi = 300)

# ---------- LU RDA YEARLY ----------

# Select years
years <- c(2011, 2013, 2015, 2017, 2019, 2021)

# Run yearly RDA to loop through years
run_yearly_rda <- function(year) {
  subset_data <- wq_lu_merge %>% filter(year_collected == year)

  # Ensure complete cases across both sets of variables
  complete_rows <- complete.cases(subset_data[, c(reduced_vars, lu_vars)])
  subset_data <- subset_data[complete_rows, ]

  # Extract matrices
  wq_matrix <- subset_data %>% select(all_of(reduced_vars)) %>% scale()
  lu_matrix <- subset_data %>% select(all_of(lu_vars))

  # Escape if data is sparse
  if (nrow(lu_matrix) < 2 || ncol(lu_matrix) == 0 ||
      nrow(wq_matrix) < 2 || ncol(wq_matrix) == 0) {
    warning("Insufficient data for year: ", year)
    return(tibble(Year = year, R2 = NA, AdjR2 = NA, F = NA, p_value = NA))
  }

  tryCatch({
    rda_model <- rda(wq_matrix ~ ., data = lu_matrix)
```

```
    rsq <- RsquareAdj(rda_model)
    r2 <- rsq$r.squared
    adj_r2 <- rsq$adj.r.squared

    overall_test <- anova(rda_model, permutations = 999)
    term_tests <- anova(rda_model, by = "term", permutations = 999)
    term_df <- as.data.frame(term_tests) %>%
      mutate(Year = year, Term = rownames(term_tests))

    write.csv(as.data.frame(overall_test),
              paste0("statistical_analysis/rda_results/rda_lu_yearly_overall", year, ".csv"),
              row.names = FALSE)

    write.csv(term_df,
              paste0("statistical_analysis/rda_results/rda_lu_yearly_terms", year, ".csv"),
              row.names = FALSE)

    tibble(
      Year = year,
      R2 = r2,
      AdjR2 = adj_r2,
      F = overall_test$F[1],
      p_value = overall_test$`Pr(>F)`[1]
    )
  }, error = function(e) {
    warning("Error for year ", year, ": ", e$message)
    tibble(Year = year, R2 = NA, AdjR2 = NA, F = NA, p_value = NA)
  })
}

yearly_results <- purrr::map_dfr(years, run_yearly_rda)

# Save output summary
write.csv(yearly_results, "statistical_analysis/rda_results/rda_lu_yearly_summary.csv",
row.names = FALSE)

# ---------------------------------------------------- MACRO RDA ----------------------------------------------------

# ---------------- MACRO RDA POOLED -----------------

# Select reduced water quality variables
wq_matrix <- wq_eh_merge %>%
  select(all_of(reduced_vars)) %>%
  drop_na()
```

```
# Remove na's
wq_eh_merge <- wq_eh_merge %>%
  drop_na(all_of(c(reduced_vars, "richness", "abundance", "aspt")))

# Select land use predictors
macro_vars <- c("richness", "abundance", "aspt")

# Select WQ and EH matrices from the same rows
wq_matrix <- wq_eh_merge %>% select(all_of(reduced_vars))
eh_matrix <- wq_eh_merge %>% select(all_of(macro_vars)) %>% drop_na()

# Run the RDA: EH as response, WQ as predictor with wq data normalized
wq_matrix_scaled <- scale(wq_matrix)
rda_model_scaled <- rda(eh_matrix ~ ., data = as.data.frame(wq_matrix_scaled))

# Significance tests
anova(rda_model_scaled, permutations = 999)
anova(rda_model_scaled, by = "term", permutations = 999)  # Tests individual predictors

# Export the results
## Overall RDA test
rda_overall <- anova(rda_model_scaled, permutations = 999)
write.csv(as.data.frame(rda_overall),
"statistical_analysis/rda_results/rda_macro_pooled_overall.csv")

## Term-specific RDA test
rda_terms <- anova(rda_model_scaled, by = "term", permutations = 999)
write.csv(as.data.frame(rda_terms),
"statistical_analysis/rda_results/rda_macro_pooled_terms.csv")

# ---------- MACRO RDA PLOT ----------

# Extract scores for plotting
scores_rda <- scores(rda_model_scaled, display = c("species", "sites", "bp"), scaling = 2)

# Convert to data frames
species_scores <- as.data.frame(scores_rda$species) %>% rownames_to_column("Metric")
biplot_scores <- as.data.frame(scores_rda$biplot) %>% rownames_to_column("WQ_VAR")

# Define plot labels and colors
wq_metric_labels <- c(
  "codmn_mgl" = "COD",
  "cond_msm" = "Conductivity",
  "do_mgl" = "DO",
```

```
  "ph" = "pH",
  "temp_c" = "Temp",
  "totn_mgl" = "TN",
  "totp_mgl" = "TP",
  "tss_mgl" = "TSS"
)

wq_colors <- c(
  "codmn_mgl" = "#1b9e77",
  "cond_msm" = "#d95f02",
  "do_mgl" = "#7570b3",
  "ph" = "#e7298a",
  "temp_c" = "#66a61e",
  "totn_mgl" = "#e6ab02",
  "totp_mgl" = "#a6761d",
  "tss_mgl" = "#666666"
)

metric_labels <- c(
  "aspt" = "ASPT",
  "abundance" = "Abundance",
  "richness" = "Richness"
)

# Apply metric_labels to species_scores
species_scores <- species_scores %>%
  mutate(Label = metric_labels[Metric])

# Make arrows larger
stretch_factor <- 20

biplot_scores_stretched <- biplot_scores %>%
  mutate(RDA1 = RDA1 * stretch_factor,
         RDA2 = RDA2 * stretch_factor)

# Plot biplot
rda_biplot <- ggplot() +
  geom_segment(
    data = biplot_scores_stretched,
    aes(x = 0, y = 0, xend = RDA1, yend = RDA2, color = WQ_VAR),
    arrow = arrow(length = unit(0.2, "cm")), size = 2
  ) +
  geom_text_repel(
    data = species_scores,
```

```
  aes(x = RDA1, y = RDA2, label = Label),
  color = "darkblue", size = 7,
  max.overlaps = Inf
 ) +
 scale_color_manual(
  values = wq_colors,
  labels = wq_metric_labels,
  breaks = names(wq_metric_labels),
  name = "Water Quality"
 ) +
 coord_equal() +
 labs(
  title = "RDA Biplot: Scaled WQ Variables and Macroinvertebrate Responses",
  x = "RDA 1",
  y = "RDA 2",
  color = "Water Quality Parameters"
 ) +
 theme(
  plot.title = element_text(size = 22, face = "bold"),
  axis.title = element_text(size = 20),
  axis.text = element_text(size = 16),
  legend.title = element_text(size = 20),
  legend.position = "bottom",
  legend.text = element_text(size = 18)
 )

print(rda_biplot)

# Save plot
ggsave("figures/rda_macro_pooled.png", rda_biplot, width = 10, height = 6, dpi = 300)
```

Appendix J. Multiple Linear Regression Code

Code from “dissfinal/code/appendix j mlr.R”

```
# ---------------------------- MACRO MULTIPLE LINEAR REGRESSION ---------------------------------------

# ---------------- MACRO MLR POOLED -----------------

# Define macro predictor variables
macro_vars <- c("aspt", "abundance", "richness")

# Initialize list to store results
mlr_results <- list()

# Loop through each WQ variable in reduced_vars
for (wq_var in reduced_vars) {
  # Create modeling dataset
  model_df <- wq_eh_merge %>%
    select(all_of(c(wq_var, macro_vars))) %>%
    drop_na()

  # Build formula and fit model
  mlr_formula <- as.formula(paste(wq_var, "~", paste(macro_vars, collapse = " + ")))
  mlr_model <- lm(mlr_formula, data = model_df)

  # Check for colinearity and print results
  vif_values <- vif(mlr_model)
  print(paste("VIFs for", wq_var))
  print(vif_values)

  # Tidy output and add WQ variable as ID
  mlr_output <- tidy(mlr_model) %>%
    mutate(response_variable = wq_var)

  # Store in list
  mlr_results[[wq_var]] <- mlr_output
}

# Combine all outputs into a single dataframe
mlr_all_output <- bind_rows(mlr_results)

# Save to CSV
dir.create("statistical_analysis/mlr_macro")
```

```
write.csv(mlr_all_output, "statistical_analysis/mlr_macro/mlr_macro_pooled.csv", row.names =
FALSE)

# ---------- MACRO MLR YEARLY ----------

# Define macro predictor variables
macro_vars <- c("aspt", "abundance", "richness")

# Define years of interest
years <- c(2011, 2013, 2015, 2017, 2019, 2021)

# Initialize list to store all results
mlr_yearly_results <- list()

# Loop through each year
for (yr in years) {
  # Filter data for the year
  yearly_data <- wq_eh_merge %>% filter(year_collected == yr)

  # Loop through each WQ variable
  for (wq_var in reduced_vars) {
    model_df <- yearly_data %>%
      select(all_of(c(wq_var, macro_vars))) %>%
      drop_na()
    # Skip model if not enough data
    if (nrow(model_df) < 3) next
    # Create formula and fit model
    mlr_formula <- as.formula(paste(wq_var, "~", paste(macro_vars, collapse = " + ")))
    mlr_model <- lm(mlr_formula, data = model_df)
    # Tidy output and add identifiers
    mlr_output <- tidy(mlr_model) %>%
      mutate(
        response_variable = wq_var,
        year = yr
      )
    # Store in list
    mlr_yearly_results[[paste(wq_var, yr, sep = "_")]] <- mlr_output
  }
}

# Combine and export
mlr_yearly_output <- bind_rows(mlr_yearly_results)
write.csv(mlr_yearly_output, "statistical_analysis/mlr_macro/mlr_macro_yearly.csv", row.names
= FALSE)
```

Appendix K. Macroinvertebrate Predictor GAM Code

Code for all GAMs is available in this appendix. Code for each predictor set was as separate R files and can be found in “dissfinal/code/appendix k macro gam.R”, “dissfinal/code/appendix k lu gam.R”, “dissfinal/code/appendix k full gam.R” and “dissfinal/code/appendix k mvt gam.R”. Code for figures has been removed from this appendix due to space constraints but is available in corresponding R files (i.e. code for figures related to Macro GAMs would be found in “dissfinal/code/appendix k macro gam.R”).

```
# ---------------------------------------------------- MACRO GAM ----------------------------------------------------

# ---------------- MACRO GAM POOLED -----------------

# Function to run GAM and extract results
run_gam_model <- function(response_var) {
  formula <- as.formula(paste(response_var, "~ s(aspt) + s(richness) + s(abundance)"))

  # Fit GAM
  model <- gam(formula, data = wq_eh_lu_merge, method = "REML")

  # Summary stats
  gam_summary <- summary(model)
  smooth_terms <- gam_summary$s.table

  tibble(
    Response = response_var,
    Adj_R2 = gam_summary$r.sq,
    Dev_Explained = gam_summary$dev.expl * 100,
    ASPT_p = smooth_terms["s(aspt)", "p-value"],
    Richness_p = smooth_terms["s(richness)", "p-value"],
    Abundance_p = smooth_terms["s(abundance)", "p-value"]
  )
}

# Fit a GAM model
gam_model <- gam(totp_mgl ~ s(aspt) + s(richness) + s(abundance),
                 data = wq_eh_lu_merge, method = "REML")

# Run across all WQ variables
gam_macro_pooled <- map_dfr(reduced_vars, run_gam_model)

# Save results
write.csv(gam_macro_pooled, "statistical_analysis/gams/gam_macro_pool.csv", row.names =
FALSE)
```

```
# ---------------- MACRO GAM YEARLY -----------------

# Function to loop through all years
run_gam_by_year <- function(year, response_var) {
  data_subset <- wq_eh_lu_merge %>%
    filter(year_collected == year) %>%
    select(aspt, richness, abundance, all_of(response_var)) %>%
    drop_na()

  if (nrow(data_subset) < 10) return(NULL)

  formula <- as.formula(paste(response_var, "~ s(aspt) + s(richness) + s(abundance)"))
  model <- gam(formula, data = data_subset, method = "REML")
  summary_model <- summary(model)

  tibble(
    Year = year,
    Response = response_var,
    Adj_R2 = summary_model$r.sq,
    Dev_Explained = summary_model$dev.expl * 100,
    ASPT_p = summary_model$s.table["s(aspt)", "p-value"],
    Richness_p = summary_model$s.table["s(richness)", "p-value"],
    Abundance_p = summary_model$s.table["s(abundance)", "p-value"]
  )
}

# Loop across years and WQ variables
gam_macro_by_year <- map_dfr(unique(wq_eh_lu_merge$year_collected), function(y) {
  map_dfr(reduced_vars, ~run_gam_by_year(y, .x))
})

# Save
write.csv(gam_macro_by_year, "statistical_analysis/gams/gam_macro_yearly.csv", row.names =
FALSE)

# ---------------------------------- MACRO GAM W/O RICHNESS ----------------------------------------

# Set up model
compare_gam_models <- function(response_var, data = wq_eh_lu_merge) {
  # Full model with all predictors
  full_formula <- as.formula(paste(response_var, "~ s(aspt) + s(richness) + s(abundance)"))
  full_model <- gam(full_formula, data = data, method = "REML")
  full_summary <- summary(full_model)
```

```
  # Reduced model without richness
  reduced_formula <- as.formula(paste(response_var, "~ s(aspt) + s(abundance)"))
  reduced_model <- gam(reduced_formula, data = data, method = "REML")
  reduced_summary <- summary(reduced_model)

  # Combine results
  tibble(
    Response = response_var,
    Full_Adj_R2 = full_summary$r.sq,
    Full_Dev_Explained = full_summary$dev.expl * 100,
    Reduced_Adj_R2 = reduced_summary$r.sq,
    Reduced_Dev_Explained = reduced_summary$dev.expl * 100,
    Δ_Adj_R2 = round(reduced_summary$r.sq - full_summary$r.sq, 3),
    Δ_Dev_Explained = round((reduced_summary$dev.expl - full_summary$dev.expl) * 100, 2)
  )
}

# Apply across all reduced_vars
gam_comparison_results <- map_dfr(reduced_vars, ~compare_gam_models(.x))

# View or export
print(gam_comparison_results)
write.csv(gam_comparison_results, "statistical_analysis/gams/gam_macro_no_richness.csv",
row.names = FALSE)

# ----------------------------------------------------. LU GAM ----------------------------------------------------

# ---------------- LU GAM POOLED -----------------

# Define land use predictors
lu_vars <- c("urban", "agriculture", "forest")

# GAM function
run_gam_lu <- function(response_var) {
  formula <- as.formula(paste(response_var, "~", paste0("s(", lu_vars, ")", collapse = " + ")))

  model <- gam(formula, data = wq_eh_lu_merge, method = "REML")
  summary_model <- summary(model)

  term_table <- summary_model$s.table

  tibble(
    Response = response_var,
    Adj_R2 = summary_model$r.sq,
```

```
    Dev_Explained = summary_model$dev.expl * 100,
    Urban_p = term_table["s(urban)", "p-value"],
    Agriculture_p = term_table["s(agriculture)", "p-value"],
    Forest_p = term_table["s(forest)", "p-value"]
  )
}

# Apply GAM function across multiple response variables and combine output into a single tibble
gam_lu_pooled <- map_dfr(reduced_vars, run_gam_lu)

# Save
write.csv(gam_lu_pooled, "statistical_analysis/gams/gam_lu_pool.csv", row.names = FALSE)

# ---------- LU GAM YEARLY ----------

# Loop through each year
run_gam_lu_by_year <- function(year, response_var) {
  data_subset <- wq_eh_lu_merge %>%
    filter(year_collected == year) %>%
    select(all_of(c(response_var, lu_vars))) %>%
    drop_na()

  if (nrow(data_subset) < 10) return(NULL)

  formula <- as.formula(paste(response_var, "~", paste0("s(", lu_vars, ")", collapse = " + ")))
  model <- gam(formula, data = data_subset, method = "REML")
  summary_model <- summary(model)

  tibble(
    Year = year,
    Response = response_var,
    Adj_R2 = summary_model$r.sq,
    Dev_Explained = summary_model$dev.expl * 100,
    Urban_p = summary_model$s.table["s(urban)", "p-value"],
    Agriculture_p = summary_model$s.table["s(agriculture)", "p-value"],
    Forest_p = summary_model$s.table["s(forest)", "p-value"]
  )
}

# loop across all years / variables
gam_lu_by_year <- map_dfr(unique(wq_eh_lu_merge$year_collected), function(y) {
  map_dfr(reduced_vars, ~run_gam_lu_by_year(y, .x))
})
```

```
# Save
write.csv(gam_lu_by_year, "statistical_analysis/gams/gam_lu_yearly.csv", row.names = FALSE)


# ---------------------------------------------------- FULL GAM ----------------------------------------------------

# ------------------------- FULL GAM POOLED -------------------------

# Define predictors
eh_vars <- c("aspt", "abundance", "richness")
lu_vars <- c("urban", "agriculture", "forest")
all_predictors <- c(eh_vars, lu_vars)

# Fit gam model to return model object (necessary for predicting sites)
fit_gam_model <- function(response_var) {
  formula <- as.formula(paste(response_var, "~", paste0("s(", all_predictors, ")", collapse = " + ")))
  gam(formula, data = wq_eh_lu_merge, method = "REML")
}

# Summarise model performance
summarise_gam_model <- function(model, response_var) {
  summary_model <- summary(model)
  term_table <- summary_model$s.table

  tibble(
    Response = response_var,
    Adj_R2 = summary_model$r.sq,
    Dev_Explained = summary_model$dev.expl * 100,
    ASPT_p = term_table["s(aspt)", "p-value"],
    Richness_p = term_table["s(richness)", "p-value"],
    Abundance_p = term_table["s(abundance)", "p-value"],
    Urban_p = term_table["s(urban)", "p-value"],
    Agriculture_p = term_table["s(agriculture)", "p-value"],
    Forest_p = term_table["s(forest)", "p-value"]
  )
}

# Run model and export summary
gam_pooled_results <- map_dfr(reduced_vars, function(response_var) {
  model <- fit_gam_model(response_var)
  summarise_gam_model(model, response_var)
})
```

```
# Save
write.csv(gam_pooled_results, "statistical_analysis/gams/gam_full_pooled.csv", row.names =
FALSE)

# -------------------------- PREDICT WQ FROM FULL MODEL -------------------------

# Rerun joins with predictor_site_match
# ------ Join WQ and EH
# Drop NA's and join data
wq_eh_merge_predictor <- predictor_site_match %>%
  left_join(select(dry_median, statid, year_collected),
            by = c("wq_statid" = "statid")) %>%
  left_join(macro_metrics, by = c("eh_code" = "code", "year_collected")) %>%
  left_join(select(dry_median, statid, year_collected, all_of(reduced_vars)),
            by = c("wq_statid" = "statid", "year_collected")) %>%
  filter(!is.na(aspt))

# ----- Join WQ and LU
wq_lu_merge_predictor <- dry_median %>%
  left_join(land_use, by = c("statid", "year_collected")) %>%
  drop_na(forest, agriculture, urban, semi_natural)

# ----- Join WQ, EH, LU
wq_eh_lu_merge_predictor <- wq_eh_merge_predictor %>%
  left_join(select(land_use, statid, year_collected, urban, agriculture, forest),
            by = c("wq_statid" = "statid", "year_collected")) %>%
  select(wq_statid, year_collected, all_of(reduced_vars), aspt, richness, abundance, urban,
agriculture, forest, type) %>%
  drop_na()

# Define withheld sites
model_sites <- c("H011200", "H013401", "H019806", "H910108")

# Filter data for prediction
new_sites_data <- wq_eh_lu_merge_predictor %>% filter(wq_statid %in% model_sites)

# Predict loop
prediction_results <- map_dfr(reduced_vars, function(response_var) {
  model <- fit_gam_model(response_var)

  preds <- predict(model, newdata = new_sites_data, type = "response")

  tibble(
    wq_statid = new_sites_data$wq_statid,
```

```
    year_collected = new_sites_data$year_collected,
    Response = response_var,
    Prediction = preds,
    Actual = new_sites_data[[response_var]]
  )
})

# Save predictions
write.csv(prediction_results, "statistical_analysis/gams/gam_full_predictions.csv", row.names =
FALSE)

# Calculate residual
prediction_results <- prediction_results %>%
  mutate(Residual = Actual - Prediction)

# Calculate RMSE per metric
rmse_summary <- prediction_results %>%
  group_by(Response) %>%
  summarise(
    RMSE = sqrt(mean(Residual^2, na.rm = TRUE)),
    Mean_Residual = mean(Residual, na.rm = TRUE),
    SD_Residual = sd(Residual, na.rm = TRUE),
    .groups = "drop"
  )

# View RMSE per metric and save
print(rmse_summary)
write.csv(rmse_summary, "statistical_analysis/gams/gam_full_predictions_rmse.csv",
row.names = FALSE)

# ------------------------- EXAMINE OUTLIER SITES -------------------------

# Rerun EH, WQ, LU join once more with statid and eh_code
wq_eh_lu_merge_outlier <- wq_eh_merge %>%
  left_join(select(land_use, statid, year_collected, urban, agriculture, forest),
            by = c("wq_statid" = "statid", "year_collected")) %>%
  select(wq_statid, eh_code, year_collected, all_of(reduced_vars), aspt, richness, abundance,
urban, agriculture, forest, type) %>%
  drop_na()

# Define predictors and metrics
eh_vars <- c("aspt", "abundance", "richness")
lu_vars <- c("urban", "agriculture", "forest")
all_predictors <- c(eh_vars, lu_vars)
```

```
wq_metrics <- c("temp_c", "ph", "tss_mgl", "cond_msm",
                "totn_mgl", "totp_mgl", "do_mgl", "codmn_mgl")

# Extract outliers
check_combined_outliers <- function(response_var) {
  formula <- as.formula(paste(response_var, "~", paste0("s(", all_predictors, ")", collapse = " + ")))
  model <- gam(formula, data = wq_eh_lu_merge_outlier, method = "REML")

  preds <- predict(model, type = "response", se.fit = TRUE)
  upper <- preds$fit + (1.96 * preds$se.fit)
  lower <- preds$fit - (1.96 * preds$se.fit)

  wq_eh_lu_merge_outlier %>%
    mutate(response = .data[[response_var]],
           predicted = preds$fit,
           lower_CI = lower,
           upper_CI = upper,
           outside_CI = response < lower_CI | response > upper_CI,
           metric = response_var) %>%
    select(eh_code, year_collected, metric, response, predicted, lower_CI, upper_CI, outside_CI)
}

# Run accross metrics and combine
eh_lu_outliers <- map_dfr(wq_metrics, check_combined_outliers)

# Summarize per site
eh_lu_site_summary <- eh_lu_outliers %>%
  group_by(eh_code) %>%
  summarise(
    total_outliers = sum(outside_CI, na.rm = TRUE),
    total_checks = n(),
    outlier_pct = round((total_outliers / total_checks) * 100, 1),
    .groups = "drop"
  ) %>%
  arrange(desc(outlier_pct))

# Save
write.csv(eh_lu_site_summary, "statistical_analysis/gams/gam_full_outliers.csv", row.names =
FALSE)

# Classify outliers by zone
zonation <- read_csv("join_metadata/wq_corresponding_eh_final.csv")  # contains eh_code,
zone, river_system
```

```
eh_lu_site_summary_zoned <- eh_lu_site_summary %>%
  left_join(zonation, by = "eh_code")

# Summarize outliers by zone
zone_summary <- eh_lu_site_summary_zoned %>%
  group_by(type) %>%
  summarise(
    avg_outlier_pct = round(mean(outlier_pct, na.rm = TRUE), 1),
    high_outlier_sites = sum(outlier_pct > 70),
    total_sites = n()
  )

# Save results
write.csv(zone_summary, "statistical_analysis/gams/gam_full_outliers_zone.csv")

# Summarize outliers by river system
river_summary <- eh_lu_site_summary_zoned %>%
  group_by(wq_River_Names_og) %>%
  summarise(
    avg_outlier_pct = round(mean(outlier_pct, na.rm = TRUE), 1),
    high_outlier_sites = sum(outlier_pct > 70),
    total_sites = n()
  )

# Save results
write.csv(river_summary, "statistical_analysis/gams/gam_full_outliers_river.csv")

# Summarize outliers by spatial bins
coords_summary <- eh_lu_site_summary_zoned %>%
  mutate(
    lat_bin = round(wq_lat, 1),
    lon_bin = round(wq_long, 1)
  ) %>%
  group_by(lat_bin, lon_bin) %>%
  summarise(
    avg_outlier_pct = mean(outlier_pct, na.rm = TRUE),
    total_sites = n(),
    .groups = "drop"
  )

# Save results
write.csv(coords_summary, "statistical_analysis/gams/gam_full_outlier_bins.csv")
```

```
# -------------------------- FULL GAM YEARLY -------------------------

# RUN !!
run_gam_wq_eh_lu_by_year <- function(year, response_var) {
  data_subset <- wq_eh_lu_merge %>%
    filter(year_collected == year) %>%
    select(all_of(c(response_var, all_predictors))) %>%
    drop_na()

  if (nrow(data_subset) < 10) return(NULL)

  formula <- as.formula(paste(response_var, "~", paste0("s(", all_predictors, ")", collapse = " + ")))
  model <- gam(formula, data = data_subset, method = "REML")
  summary_model <- summary(model)

  term_table <- summary_model$s.table

  tibble(
    Year = year,
    Response = response_var,
    Adj_R2 = summary_model$r.sq,
    Dev_Explained = summary_model$dev.expl * 100,
    ASPT_p = term_table["s(aspt)", "p-value"],
    Richness_p = term_table["s(richness)", "p-value"],
    Abundance_p = term_table["s(abundance)", "p-value"],
    Urban_p = term_table["s(urban)", "p-value"],
    Agriculture_p = term_table["s(agriculture)", "p-value"],
    Forest_p = term_table["s(forest)", "p-value"]
  )
}

gam_wq_eh_lu_by_year <- map_dfr(unique(wq_eh_lu_merge$year_collected), function(y) {
  map_dfr(reduced_vars, ~run_gam_wq_eh_lu_by_year(y, .x))
})

# Save
write.csv(gam_wq_eh_lu_by_year, "statistical_analysis/gams/gam_full_yearly.csv", row.names =
FALSE)

# -------------------------- FULL GAM MAINSTEM V TRIBUTARY -----------------------------------------

# Sort rivers into mainstem or tributary
mainstem_rivers <- c("Mekong River", "Mekong  River")
tributary_rivers <- c("Sekong River ", "Se San River", "Srepork", "Se Bangfai River",
```

```
                "Nam Ou River", "Mae Kok River", "Mun", "Song Khram River",
                "Tonle Sap Lake","Tonle Sap River", "Bassac River")

# same merge as in section 1 with additional column
wq_eh_lu_merge <- wq_eh_merge %>%
  left_join(select(land_use, statid, year_collected, urban, agriculture, forest),
          by = c("wq_statid" = "statid", "year_collected")) %>%
  select(wq_statid, year_collected, all_of(reduced_vars), wq_River_Names_og, aspt, richness,
abundance, urban, agriculture, forest, type) %>%
  drop_na()

# Create new column based on river name
wq_eh_lu_merge$site_type <- ifelse(wq_eh_lu_merge$wq_River_Names_og %in%
mainstem_rivers,
                           "mainstem",
                           "tributary")

# Subset data into site groups
mainstem_data <- filter(wq_eh_lu_merge, site_type == "mainstem")
tributary_data <- filter(wq_eh_lu_merge, site_type == "tributary")

# Extract statid-year pairs from each data subset
mainstem_keys <- mainstem_data %>%
  select(wq_statid, year_collected)
tributary_keys <- tributary_data %>%
  select(wq_statid, year_collected)

# Choose predictors
lu_vars <- c("urban", "agriculture", "forest")
eh_vars <- c("aspt", "abundance", "richness")
all_predictors <- c(eh_vars, lu_vars)

# Mainstem
wq_main <- mainstem_data %>% select(all_of(reduced_vars))
predictors_main <- mainstem_data %>% select(all_of(all_predictors))

# Tributary
wq_trib <- tributary_data %>% select(all_of(reduced_vars))
predictors_trib <- tributary_data %>% select(all_of(all_predictors))

# ---------- GAM POOLED M v. T ----------

# mainstem
run_gam_pooled_main <- function(response_var) {
```

```
  formula <- as.formula(paste(response_var, "~", paste0("s(", colnames(predictors_main), ")",
collapse = " + ")))

  model <- gam(formula, data = mainstem_data, method = "REML")
  summary_model <- summary(model)
  term_table <- summary_model$s.table

  tibble(
    Response = response_var,
    Adj_R2 = summary_model$r.sq,
    Dev_Explained = summary_model$dev.expl * 100,
    ASPT_p = term_table["s(aspt)", "p-value"],
    Richness_p = term_table["s(richness)", "p-value"],
    Abundance_p = term_table["s(abundance)", "p-value"],
    Urban_p = term_table["s(urban)", "p-value"],
    Agriculture_p = term_table["s(agriculture)", "p-value"],
    Forest_p = term_table["s(forest)", "p-value"]
  )
}

# trib
run_gam_pooled_trib <- function(response_var) {
  formula <- as.formula(paste(response_var, "~", paste0("s(", colnames(predictors_trib), ")",
collapse = " + ")))

  model <- gam(formula, data = tributary_data, method = "REML")
  summary_model <- summary(model)
  term_table <- summary_model$s.table

  tibble(
    Response = response_var,
    Adj_R2 = summary_model$r.sq,
    Dev_Explained = summary_model$dev.expl * 100,
    ASPT_p = term_table["s(aspt)", "p-value"],
    Richness_p = term_table["s(richness)", "p-value"],
    Abundance_p = term_table["s(abundance)", "p-value"],
    Urban_p = term_table["s(urban)", "p-value"],
    Agriculture_p = term_table["s(agriculture)", "p-value"],
    Forest_p = term_table["s(forest)", "p-value"]
  )
}

# Run for all WQ variables
gam_pooled_main <- map_dfr(reduced_vars, run_gam_pooled_main)
```

```
gam_pooled_trib <- map_dfr(reduced_vars, run_gam_pooled_trib)

# Save results
write.csv(gam_pooled_main, "statistical_analysis/gams/gam_pooled_mainstem.csv", row.names
= FALSE)
write.csv(gam_pooled_trib, "statistical_analysis/gams/gam_pooled_triburaty.csv", row.names =
FALSE)
```

Appendix L. Variance Partitioning Code

Code from “dissfinal/code/appendix l vp.R”

```
# ---------------------------------------- VARIANCE PARTITIONING ----------------------------------------

# ---------------- VARIANCE PARTITIONING POOLED -----------------

# Define matrices
wq_matrix_scaled <- wq_eh_lu_merge %>% select(all_of(reduced_vars)) %>% scale()
landuse_block <- wq_eh_lu_merge %>% select(urban, agriculture, forest) # Predictor set 1: LU
eh_block <- wq_eh_lu_merge %>% select(aspt, richness, abundance) # Predictor set 2: Macros

# Run variance partitioning
vp_result <- varpart(wq_matrix_scaled, landuse_block, eh_block)

# Fit partial RDA models for permutation testing
rda_land_partial <- rda(wq_matrix_scaled, landuse_block, eh_block)  # Land use conditioned on macro
rda_eh_partial <- rda(wq_matrix_scaled, eh_block, landuse_block)    # Macro conditioned on land use

# Permutation tests for unique fractions
anova_land <- anova(rda_land_partial, permutations = 999)
anova_eh <- anova(rda_eh_partial, permutations = 999)

# Export results
## Variance fractions
vp_fractions <- as.data.frame(vp_result$part$indfract)
dir.create("statistical_analysis/vp")
write.csv(vp_fractions, "statistical_analysis/vp/vp_pooled_fractions.csv", row.names = TRUE)

## Significance tests for unique fractions
write.csv(as.data.frame(anova_land), "statistical_analysis/vp/vp_pooled_unique_landuse.csv",
row.names = FALSE)
write.csv(as.data.frame(anova_eh),
"statistical_analysis/vp/vp_pooled_unique_macroinverts.csv", row.names = FALSE)

# ---------------- VARIANCE PARTITIONING YEARLY -----------------

# Define variables
reduced_vars <- c("temp_c", "ph", "tss_mgl", "cond_msm", "totn_mgl", "totp_mgl", "do_mgl",
"codmn_mgl")
lu_vars <- c("urban", "agriculture", "forest")
```

```
eh_vars <- c("aspt", "richness", "abundance")
years <- c(2011, 2013, 2015, 2017, 2019, 2021)

# Function to run VP per year
run_variance_partitioning <- function(year) {
  data_subset <- wq_eh_lu_merge %>% filter(year_collected == year)

  # Define all required variables
  vars_needed <- c(reduced_vars, lu_vars, eh_vars)
  vars_available <- intersect(vars_needed, colnames(data_subset))

  # If no variables are available at all, exit early
  if (length(vars_available) == 0) {
    warning("No predictor variables found for year: ", year)
    return(tibble(
      Year = year,
      LU_Unique = NA,
      EH_Unique = NA,
      Shared = NA,
      Residual = NA,
      Total_Explained = NA
    ))
  }

  # Filter to complete rows only (only among available vars)
  data_subset <- data_subset %>% drop_na(all_of(vars_available))

  if (nrow(data_subset) < 5) {
    warning("Too few complete rows for year: ", year)
    return(tibble(
      Year = year,
      LU_Unique = NA,
      EH_Unique = NA,
      Shared = NA,
      Residual = NA,
      Total_Explained = NA
    ))
  }

  print(paste("Running year:", year, "with", nrow(data_subset), "complete cases"))

  # Define predictor sets
  wq_vars <- intersect(reduced_vars, colnames(data_subset))
  lu_vars_present <- intersect(lu_vars, colnames(data_subset))
```

```
eh_vars_present <- intersect(eh_vars, colnames(data_subset))

# Double-check blocks are non-empty
if (length(wq_vars) == 0 || length(lu_vars_present) == 0 || length(eh_vars_present) == 0) {
  warning("Missing variable block(s) for year: ", year)
  return(tibble(
    Year = year,
    LU_Unique = NA,
    EH_Unique = NA,
    Shared = NA,
    Residual = NA,
    Total_Explained = NA
  ))
}

print("Available columns:")
print(colnames(data_subset))

print("WQ vars used:")
print(wq_vars)

print("LU vars used:")
print(lu_vars_present)

print("EH vars used:")
print(eh_vars_present)

# Build matrices
wq_scaled <- scale(data_subset[, wq_vars, drop = FALSE])
lu_block  <- data_subset[, lu_vars_present, drop = FALSE]
eh_block  <- data_subset[, eh_vars_present, drop = FALSE]

# Run variance partitioning
vp_result <- tryCatch({
  varpart(wq_scaled, lu_block, eh_block)
}, error = function(e) {
  warning("varpart() failed for year: ", year)
  return(NULL)
})

if (is.null(vp_result)) {
  return(tibble(
    Year = year,
    LU_Unique = NA,
```

```
    EH_Unique = NA,
    Shared = NA,
    Residual = NA,
    Total_Explained = NA
  ))
}

# Fit partial RDAs for significance testing
rda_land <- rda(wq_scaled, lu_block, eh_block)
rda_eh   <- rda(wq_scaled, eh_block, lu_block)

# Export results
write.csv(as.data.frame(vp_result$part$indfract),
          paste0("statistical_analysis/vp/vp_fractions_", year, ".csv"), row.names = TRUE)

write.csv(as.data.frame(anova(rda_land, permutations = 999)),
          paste0("statistical_analysis/vp/vp_landuse_", year, ".csv"), row.names = FALSE)

write.csv(as.data.frame(anova(rda_eh, permutations = 999)),
          paste0("statistical_analysis/vp/vp_macros_", year, ".csv"), row.names = FALSE)

# Extract fractions safely
adj_r2 <- vp_result$part$indfract$Adj.R.squared
names(adj_r2) <- c("[a]", "[b]", "[ab]", "Residual")

tibble(
  Year = year,
  LU_Unique = adj_r2["[a]"],
  EH_Unique = adj_r2["[b]"],
  Shared    = adj_r2["[ab]"],
  Residual  = adj_r2["Residual"],
  Total_Explained = sum(adj_r2[c("[a]", "[b]", "[ab]")], na.rm = TRUE)
)
}

# Run across all years
vp_yearly_results <- map_dfr(years, run_variance_partitioning)

# Export yearly summary
write.csv(vp_yearly_results, "statistical_analysis/vp/vp_yearly_summary.csv", row.names =
FALSE)
```

Appendix M. Outlier Analysis for Full GAM

Table 4: Summary of site-level outlier detection across full GAM model integrating macroinvertebrate and land use predictors. *Results indicate the total number and proportion of response variable observations falling outside the 95% confidence interval per ecological health site, based on GAM predictions fitted to eight water quality metrics using ASPT, richness, abundance, and land use variables. The column "total_outliers" represents the number of water quality observations per site that fall outside the 95% confidence intervals predicted by the GAM models, while "total_checks" indicates the total number of observations assessed for outlier detection at each site.*

eh_code	total_outliers	total_checks	outlier_pct
LMX	41	48	85.4
CSP	36	48	75
VKB	35	48	72.9
TCS	34	48	70.8
LPB	28	40	70
CPT	32	48	66.7
LBH	32	48	66.7
TSM	31	48	64.6
LVT	30	48	62.5
CTU	29	48	60.4
LSD	29	48	60.4
TMU	29	48	60.4
CKT	28	48	58.3
CSK	28	48	58.3
CPP	27	48	56.2
CMR	26	48	54.2
LDN	26	48	54.2
VDP	26	48	54.2
TKC	25	48	52.1
CSJ	23	48	47.9
CKK	21	48	43.8
CKL	21	48	43.8
VCT	21	48	43.8
VLX	19	48	39.6
TNP	17	48	35.4
VCL	17	48	35.4
CCK	14	40	35
TKO	16	48	33.3
VTT	16	48	33.3
CNL	15	48	31.2
VVL	13	48	27.1
VTP	11	48	22.9

Table 5: Spatial bins of outlier prevalence across sites. *Summary of average outlier percentages by rounded latitude and longitude bins, based on site-level assessments of water quality response variables using full GAM model. Each bin represents the mean percentage of outliers detected across all sites within its spatial extent.*

	lat_bin	lon_bin	avg_outlier_pct	total_sites
1	10.7	105.1	43.8	1
2	10.8	105.2	27.1	1
3	10.8	105.3	46.9	2
4	10.9	105.5	34.35	2
5	11.1	105.1	72.9	1
6	11.1	105.2	22.9	1
7	11.3	105	43.8	1
8	11.6	104.9	43.7	2
9	11.8	104.8	60.4	1
10	12.5	106	66.7	1
11	12.6	104.2	43.8	1
12	13.3	103.4	58.3	1
13	13.3	103.8	35	1
14	13.5	106	58.3	1
15	13.5	106.5	47.9	1
16	13.8	107.4	75	1
17	14.1	106.4	54.2	1
18	14.2	107.8	75	1
19	15.1	105.8	57.3	2
20	15.3	105.5	56.25	2
21	17.1	105	66.7	1
22	17.4	104.8	35.4	1
23	17.6	104.5	64.6	1
24	18	102.6	62.5	1
25	19.9	99.8	33.3	1
26	20.1	102.3	70	1
27	20.3	100.1	70.8	1
28	21.5	101.2	85.4	1

Table 6: Summary of outlier patterns in GAM residuals, stratified by zone type (mainstem vs tributary). *This breakdown highlights spatial disparities in model reliability, suggesting potential zone-specific influences.*

	type	avg_outlier_pct	high_outlier_sites	total_sites
1	m	50.3	2	14
2	t	55.1	3	19

Table 7: Outlier distribution in GAM residuals across individual rivers.

	wq_River_Names_og	avg_outlier_pct	high_outlier_sites	total_sites
1	Bassac River	50.9	1	5
2	Mae Kok River	33.3	0	1
3	Mekong River	49.3	2	13
4	MekongÂ River	62.5	0	1
5	Mun	52.1	0	1
6	Nam Ou River	70	0	1
7	Se Bangfai River	66.7	0	1
8	Se San River	75	2	2
9	Sekong RiverÂ	54.2	0	1
10	Song Khram River	64.6	0	1
11	Srepork	47.9	0	1
12	Tonle Sap Lake	45.7	0	3
13	Tonle Sap River	58.3	0	2

Appendix N. RDA Results

Table 8: Overall results of land use and water quality pooled RDA.

	Df	Variance	F	Pr(>F)
Model	3	1.261939	12.29839	0.001
Residual	197	6.738061	NA	NA

Table 9: Term results of land use and water quality pooled RDA.

	Df	Variance	F	Pr(>F)
urban	1	0.622367	18.19608	0.001
agriculture	1	0.464683	13.5859	0.001
forest	1	0.174888	5.113197	0.002
Residual	197	6.738061	NA	NA

Table 10: Overall results of macroinvertebrates and water quality RDA. *Macroinvertebrates were treated as the response variable to evaluate how shifts in water quality influence ecological communities.*

	Df	Variance	F	Pr(>F)
Model	8	128852.2	13.42004	0.002
Residual	181	217233.5	NA	NA

Table 11: Term results of macroinvertebrates and water quality RDA.

	Df	Variance	F	Pr(>F)
temp_c	1	14709.9	12.25636	0.001
ph	1	4402.248	3.667975	0.037
tss_mgl	1	102218.1	85.16865	0.003
cond_msm	1	624.2879	0.52016	0.531
totn_mgl	1	3150.249	2.624803	0.088
totp_mgl	1	1723.647	1.436151	0.21
do_mgl	1	1835.542	1.529383	0.192
codmn_mgl	1	188.1958	0.156806	0.841
Residual	181	217233.5	NA	NA

Appendix O. Single Predictor GAM Results

Table 12: Summary of pooled single predictor land use GAM.

Response	Adj_R2	Dev_Explained	Urban_p	Agriculture_p	Forest_p
temp_c	0.292	31.062	0.188	0.265	0.139
ph	0.245	28.394	0.001	0.045	0.034
tss_mgl	0.098	14.164	0.035	0.129	0.081
cond_msm	0.520	57.204	0.000	0.000	0.000
totn_mgl	0.280	32.033	0.053	0.000	0.015
totp_mgl	0.114	15.180	0.285	0.309	0.463
do_mgl	0.362	38.846	0.001	0.000	0.000
codmn_mgl	0.374	39.240	0.000	0.019	0.000

Table 13a: Summary of yearly single predictor land use GAM, 2011-2015

Year	Response	Adj_R2	Dev_Explained	Urban_p	Agriculture_p	Forest_p
2011	temp_c	0.274	34.388	0.896	0.505	0.131
2011	ph	0.167	24.783	0.382	0.684	0.159
2011	tss_mgl	0.085	17.377	0.481	0.989	0.357
2011	cond_msm	-0.021	10.520	0.625	0.629	0.810
2011	totn_mgl	0.292	37.379	0.030	1.000	0.222
2011	totp_mgl	0.315	40.897	0.130	0.162	0.235
2011	do_mgl	0.205	29.717	0.759	0.579	0.869
2011	codmn_mgl	0.462	51.968	0.008	0.775	0.036
2013	temp_c	0.255	37.032	0.304	0.294	0.481
2013	ph	0.028	12.514	0.307	0.831	0.575
2013	tss_mgl	0.257	33.992	0.216	0.065	0.012
2013	cond_msm	0.359	52.143	0.038	0.159	0.117
2013	totn_mgl	0.820	87.254	0.001	0.180	0.355
2013	totp_mgl	0.095	18.527	0.042	0.501	0.148
2013	do_mgl	0.498	56.993	0.025	0.109	0.028
2013	codmn_mgl	0.768	84.617	0.001	0.914	0.005
2015	temp_c	0.307	37.427	0.986	0.996	0.274
2015	ph	0.644	74.287	0.338	0.770	0.197
2015	tss_mgl	0.049	14.113	0.194	0.286	0.101
2015	cond_msm	-0.058	4.956	0.744	0.335	0.504
2015	totn_mgl	0.437	53.080	0.013	0.141	0.059
2015	totp_mgl	0.114	23.309	0.398	0.348	0.988
2015	do_mgl	0.638	73.360	0.076	0.063	0.195
2015	codmn_mgl	0.132	21.699	0.590	0.497	0.992

Table 13b: Summary of yearly single predictor land use GAM table continued, 2017-2021.

2017	temp_c	0.181	25.991	0.418	0.505	0.944
2017	ph	0.531	61.662	0.079	0.004	0.001
2017	tss_mgl	-0.056	6.139	0.627	0.752	0.832
2017	cond_msm	0.001	12.723	0.547	0.867	0.332
2017	totn_mgl	-0.018	10.948	0.423	0.908	0.975
2017	totp_mgl	0.138	24.448	0.955	0.209	0.267
2017	do_mgl	0.355	47.015	0.231	0.055	0.050
2017	codmn_mgl	0.154	23.573	0.057	0.214	0.034
2019	temp_c	0.349	41.386	0.894	0.939	0.311
2019	ph	0.897	94.344	0.006	0.000	0.000
2019	tss_mgl	-0.053	5.227	0.239	0.455	0.335
2019	cond_msm	-0.099	1.332	0.984	0.846	0.901
2019	totn_mgl	0.096	24.137	0.248	0.537	0.995
2019	totp_mgl	0.230	32.827	0.131	0.366	0.136
2019	do_mgl	0.540	61.885	0.067	0.006	0.001
2019	codmn_mgl	0.741	79.185	0.002	0.004	0.001
2021	temp_c	0.178	25.714	0.433	0.475	0.965
2021	ph	0.366	45.274	0.047	0.164	0.007
2021	tss_mgl	0.000	9.644	0.315	0.993	0.556
2021	cond_msm	-0.087	1.781	0.804	0.994	0.806
2021	totn_mgl	0.318	43.588	0.094	0.047	0.014
2021	totp_mgl	0.134	21.771	0.216	0.988	0.327
2021	do_mgl	0.720	76.819	0.000	0.000	0.000
2021	codmn_mgl	0.591	64.968	0.000	0.004	0.000

Table 14: Summary of pooled single predictor macroinvertebrate GAM.

Response	Adj_R2	Dev_Explained	ASPT_p	Richness_p	Abundance_p
temp_c	0.132	15.568	0.000	0.001	0.071
ph	0.045	7.022	0.027	0.124	0.345
tss_mgl	0.802	81.075	0.000	0.751	0.339
cond_msm	-0.005	1.646	0.675	0.719	0.846
totn_mgl	0.043	7.365	0.942	0.359	0.108
totp_mgl	0.225	26.076	0.003	0.023	0.004
do_mgl	0.028	5.405	0.088	0.427	0.437
codmn_mgl	0.030	5.204	0.553	0.353	0.161

Table 15: Comparison of pooled single predictor macroinvertebrate GAM results without richness. *"Full" columns show the original results with the inclusion of richness, "Reduced" columns show how much the results changed when richness was removed, and the final two columns show the GAM results without richness.*

Response	Full_Adj_R2	Full_Dev_ Explained	Reduced_Adj_R2	Reduced_Dev_ Explained	Adj_R2	Dev_Explained
temp_c	0.132	15.568	0.065	8.319	-0.067	-7.250
ph	0.045	7.022	0.032	5.011	-0.013	-2.010
tss_mgl	0.802	81.075	0.803	81.045	0.001	-0.030
cond_msm	-0.005	1.646	-0.009	0.169	-0.004	-1.480
totn_mgl	0.043	7.365	0.026	4.302	-0.017	-3.060
totp_mgl	0.225	26.076	0.204	23.555	-0.021	-2.520
do_mgl	0.028	5.405	0.023	4.150	-0.005	-1.250
codmn_mgl	0.030	5.204	0.029	4.489	-0.002	-0.720

Table 16a: Summary of yearly single predictor macroinvertebrate GAM, 2011-2013.

Year	Response	Adj_R2	Dev_Explained	ASPT_p	Richness_p	Abundance_p
2011	temp_c	0.867	92.588	0.000	0.000	0.215
2011	ph	0.491	60.031	0.005	0.018	0.266
2011	tss_mgl	0.002	9.882	0.304	0.717	0.877
2011	cond_msm	0.311	46.273	0.018	0.648	0.219
2011	totn_mgl	0.038	16.115	0.524	0.563	0.296
2011	totp_mgl	-0.022	11.082	0.862	0.406	0.576
2011	do_mgl	0.113	19.852	0.992	0.306	0.686
2011	codmn_mgl	-0.053	4.869	0.696	0.800	0.910
2013	temp_c	0.092	20.302	0.233	0.552	0.943
2013	ph	0.089	18.052	0.096	0.061	0.539
2013	tss_mgl	0.011	12.618	0.122	0.330	1.000
2013	cond_msm	-0.048	5.713	0.250	0.434	0.850
2013	totn_mgl	0.085	17.677	0.165	0.925	0.327
2013	totp_mgl	-0.050	5.461	0.763	0.381	0.333
2013	do_mgl	0.073	16.567	0.592	0.430	0.152
2013	codmn_mgl	0.287	38.936	0.763	0.840	0.064

Table 16b: Summary of yearly single predictor macroinvertebrate GAM continued, 2015-2021.

2015	temp_c	0.481	61.351	0.095	0.010	0.070
2015	ph	-0.043	6.892	0.655	0.346	0.523
2015	tss_mgl	0.137	23.597	0.672	0.473	0.051
2015	cond_msm	0.098	22.739	0.116	0.236	0.751
2015	totn_mgl	-0.007	9.088	0.419	0.860	0.118
2015	totp_mgl	0.553	68.152	0.043	0.264	0.023
2015	do_mgl	0.289	39.183	0.018	0.006	0.914
2015	codmn_mgl	0.014	12.700	0.630	0.837	0.342
2017	temp_c	0.254	36.881	0.206	0.275	0.207
2017	ph	0.265	39.580	0.264	0.865	0.054
2017	tss_mgl	0.012	13.152	0.482	0.716	0.702
2017	cond_msm	0.435	53.070	0.527	0.141	0.442
2017	totn_mgl	0.164	24.621	0.103	0.047	0.288
2017	totp_mgl	0.443	52.400	0.392	0.013	0.332
2017	do_mgl	0.364	47.931	0.015	0.212	0.078
2017	codmn_mgl	0.112	25.689	0.958	0.369	0.295
2019	temp_c	0.318	41.013	0.279	0.025	0.979
2019	ph	0.297	39.864	0.009	0.007	0.942
2019	tss_mgl	0.407	55.484	0.779	0.032	0.164
2019	cond_msm	0.051	15.457	0.212	0.055	0.232
2019	totn_mgl	0.327	43.717	0.996	0.802	0.040
2019	totp_mgl	0.037	13.304	0.764	0.627	0.077
2019	do_mgl	0.227	33.360	0.044	0.007	0.030
2019	codmn_mgl	0.107	21.751	0.189	0.033	0.136
2021	temp_c	0.009	10.468	0.361	0.465	0.468
2021	ph	0.011	10.644	0.203	0.305	0.854
2021	tss_mgl	0.894	91.488	0.937	0.000	0.159
2021	cond_msm	-0.075	2.888	0.815	0.710	0.959
2021	totn_mgl	0.007	12.032	0.654	0.851	0.322
2021	totp_mgl	0.340	40.794	0.656	0.839	0.330
2021	do_mgl	-0.081	2.348	0.718	0.804	0.535
2021	codmn_mgl	-0.013	8.529	0.897	0.667	0.147

Appendix P: Full GAM Results

A summary table of results for the full pooled model can be reviewed in Table 1 of the Research Paper.

Table 17a: Summary of full yearly model, 2011-2017.

Year	Response	Adj_R2	Dev_ Explained	ASPT_p	Richness _p	Abundance _p	Urban _p	Agriculture _p	Forest _p
2011	temp_c	0.925	96.632	0.000	0.000	0.209	0.137	0.624	0.040
2011	ph	0.762	84.205	0.001	0.000	0.650	0.029	0.003	0.000
2011	tss_mgl	0.146	31.497	0.258	0.658	0.581	0.299	0.432	0.125
2011	cond_msm	0.760	86.460	0.000	0.011	0.493	0.000	0.013	0.008
2011	totn_mgl	0.413	55.751	0.433	0.692	0.240	0.012	0.982	0.150
2011	totp_mgl	0.378	53.409	0.278	0.080	0.252	0.128	0.039	0.078
2011	do_mgl	0.388	51.138	0.888	0.151	0.588	0.137	0.977	0.176
2011	codmn_mgl	0.536	63.449	0.702	0.431	0.836	0.002	0.377	0.026
2013	temp_c	0.244	40.741	0.128	0.160	0.524	0.313	0.110	0.402
2013	ph	0.080	28.404	0.268	0.260	0.834	0.379	0.833	0.397
2013	tss_mgl	0.299	46.264	0.383	0.261	0.643	0.285	0.105	0.032
2013	cond_msm	-0.052	18.743	0.156	0.385	0.889	0.498	0.227	0.277
2013	totn_mgl	0.815	88.360	0.399	0.791	0.145	0.006	0.592	0.550
2013	totp_mgl	-0.001	19.885	0.995	0.935	0.567	0.065	0.515	0.203
2013	do_mgl	0.505	64.037	0.641	0.959	0.616	0.052	0.123	0.064
2013	codmn_mgl	0.730	84.309	0.459	0.572	0.427	0.008	0.892	0.791
2015	temp_c	0.524	62.451	0.021	0.001	0.254	0.681	0.830	0.282
2015	ph	0.513	64.774	0.303	0.205	0.798	0.008	0.011	0.005
2015	tss_mgl	0.296	48.245	0.538	0.310	0.424	0.183	0.349	0.054
2015	cond_msm	0.073	30.095	0.170	0.153	0.714	0.703	0.496	0.954
2015	totn_mgl	0.558	69.169	0.325	0.496	0.016	0.104	0.313	0.058
2015	totp_mgl	0.814	90.067	0.001	0.003	0.000	0.648	0.003	0.098
2015	do_mgl	0.592	71.899	0.089	0.063	0.646	0.901	0.164	0.061
2015	codmn_mgl	0.317	49.906	0.712	0.667	0.168	0.419	0.890	0.525
2017	temp_c	0.400	54.258	0.136	0.582	0.551	0.126	0.442	0.774
2017	ph	0.624	75.514	0.651	0.489	0.108	0.400	0.008	0.031
2017	tss_mgl	0.456	68.214	0.485	0.025	0.157	0.861	0.595	0.789
2017	cond_msm	0.569	70.252	0.756	0.042	0.086	0.732	0.545	0.331
2017	totn_mgl	0.138	32.395	0.220	0.087	0.354	0.873	0.883	0.895
2017	totp_mgl	0.445	58.358	0.410	0.136	0.847	0.799	0.146	0.618
2017	do_mgl	0.480	63.970	0.069	0.386	0.130	0.866	0.156	0.302
2017	codmn_mgl	0.186	36.645	0.993	0.692	0.165	0.078	0.089	0.018

Table 17b: Summary of full yearly model continued, 2017-2021.

2019	temp_c	0.522	65.173	0.794	0.292	0.215	0.493	0.819	0.396
2019	ph	0.621	75.420	0.204	0.069	0.870	0.471	0.972	0.006
2019	tss_mgl	0.386	60.395	0.595	0.064	0.353	0.158	0.295	0.183
2019	cond_msm	0.101	29.214	0.061	0.008	0.058	0.517	0.543	0.933
2019	totn_mgl	0.374	56.123	0.473	0.581	0.284	0.936	0.807	0.801
2019	totp_mgl	0.253	43.033	0.415	0.688	0.387	0.662	0.538	0.200
2019	do_mgl	0.685	78.491	0.559	0.091	0.054	0.013	0.006	0.000
2019	codmn_mgl	0.833	89.705	0.181	0.122	0.216	0.009	0.001	0.001
2021	temp_c	0.279	42.626	0.466	0.328	0.084	0.266	0.479	0.994
2021	ph	0.386	52.947	0.226	0.259	0.251	0.088	0.210	0.010
2021	tss_mgl	0.938	96.533	0.933	0.000	0.147	0.433	0.288	0.946
2021	cond_msm	-0.167	5.865	0.915	0.754	0.778	0.712	0.949	0.787
2021	totn_mgl	0.332	52.318	0.664	0.704	0.253	0.342	0.050	0.013
2021	totp_mgl	0.766	85.892	0.302	0.002	0.004	0.706	0.058	0.278
2021	do_mgl	0.717	80.100	0.590	0.471	0.995	0.000	0.000	0.000
2021	codmn_mgl	0.583	68.399	0.301	0.265	0.226	0.000	0.004	0.000

Table 18: Prediction accuracy metrics for the full GAM modes across water quality variables. *RMSE, mean residual, and standard deviation of residuals were calculated for each response variable based on GAM predictions at four withheld ecological health sites. The complete set of site-level prediction results is available in CSV format but omitted here due to length.*

Response	RMSE	Mean_Residual	SD_Residual
codmn_mgl	2.718	1.498	2.319
cond_msm	50.013	24.727	44.450
do_mgl	2.006	-1.007	1.774
ph	0.351	-0.081	0.349
temp_c	3.270	-1.571	2.933
totn_mgl	1.228	0.655	1.062
totp_mgl	0.048	0.005	0.048
tss_mgl	25.847	5.138	25.901

Appendix Q: Variance Partitioning on Full GAM Results

Table 19: Variance fractions explained by land use and macroinvertebrate predictors. *Shared and unique fractions are reported as adjusted R^2 values. X1|X2 refers to variation explained uniquely by predictor set X1 (land use), after accounting for X2 (macroinvertebrates).*

	Df	R.squared	Adj.R.squared	Testable
[a] = X1|X2	3	NA	0.133	TRUE
[b] = X2|X1	3	NA	0.076	TRUE
[c] = Shared	0	NA	0.000	FALSE
[d] = Residuals	NA	NA	0.791	FALSE

Table 20: Permutation test results for the unique effect of land use predictors. *Significance testing of the variance fraction uniquely explained by land use, accounting for macroinvertebrate metrics.*

Df	Variance	F	Pr(>F)
3	1.149	11.450	0.001
183	6.124	NA	NA

Table 21: Permutation test results for the unique effect of macroinvertebrate predictors.

Df	Variance	F	Pr(>F)
3	0.702	6.990	0.001
183	6.124	NA	NA

Table 22: Year-wise variance partitioning of water quality responses explained by land use and macroinvertebrate predictors.

Year	LU_Unique	EH_Unique	Shared	Residual	Total_Explained
2011	0.264	0.121	-0.072	0.687	0.313
2013	0.136	-0.021	0.061	0.824	0.176
2015	0.145	0.062	-0.009	0.801	0.199
2017	0.071	0.133	-0.016	0.811	0.189
2019	0.091	0.041	0.048	0.820	0.180
2021	0.178	0.097	0.009	0.715	0.285

Appendix R: Mainstem and Tributary GAM Results

Table 23: Full pooled GAM results with only mainstem stations.

Response	Adj_R2	Dev_ Explained	ASPT_p	Richness _p	Abundance _p	Urban_p	Agriculture _p	Forest_p
temp_c	0.636	69.186	0.173	0.052	0.486	0.492	0.001	0.358
ph	0.305	37.132	0.001	0.040	0.167	0.671	0.000	0.003
tss_mgl	0.287	34.272	0.099	0.337	0.215	0.888	0.000	0.004
cond_msm	0.554	61.088	0.165	0.816	0.841	0.453	0.000	0.290
totn_mgl	0.222	28.566	0.261	0.755	0.015	0.633	0.000	0.181
totp_mgl	0.030	10.551	0.298	0.947	0.521	0.547	0.084	0.268
do_mgl	0.585	64.447	0.486	0.415	0.147	0.030	0.000	0.000
codmn_mgl	0.530	59.602	0.860	0.732	0.340	0.667	0.000	0.048

Table 24: Full pooled GAM results with only tributary stations.

Response	Adj_R2	Dev_ Explained	ASPT_p	Richness _p	Abundance _p	Urban_p	Agriculture _p	Forest_p
temp_c	0.562	61.327	0.000	0.000	0.462	0.000	0.000	0.000
ph	0.392	46.172	0.608	0.556	0.637	0.231	0.010	0.016
tss_mgl	0.889	90.520	0.000	0.906	0.266	0.000	0.002	0.015
cond_msm	0.696	75.056	0.205	0.252	0.442	0.000	0.026	0.002
totn_mgl	0.452	52.442	0.452	0.434	0.042	0.094	0.123	0.071
totp_mgl	0.473	54.093	0.002	0.034	0.014	0.009	0.010	0.032
do_mgl	0.456	51.896	0.188	0.641	0.121	0.183	0.022	0.049
codmn_mgl	0.453	50.517	0.791	0.349	0.060	0.230	0.614	0.269