Robustness of pesticide and other environmental stressors as key drivers of stream macroinvertebrates in small agricultural catchments

Hanh H. Nguyen[*,1,2], Verena C. Schreiner[1,2], Ralf B. Schäfer[1,2]

[1]Faculty of Biology, University of Duisburg-Essen, Essen, Germany

[2]Research Center One Health Ruhr, University Alliance Ruhr, Essen, Germany

*Corresponding author: honghanh.nguyen@uni-due.de.

ABSTRACT

Robustness of multiple stressor rankings is essential for credible ecotoxicological assessments and policy guidance. A widely cited study of 101 small agricultural streams across Germany identified pesticide mixtures as the dominant stressor for stream macroinvertebrates, but its analytical robustness has since been questioned. Using a fully reproducible workflow, we reanalysed this dataset to evaluate how data aggregation and modelling choices shape conclusions about stressor importance. Pesticide toxicity remained consistently identified as a key stressor regardless of modelling approach. Summed toxic units combining event and grab sampling strengthened the pesticide-macroinvertebrate association compared to maximum toxic units or single sampling methods, suggesting that mixture-level exposure mechanisms and substance-specific toxicokinetics matter. Beyond pesticides, data aggregation choices influenced the relative importance of multiple stressors, with agricultural land use, nutrients, and hydromorphological degradation showing stronger effects than previously reported and no single stressor dominating ecological responses. Different ecological metrics responded to distinct stressor sets, highlighting metric choice as a relevant consideration in interpreting multiple stressor effects. Our findings reveal that conclusions about dominant stressors are sensitive to analytical decisions, calling for transparent, multi-metric, multi-model approaches to enable more defensible evaluations of chemical mixture and multiple stressor effects on freshwater biodiversity under increasing global pollution.

KEYWORDS: *Chemical mixture, toxicants, land uses, biomonitoring, stability selection*

INTRODUCTION

Agriculture is a main driver of biodiversity loss, with species exposed to agricultural landscapes exhibiting the strongest declines[1,2]. Decline trends have been documented across multiple organism groups, including terrestrial insects[3,4] and stream macroinvertebrates[5,6]. These declines have been attributed to multiple stressors, including habitat degradation and excessive agricultural inputs of nutrients and pesticides into adjacent non-target ecosystems[7–9]. Yet, comparisons of the relative importance of pesticide exposure versus other environmental stressors in real-world ecosystems remain scarce, limiting the evidence base for chemical risk assessment and water quality regulation. Small agricultural streams are particularly vulnerable to pesticide pollution due to lower dilution potential than large rivers, often resulting in higher pesticide concentrations[10], and short, rainfall-driven exposure pulses that can exert ecotoxicological effects within hours to days[11,12]. In this context, Liess et al.[13] conducted the first field study on a national scale to employ event sampling to capture potential pesticide peaks, reporting a strong association between pesticide toxicity estimates based on the most toxic substance and multiple ecological metrics, and identifying pesticide toxicity as the dominant stressor driving declines in stream macroinvertebrates. Recently, Moore and Rathjens[14] questioned the reproducibility of Liess et al.[13] and argued that conclusions are sensitive to data aggregation choices, including toxicity metric selection and the temporal alignment of chemical and biological sampling. Pesticide toxicity estimates are typically characterized using metrics representing either maximum (TUmax) or cumulative (TUsum) mixture toxicity[15]. Liess et al.[13] used TUmax as the primary metric, considering only the most toxic substance rather than the full mixture, and omitted samples with high toxicity peaks. Schriever et al.[16] similarly questioned the robustness of these conclusions, suggesting that other stressors such as land use may better explain biodiversity decline.

Conclusions about stressor importance can also be sensitive to modelling choices, particularly variable selection[17]. High-dimensional models with interaction terms may capture complex ecological relationships but produce unstable parameter estimates and are prone to overfitting under the constrained sample sizes typical of ecotoxicological field studies [18,19]. Stability analysis and bootstrapping have been advocated to quantify robustness by identifying variables consistently selected across perturbed datasets[20,21]. Despite these methodological advances, how sensitive conclusions about dominant stressors are to data aggregation and modelling choices has not been systematically evaluated.

Here, we use the Liess et al.[13,22] dataset to address this gap through a fully reproducible reanalysis. We evaluated alternative toxicity metrics, sampling combinations, and statistical modelling approaches with bootstrap stability analysis to assess how data aggregation and modelling choices shape conclusions about pesticide and multiple stressor effects on stream macroinvertebrates. By doing so, we move beyond reproduction to provide a new interpretation of existing data with implications for chemical risk assessment, water quality regulation, and the design of monitoring programs intended to identify key stressors driving biodiversity decline.

MATERIALS AND METHODS

Background of the abiotic and biotic datasets and metrics

The Kleingewässermonitoring (KgM) dataset[22] includes 101 small streams with catchment areas < 100 km^2 across Germany, spanning a broad gradient of agricultural intensity (86 streams with 20–100% agricultural land cover). Chemical sampling occurred during April–July 2018 and 2019 to capture the main pesticide application periods. Two types of pesticide samples were collected at each stream: (i) grab samples every third week under baseline weather conditions, defined as the absence of strong rain events with > 10 mm/day, and (ii) event samples following strong rain events[13]. In total, 840 samples were analysed for 76 pesticides (40 herbicides, 24 fungicides, and 12 insecticides) and 32 pesticide metabolites. Additionally, nutrients and trace elements (arsenic, cadmium, copper, zinc, lead, and mercury) were analysed in each sample, whereas physico-chemical variables (e.g., pH, dissolved oxygen, flow velocity) were measured *in situ* in parallel to the grab samples. Habitat condition was characterised through hydromorphological variables (e.g., stream depth and width) and substrate types measured *in situ*, as well as land use types of the upstream catchments derived from CORINE land cover 2018 data. Benthic macroinvertebrates were sampled in June 2018–2019, coinciding with peak pesticide applications. For complete methodology, see Liess et al.[13,22] and Halbach et al.[23]. We focussed on the three ecological metrics that explained the highest variance in Liess et al. (2021a): $SPEAR_{pesticides}$ (Species at risk from pesticides[24]), which represents the fraction of pesticide-vulnerable stream macroinvertebrates and specifically indicates pesticide effects; the saprobic index, which indicates oxygen deficiency from organic pollution[25]; and %EPT, the percentage of Ephemeroptera, Plecoptera and Trichoptera taxa relative to the total number of taxa in the macroinvertebrate community[26]. To allow site-level analysis, abiotic and biotic metrics per site over two years were aggregated using mean aggregation for the 11 of 101 sites where data were available for both years, consistent with Liess et al.[13].

Reproducibility of pesticide toxicity estimates

We reproduced the pesticide toxicity estimates for all 101 study sites by Liess et al.[13] (see Fig. A1) and compared our reproduction with both the original study and the deviating estimations by Moore and Rathjens[14]. We briefly outline the basics of pesticide toxicity estimation and then address reproducibility issues. Pesticide concentrations were converted into estimates of their toxicity towards stream macroinvertebrates by calculating toxic units (TUs) for each substance. TUs were calculated by dividing measured pesticide concentrations by their respective LC_{50} values of the more sensitive taxon from either *Daphnia magna* or *Chironomus sp.*, provided in the KgM dataset[22]. For each water sample, the maximum TU (TUmax) was determined as the highest TU across all detected substances, and per site, the highest TUmax across all samples within a year was selected and log10-transformed for further analyses.

We identified three issues that explained the mismatch between studies and hampered reproducibility. Firstly, Liess et al.[13] considered all pesticides with values equal to or above the limit of detection (LOD), whereas Moore and Rathjens [14] only considered those with concentrations above the limit of quantification (LOQ). Although values >LOD and <LOQ exhibit a lower accuracy, omitting them from the data set is equivalent to replacing them with a constant (i.e., with 0), which performed poorly in simulation studies[27]. We followed Liess et al.[13] in our analysis, as lower accuracy incurs less bias than setting values to zero as done by Moore and Rathjens [14]. Secondly, the definition of exceptionally high TUmax values in Liess et al. [13] was imprecise. They stated a sample was exceptionally toxic when it exceeded the mean of "five subsequent samples" by a factor of 100, but did not restrict this criterion to sites or years having more than five subsequent samples. We scrutinised their decision to exclude exceptionally toxic samples and to use the second-highest TUmax for a site-year combination when exceeding the mean of up to five subsequent samples by a factor of 100. Lastly, Liess et al. [13] did not document that TUmax values of samples below -5 were replaced by this value, which hampered reproducibility at sites with very low toxic units. This pragmatic cutoff has been commonly used because no acute mortality in field or laboratory has been observed below 1/10,000 of the LC_{50} of standard test organisms[15,24]. We followed their approach and truncated the pesticide toxicity estimates at −5. The reproduction issues were communicated with the main authors of Liess et al.[13] and Moore and Rathjens[14] to ensure transparency and identify potential discrepancies.

Selection of metrics for pesticide toxicity estimation

To evaluate the influence of pesticide toxicity metrics on the relationship between pesticide toxicity and stream macroinvertebrates, an issue raised by Moore and Rathjens[14], we compared TUsum, which is the sum of all pesticides and metabolites per sample following the concept of concentration addition[15], versus the following metrics (Fig. 1): first, TUmax, representing a sample aggregation approach based on the maximum toxicity across all samples and sample types per site; second, TUsum separately for event and grab samples, to scrutinise the influence of chemical sample types on the results; and third, TUsum without exceptionally high peaks, following the exclusion approach described by Liess et al.[13] for TUmax, applied here to TUsum. The TUsum for pesticide mixtures was computed by summing the TUs of all detected substances within each sample, with the highest TUsum across all samples per site selected and log10-transformed as the representative toxicity estimate.

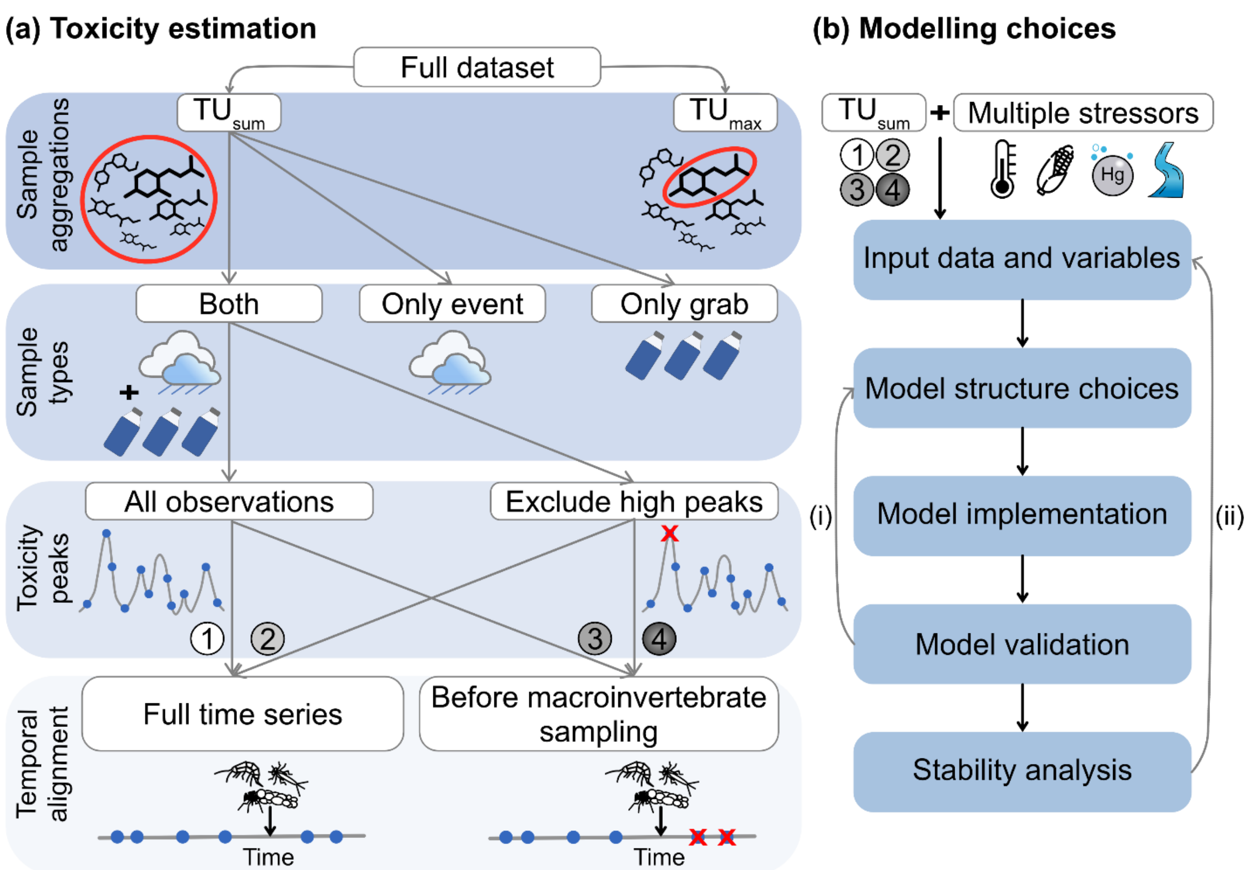

Figure 1. Study workflow to address effects of pesticide toxicity estimates (a) and statistical modelling choices (b) on stream macroinvertebrates. TUsum = sum toxic unit. TUmax = max toxic unit. (i) Model validation verifies choices of model structures. (ii) Stability analysis using different statistical models strengthens conclusions about influential data and explanatory variables.

In addition, to address the temporal alignment concern raised by Moore and Rathjens[14], we compared the TUsum across all samples taken over the sampling period (metric 1; Fig. 1a) with the TUsum restricted to samples collected before the June macroinvertebrate sampling date, thereby excluding pesticide measurements from late June and July (metric 3; Fig. 1a). Finally, we calculated the TUsum considering both the exclusion of high toxicity peaks and temporal alignment of chemical and biological samples (metric 4; Fig. 1a) and compared it to the TUsum excluding only exceptionally high peaks (metric 2; Fig. 1a). In total, we derived seven pesticide toxicity metrics from different combinations of sample aggregation (TUsum versus TUmax), sample types (only grab, only event, or both), and data filtering choices (inclusion or exclusion of high peaks and temporal alignment; Fig. 1a). We used linear regression models to compare the explained variance (R^2) across toxicity metrics for their relationship with $SPEAR_{pesticides}$., as this metric was originally designed to indicate pesticide-related community changes[24]. Together, these comparisons inform the selection of pesticide toxicity metrics (e.g., metrics 1-4) for assessing variable importance within a multiple-stressor framework addressed in 2.4.2.

Statistical modelling of the relative importance of pesticide toxicity estimates

We first reproduced the statistical modelling of Liess et al.[13] and then scrutinised how choices of pesticide toxicity metrics and statistical modelling influenced the relative importance of pesticide toxicity estimates for stream macroinvertebrate metrices (Fig.1b).

Reproducibility of the statistical modelling

We followed the modelling workflow and methods of stressor aggregation of Liess et al.[13] to reproduce their results on relative variable importance. Briefly, prior to model selection, explanatory variables considered in Liess et al.[13] were screened for multicollinearity using variance inflation factors (VIF), and variables with VIF > 2 were excluded. Linear regression models were fitted separately for three ecological metrics (see 2.1). Forward stepwise model selection based on the Akaike Information Criterion (AIC) was applied, starting from an intercept-only model, where the full model contained all main effects and two-way interactions. The final model was obtained through manual term selection, retaining only significant main effects ($p < 0.05$) and including interaction terms only when both corresponding main effects were significant. We quantified the total explained variance of the best-fit model using the adjusted R^2, while the relative importance of explanatory variables was determined using hierarchical partitioning, i.e., the *lmg* metric from *relaimpo* package in R[28]. However, we initially obtained different model estimates and explained variances compared to Liess et al.[13], which were resolved following consultation with the main authors and a review of the shared data and computer code, though some effect directions remain irreproducible (Fig. A2). The main barrier was an unstated list of variables included in the full model, with the original analysis using a subset of 13 of the 18 explanatory variables described in the manuscript (i.e., pesticide toxicity, trace element toxicity, and ammonium, nitrite, total phosphorous, flow velocity, stream width, stream depth, hydromorphology, bed habitat structure, pH, dissolved oxygen, and temperature); this choice was not documented in the original study. The difference in effect directions from original study outputs was due to the relabelling of some stressors as deficits to emphasise their stressor potential.

Statistical modelling considering different data aggregation and model types

We evaluated the sensitivity of pesticide toxicity rankings to different data aggregation and modelling choices. In particular, we examined the influence of using different pesticide toxicity metrics, focusing on four cases addressing toxicity peaks and temporal alignment (see 2.3, Fig. 1b). Following model implementation and validation of model structure, we included the stability analysis, which compares variable selection from the original AIC-based model against alternative statistical models to evaluate whether pesticides are consistently identified as an influential driver among multiple stressors.

In addition, we applied modifications to some abiotic data aggregation and original model structures, addressing issues raised by Moore and Rathjens[14] and identified in 2.4.1, respectively. For abiotic variables accompanying the toxicity metrics (see 2.3), we excluded dissolved oxygen and temperature due to extensive missing data (> 50% of sites), where Liess et al. [13] had used imputed values. Flow values were aggregated per site using the arithmetic mean rather than geometric mean[13], as the data distribution showed minimal skewness and included ecologically meaningful zeros (0.0–1.3 m/s). Total nitrogen (TN) and total phosphorus (TP) served as proxies for nutrient effects, consistent with the use of summed toxic units for pesticides and trace elements. Pairwise stressor–response correlations were examined for linearity and nutrient variables were log10-transformed, consistent with transformations applied to pesticide and trace elements. All predictors were standardised to a zero mean and unit variance to ensure comparability of effect sizes. For variable selection, we applied backward AIC-based model selection using the stepAIC function in the MASS package[29], prioritised over forward selection for more accurate parameter estimates and p-values[30]. In contrast to Liess et al.[13], we retained non-significant explanatory variables in the best-fit based models. Regarding model structure, Liess et al.[13] included 13 variables and all 78 two-way interactions, a specification that poses overfitting risk with only 101 sites[17], yet retained none interaction in their final models. Thus, we fitted models without interaction terms.

We also tested for spatial autocorrelation in the best-fit models using permutation-based Moran's I (999 simulations, function moran.mc in the spdep R package[31]). Given that no significant spatial dependency was detected in the responses $SPEAR_{pesticides}$, saprobic index and %EPT, nor in residuals of their corresponding best-fit models (p-value > 0.05), we omitted spatial variables from statistical modelling.

To examine model robustness, we compared bootstrap variable selection from the AIC-based linear regression model with penalized regression approaches commonly used in statistical learning: elastic net[32], Lasso[33], minimax concave penalty (MCP)[34] and a triangulated combined model integrating these methods[20]. Each approach was evaluated using 1,000 nonparametric bootstrap resamples with fixed random seeds for reproducibility. For each resample, model coefficients and selected variables were recorded to quantify stability as the proportion of iterations retaining each variable. Variables retained in ≥70% of replicates were considered stable under the AIC-based approach, whereas adaptive thresholds for the penalized regressions were derived using the stabilize and triangulate functions of the stabilizer package[20]. Consistency of the effect direction was quantified using the sign-instability (bootstrap-p) metric, defined as the proportion of signs opposing that of the mean[20]. All analyses in this study were conducted in R version 4.4.2[35]. The relationships of abiotic and biotic variables were visualized using the ggplot2 package[36]. All data and R code to reproduce the analyses presented in this study are publicly available following FAIR principles (see Data and Code Availability section).

RESULTS AND DISCUSSION

Relationship between pesticide toxicity and stream macroinvertebrates

Whole-mixture toxicity estimates better capture the association with ecological responses

The choice of the toxicity metric influenced the ecological interpretation (Fig. 2). The sum toxic unit (TUsum) and the max toxic unit (TUmax), which aggregate toxicity across all pesticides and metabolites or consider only the most toxic pesticide substance, respectively, showed similar value distributions (Fig. 2a). Nevertheless, TUsum explained slightly more variance in $SPEAR_{pesticides}$

than TUmax (R^2 = 0.36 vs. 0.33; Fig. 2d). These results are in line with a previous study that compared different toxicity metrics on multiple data sets[15] and support the concept of concentration addition, which underlies the TUsum calculation. At the same time, focusing on single substances alone that exhibited the highest toxicity estimate reduced the explanatory power only slightly, consistent with earlier studies demonstrating that toxicity estimates are often governed by very few substances[37,38]. Yet, given that the most toxic substance varies across sites, chemical monitoring needs to consider a wide spectrum of substances to yield reliable toxicity estimates[37], even if the choice of mixture aggregation may have a minor influence.

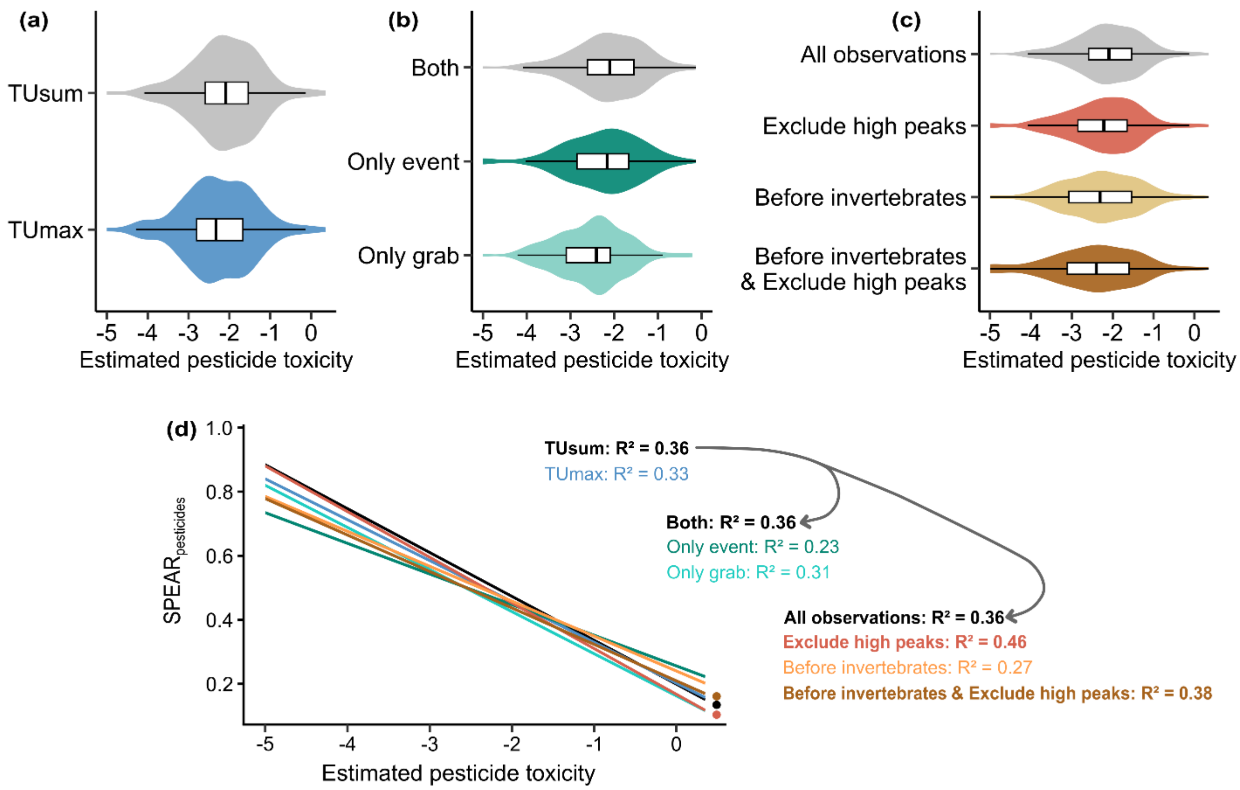

Figure 2. Pesticide toxicity estimates with different data aggregation choices: a) Sum toxic unit (TUsum) versus max toxic unit (TUmax); b) TUsum with grab (only grab), event (only event), and combined sampling (both); c) Sum toxic unit with all observations versus excluding high peaks, restricted to before macroinvertebrate sampling (before invertebrates), and both (before invertebrates and exclude high peaks); d) Association between pesticide toxicity estimates from a), b), and c) and stream macroinvertebrate responses, measured by the SPEAR$_{pesticides}$, expressed as explained variance (R^2) from linear regression. The dots indicate three regression models with the highest explained variance values. Given differences in the toxicity estimates, scatterplots with individual data points for each metric are provided in Supporting Information (Fig. A3).

Combined event and grab sampling methods capture exposure more completely

The highest toxicity values and strongest associations with $SPEAR_{pesticides}$ emerged when considering both event and grab samples in the calculation of the pesticide toxicity metric (Fig. 2b, 2d). This suggests that the ecological risk is governed jointly by pesticide pulses associated with rain events and baseline weather conditions. Restricting sampling methods to a single sampling type would therefore underestimate environmentally relevant exposure, as demonstrated across seven European countries where reliance on a single sampling type did not adequately capture concentration peaks relevant to regulatory and ecological assessments[39]. Notably, grab samples were higher than the corresponding event samples at roughly one-third (29 of 101) of sites, likely reflecting higher concentrations under low-flow conditions due to reduced dilution, combined with continuous pesticide inputs from point sources or groundwater exchange[40].

Differences in TUsum-based metrics between event and grab samples also highlight the importance of toxicokinetics when linking toxicity to ecological responses. Although TUsum was often higher in event samples, these samples exhibited a lower association with $SPEAR_{pesticides}$ than grab samples (Fig. 2d). Similarly, high-frequency monitoring studies found limited explanatory power of extreme rain-event concentrations for ecological responses[41]. This suggests that short-term peaks alone can only partially explain macroinvertebrate community changes. Rather, continuous exposure appears to be an important contributor to ecological effects. For instance, sediment-associated toxicity, which reflects continuous exposure, was a stronger predictor of macroinvertebrate effects than water samples reflecting short-term exposure[42]. Such baseline exposure, even at lower concentrations, can cause chronic effects in macroinvertebrate communities against a background of frequently recurring pulses[43]. Organisms recovering slowly from a single exposure pulse may remain vulnerable to subsequent pulses, meaning that the interplay between event-driven peaks and baseline exposure, rather than either regime alone, shapes community-level responses[43].

High toxicity peaks do not proportionally reflect ecological responses

Our comparison of event and grab samples suggested that grab samples were ecologically more important (see previous section), which was also supported by our analysis of the contribution of high toxicity peaks. We evaluated the metric introduced by Liess et al.[13], which calculates TUsum after replacing exceptionally high toxicity peaks with the next-lower value in the combined dataset, thus retaining event samples but reducing the influence of high toxicity peaks. Excluding these peaks improved the association with $SPEAR_{pesticides}$ ($R^2 = 0.46$ versus $R^2 = 0.36$ when including all observations; Fig. 2c, 2d), even though these peaks elevated toxicity estimates at approximately 20% of sites. This reinforces the relevance of toxicokinetics: uptake and elimination times vary considerably across substances[44], and when exposure duration is too short relative to organism uptake rates, even high external concentrations may not translate proportionally into internal doses and subsequent ecological effects[43]. Consequently, the combination of baseline exposure and recurring pulses, rather than high individual toxicity peaks, may drive macroinvertebrate community changes[43,45]. Our findings suggest that excluding high peaks from toxicity estimates may better represent the ecologically effective exposure, though further research is needed to determine whether the excluded high peaks are consistently driven by specific substances with distinct toxicokinetic properties.

Temporal alignment of chemical and biological monitoring improves causal inference at the cost of statistical power

To address the concern that toxicity estimates from water samples collected after biological sampling could not be causally linked to the observed macroinvertebrate responses[14], we calculated TUsum using only chemical samples collected before the biological sampling date. Temporal alignment reduced toxicity values and lowered the association with $SPEAR_{pesticides}$ ($R^2 = 0.27$; Fig. 2d) compared to the full dataset ($R^2 = 0.36$). This raises the question of why removing pesticide data from the peak application periods in late June and July, occurring after biological sampling, diminishes the explanatory power of ecological response models. Pesticide exposure after sampling cannot, by definition, explain the ecological response in a causal sense. However, these late-season pesticide sampling data may still carry valuable information: it might reflect the typical exposure intensity at a site as shaped by recurring application patterns across years or offer a more representative estimate of site-level toxicity than the limited measurements taken during the sampling months. When temporal alignment was combined with exclusion of exceptionally high toxicity peaks, the explained variance slightly increased to $R^2 = 0.38$, comparable to the full dataset metric (Fig. 2c, 2d), suggesting that high peaks may contribute disproportionately to noise in the temporally aligned dataset, where their relative influence is amplified by the reduced sample size, and that their removal allows it to capture underlying ecological responses to pesticides. Taken together, the negative association between pesticide toxicity and stream macroinvertebrate metrics is relatively robust across different toxicity estimates, though no single metric stands out. For regulatory monitoring, we recommend combining event and grab sampling with temporally aligned, sum toxic units, while also considering the influence of high peaks.

Multiple linear regression verifies influential role of pesticide mixture among multiple stressors

Our multiple stressor analysis identified pesticide toxicity as an influential stressor for stream macroinvertebrates, supporting the results from Liess et al.[13] and being consistent with findings from multi-regional assessments across the United States[46]. Eight of eleven stressors in the analysis were selected as explanatory variables in best-fit statistical models of the three ecological metrics (Fig. 3). None of the individual stressors contributed more than 50% of the explained variance across all fitted models, emphasising the relevance of considering multiple stressors to explain community patterns. The absence of a single prevailing stressor is consistent with field studies showing that pesticides are important but interact with other stressors, such as habitat degradation and land use, in shaping macroinvertebrate responses[6,41,47].

Pesticide mixtures were consistently identified as a key stressor among multiple stressors by $SPEAR_{pesticides}$ (R^2 = 0.12–0.25; Fig. 3). Pesticide toxicity represented the stressor with the highest explanatory power, termed “dominant stressor” in Liess et al.[13], in three of the four toxicity metric cases considered (Fig. 3). The exception was the temporally aligned toxicity metric, which had the lowest association with $SPEAR_{pesticides}$ in the single-stressor analysis (see 3.1.4) and similarly exhibited the lowest explained variance in the multiple stressor model (R^2 = 0.12; Fig. 3c). In this case, total phosphorus (TP) emerged as the stressor with the highest explained variance. The importance of TP is widely recognised given it is often a limiting nutrient structuring stream communities[5]. Another key difference from Liess et al.[13] is our use of aggregated nutrient metrics (TN, TP) rather than individual nutrient species (e.g., nitrate, nitrite, or ammonium); This aggregation choice contributed to increased explanatory power of TP across all models with different pesticide toxicity metrics and ecological metrics. In addition, it revealed additional influential co-occurring stressors such as agricultural land use that were not identified in the original study (Fig. 3; SI—Fig. 2A). Overall, this illustrates stressor interdependency in a multiple-stressor context, where a shift in one variable can affect the relative importance of other variables. Our results also highlight that the diagnosis of relative stressor importance depends on the choice of the ecological metric[48]. Although the three ecological metrics used in this study were highly correlated (Pearson correlation $|r| \geq 0.6$), they responded to different sets of stressors. For instance,

$SPEAR_{pesticides}$, which relies on ecological and biological traits that are hypothesised to respond to pesticide stress (including physiological sensitivity and recolonisation ability; Liess and Ohe, 2005), indicated a higher association with TP than the taxonomy-based %EPT metric across our models. This is consistent with trait-based metrics showing stronger association with individual stressors than taxonomic metrics[50,51]. Indeed, while %EPT correlated highly with $SPEAR_{pesticides}$ (r = 0.79) and the overall explained variance among the best-fit models of these two metrics was similar (Fig. 3a, c, d), %EPT responded to a broader range of stressors, including flow velocity, pH, and trace elements. This suggests that %EPT, while less responsive to individual stressors, may capture a wider spectrum of stressors. The saprobic index, a well-known indicator of nutrient and organic enrichment, correlated with $SPEAR_{pesticides}$ and %EPT (r of -0.61 and -0.59, respectively) and responded particularly to pesticide toxicity, agricultural land use, TP, and flow velocity, though it had the lowest explained variance among the three metrics (Fig. 3). Collectively, these findings underscore that the choice of ecological metric directly shapes which environmental drivers are identified as influential, reinforcing the need for multi-metric approaches in multiple stressor assessments[52,53].

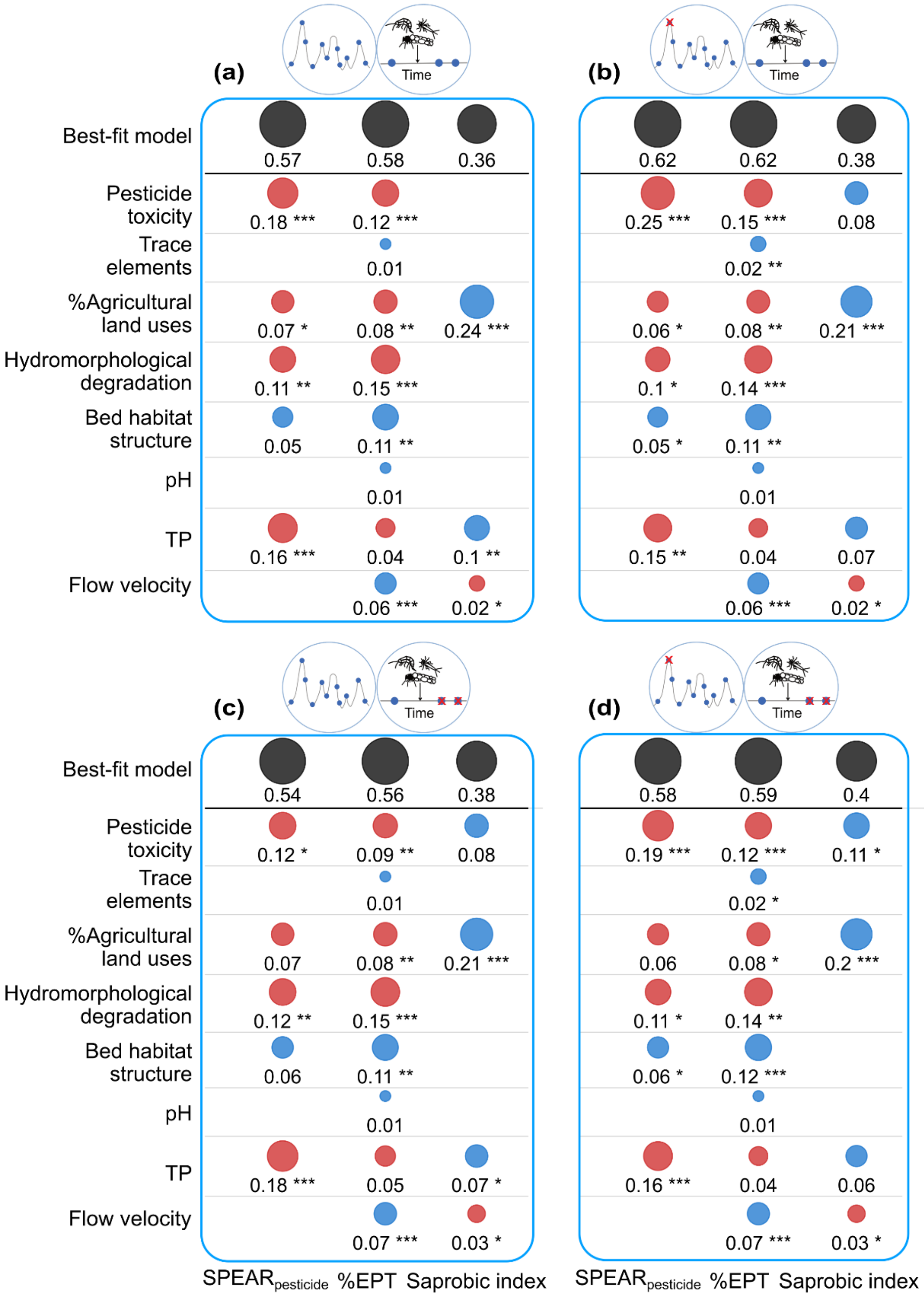

Figure 3. Relative importance of multiple stressors for three ecological metrics (SPEAR$_{pesticides}$, % EPT, saprobic index) using multiple linear regressions. Specific cases of sum toxic units: (a) All observations and full time series. (b) Excluding high peaks. (c) Before macroinvertebrate sampling. (d) Before macroinvertebrate sampling and excluding high peaks. Dots represent explained variances (numbers below dots), with red and blue indicating negative and positive estimates,

respectively. Significance: *p < 0.05; **p < 0.01; ***p < 0.001. $SPEAR_{pesticides}$ = Species at risk from pesticides. %EPT = Percentage of taxa belonging to Ephemeroptera, Plecoptera and Trichoptera relative to the total number of taxa in the benthic macroinvertebrate community. TP = total phosphorous.

Importance of pesticide toxicity confirmed by robust variable selection

Stability analysis and bootstrapping revealed a core set of variables consistently selected across perturbed datasets under five alternative statistical models (a conventional, non-penalized AIC-selected linear regression and four penalized regression methods: least absolute shrinkage and selection operator regression (Lasso), elastic net regression (Enet), minimax convex penalty (MCP), and triangulated (combined Lasso, Enet, and MCP) models; Fig. 4), indicating that the explanatory variables are robust and reproducibly associated with ecological metrics. Among these, pesticide toxicity stood out as the most consistently selected variable, which was retained among the top three stressors in all five models for each response variable (Fig. 4; Fig. A4, A5). Specifically, for the $SPEAR_{pesticides}$ metric, all five models ranked pesticide toxicity among the most consistently selected variables alongside TP (stability, i.e., selection frequency across bootstrap iterations: 97–100% and 97–100%, respectively). Pesticide toxicity estimates were similarly ranked first to third in models for %EPT (stability = 99–100%) and the saprobic index (stability = 81–95%). To test whether this pattern holds under a weaker pesticide signal, we repeated the stability analysis and bootstrapping using the TUsum before macroinvertebrate sampling, which represents the toxicity metric with the lowest explained variance (Fig. 3c). This confirmed the stable variable selection: estimated pesticide toxicity based on the TUsum before macroinvertebrate sampling was consistently among the five most important variables across ecological metrics and bootstrap resampling methods ($SPEAR_{pesticides}$ model: stability= 78–95%; %EPT model: stability = 91–98%; Saprobic index: stability = 78–92%). Although variable rankings in multiple-stressor frameworks are generally sensitive to model structure and variable collinearity [19,20,54], the consistent selection of pesticide toxicity across both penalized and non-penalized methods confirms that this finding is not an artefact of any particular modelling strategy, underscoring the value of stability analyses in such contexts.

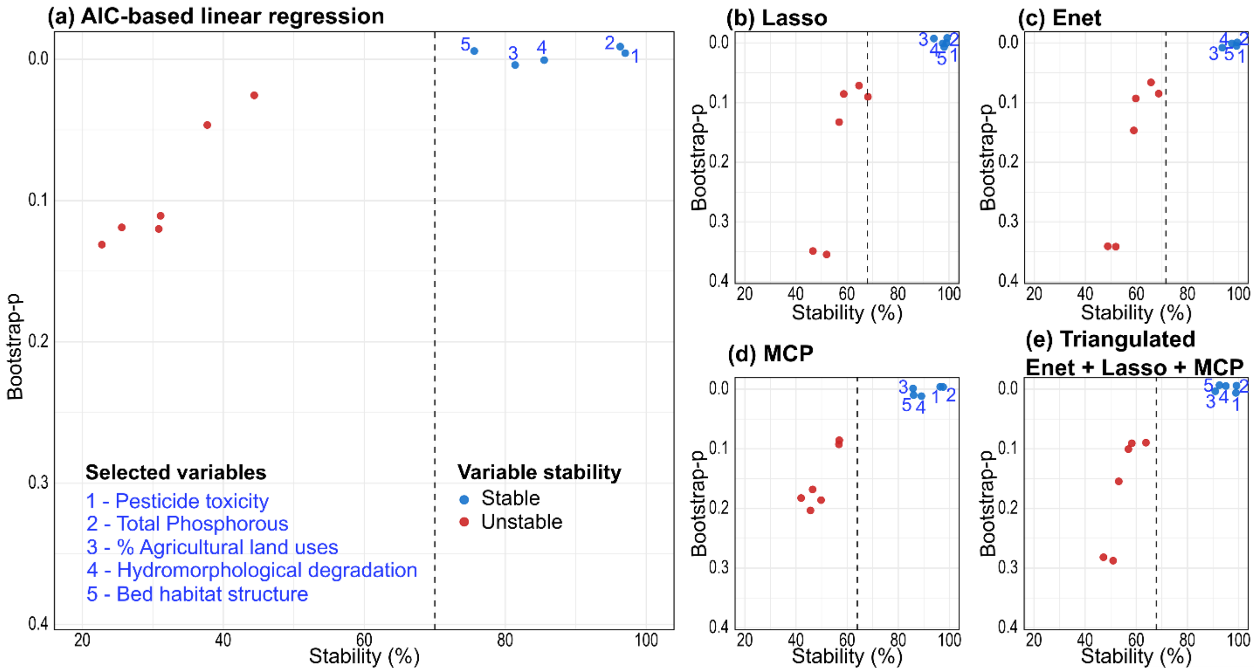

Figure 4. Stability of variable selection across five statistical models, with the case of the $\text{SPEAR}_{\text{pesticides}}$ and pesticide toxicity (i.e., TUsum before macroinvertebrate sampling and excluding high peaks): (a) AIC-based linear regression – Stepwise linear regression model using the Akaike Information Criterion, (b) Lasso – least absolute shrinkage and selection operator regression, (c) Enet – elastic net regression, (d) MCP – minimax convex penalty, and (e) triangulated method. The dashed vertical line is the threshold defining stable versus unstable variables. $\text{SPEAR}_{\text{pesticides}}$ = Species at risk from pesticides. TUsum = sum toxic unit.

CONCLUSIONS

We scrutinised the robustness and reproducibility of the central finding of Liess et al.[13] that pesticide mixtures are a key driver of macroinvertebrate community decline in small agricultural streams across different choices in data analysis. Our analysis encompassed an evaluation of alternative data aggregation and modelling choices partly raised in the critique by Moore and Rathjens[14].

Our reanalysis confirmed that pesticide toxicity estimates from Liess et al.[13] were reproducible but required reconstructing missing calculation steps and scripts through communication with the authors. We recommend that fully reproducible computer code is published along with the raw data of each study, but that, where reproducibility issues are encountered, early collaboration with original data providers is sought.

Beyond reproducibility, our evaluation of alternative data aggregation choices revealed that toxicity estimates affect ecological interpretation. Summing mixture toxicity and including both event and grab samples improved the representation of toxicity gradients. The results suggest that baseline exposure can be a more important driver of macroinvertebrate responses than short-term peaks. This very likely depends on the toxicokinetics of the substances, and future studies should test the relationship between toxicokinetic parameters, sampling strategy, and high-toxicity peaks. More broadly, future assessments of multiple stressor effects should carefully consider data aggregation choices for both pesticide toxicity metrics and other abiotic stressors, as these can substantially influence the relative importance of individual stressors. Similarly, the choice of ecological metrics needs scrutiny, as they respond to different sets of stressors, despite a high intercorrelation. Lastly, we recommend applying model validation and stability analyses to confirm findings of influential stressors from multiple linear regression analyses.

Overall, the results point to a group of core stressors, including pesticide mixtures, nutrients, agricultural land use, hydromorphological degradation, and others highlighted by different ecological metrics, which collectively shape stream macroinvertebrate communities. In light of the broader reproducibility crisis in science, these findings underscore the importance of open workflows, shared analytical code, and multi-model inference as essential best practices for enhancing data robustness and reducing analytical uncertainty[54,55].

ASSOCIATED CONTENT

CRediT author statement

Hanh H. Nguyen: Conceptualization, Methodology, Data curation, Formal analysis, Visualization, Writing- Original draft preparation **Verena C. Schreiner**: Conceptualization, Data curation, Visualization, Writing- Reviewing and Editing. **Ralf B. Schäfer**: Conceptualization, Supervision, Funding acquisition, Resources, Validation, Writing- Reviewing and Editing.

Data and Code Availability

The complete analytical workflow, including R scripts for data aggregation, toxicity estimation, multiple stressor modelling, and stability analysis, is archived on (https://github.com/hhn365/Pesticide-Multiple-stressors-ranking-KgM). Raw monitoring data were obtained from the publicly available source at Liess et al. (2021b).

ACKNOWLEDGEMENTS

We acknowledge the help of the main authors of Liess et al.[13] and Moore and Rathjens[14] in resolving issues of reproducibility. This work was supported by the European Partnership for the Assessment of Risks from Chemicals (PARC), and all authors received funding from the European Union's Horizon Europe Research and Innovation Programme under Grant Agreement No. 101057014.

REFERENCES

(1) Hallmann, C. A.; Foppen, R. P. B.; Van Turnhout, C. A. M.; De Kroon, H.; Jongejans, E. Declines in Insectivorous Birds Are Associated with High Neonicotinoid Concentrations. *Nature* **2014**, *511* (7509), 341–343. https://doi.org/10.1038/nature13531.

(2) Sayer, C. A.; Fernando, E.; Jimenez, R. R.; Macfarlane, N. B. W.; Rapacciuolo, G.; Böhm, M.; Brooks, T. M.; Contreras-MacBeath, T.; Cox, N. A.; Harrison, I.; Hoffmann, M.; Jenkins, R.; Smith, K. G.; Vié, J.-C.; Abbott, J. C.; Allen, D. J.; Allen, G. R.; Barrios, V.; Boudot, J.-P.; Carrizo, S. F.; Charvet, P.; Clausnitzer, V.; Congiu, L.; Crandall, K. A.; Cumberlidge, N.; Cuttelod, A.; Dalton, J.; Daniels, A. G.; De Grave, S.; De Knijf, G.; Dijkstra, K.-D. B.; Dow, R. A.; Freyhof, J.; García, N.; Gessner, J.; Getahun, A.; Gibson, C.; Gollock, M. J.; Grant, M. I.; Groom, A. E. R.; Hammer, M. P.; Hammerson, G. A.; Hilton-Taylor, C.; Hodgkinson, L.; Holland, R. A.; Jabado, R. W.; Juffe Bignoli, D.; Kalkman, V. J.; Karimov, B. K.; Kipping, J.; Kottelat, M.; Lalèyè, P. A.; Larson, H. K.; Lintermans, M.; Lozano, F.; Ludwig, A.; Lyons, T. J.; Máiz-Tomé, L.; Molur, S.; Ng, H. H.; Numa, C.; Palmer-Newton, A. F.; Pike, C.; Pippard, H. E.; Polaz, C. N. M.; Pollock, C. M.; Raghavan, R.; Rand, P. S.; Ravelomanana, T.; Reis, R. E.; Rigby, C. L.; Scott, J. A.; Skelton, P. H.; Sloat, M. R.; Snoeks, J.; Stiassny, M. L. J.; Tan, H. H.; Taniguchi, Y.; Thorstad, E. B.; Tognelli, M. F.; Torres, A. G.; Torres, Y.; Tweddle, D.; Watanabe, K.; Westrip, J. R. S.; Wright, E. G. E.; Zhang, E.; Darwall, W. R. T. One-Quarter of Freshwater Fauna Threatened with Extinction. *Nature* **2025**, *638* (8049), 138–145. https://doi.org/10.1038/s41586-024-08375-z.

(3) Leuenberger, W.; Doser, J. W.; Belitz, M. W.; Ries, L.; Haddad, N. M.; Thogmartin, W. E.; Zipkin, E. F. Three Decades of Declines Restructure Butterfly Communities in the Midwestern United States. *Proc. Natl. Acad. Sci.* **2025**, *122* (33), e2501340122. https://doi.org/10.1073/pnas.2501340122.

(4) Wagner, D. L.; Grames, E. M.; Forister, M. L.; Berenbaum, M. R.; Stopak, D. Insect Decline in the Anthropocene: Death by a Thousand Cuts. *Proc. Natl. Acad. Sci.* **2021**, *118* (2), e2023989118. https://doi.org/10.1073/pnas.2023989118.

(5) Nguyen, H. H.; Schürings, C.; Welti, E. A. R.; Sundermann, A.; Trepel, M.; Gericke, A.; Venohr, M. Stricter Reductions of Nutrient Pollution Support Riverine Community Recovery in Degraded Catchments. *J. Environ. Manage.* **2025**, *395*, 127909. https://doi.org/10.1016/j.jenvman.2025.127909.

(6) Schürings, C.; Globevnik, L.; Lemm, J. U.; Psomas, A.; Snoj, L.; Hering, D.; Birk, S. River Ecological Status Is Shaped by Agricultural Land Use Intensity across Europe. *Water Res.* **2024**, *251*, 121136. https://doi.org/10.1016/j.watres.2024.121136.

(7) Brauns, M.; Allen, D. C.; Boëchat, I. G.; Cross, W. F.; Ferreira, V.; Graeber, D.; Patrick, C. J.; Peipoch, M.; Von Schiller, D.; Gücker, B. A Global Synthesis of Human Impacts on the Multifunctionality of Streams and Rivers. *Glob. Change Biol.* **2022**, *28* (16), 4783–4793. https://doi.org/10.1111/gcb.16210.

(8) Schäfer, R. B.; Caquet, T.; Siimes, K.; Mueller, R.; Lagadic, L.; Liess, M. Effects of Pesticides on Community Structure and Ecosystem Functions in Agricultural Streams of Three Biogeographical Regions in Europe. *Sci. Total Environ.* **2007**, *382* (2–3), 272–285. https://doi.org/10.1016/j.scitotenv.2007.04.040.

(9) Von Gönner, J.; Gröning, J.; Grescho, V.; Neuer, L.; Gottfried, B.; Hänsch, V. G.; Molsberger-Lange, E.; Wilharm, E.; Liess, M.; Bonn, A. Citizen Science Shows That Small Agricultural Streams in Germany Are in a Poor Ecological Status. *Sci. Total Environ.* **2024**, *922*, 171183. https://doi.org/10.1016/j.scitotenv.2024.171183.

(10) Szöcs, E.; Brinke, M.; Karaoglan, B.; Schäfer, R. B. Large Scale Risks from Agricultural Pesticides in Small Streams. *Environ. Sci. Technol.* **2017**, *51* (13), 7378–7385. https://doi.org/10.1021/acs.est.7b00933.

(11) Wittmer, I. K.; Bader, H.-P.; Scheidegger, R.; Singer, H.; Lück, A.; Hanke, I.; Carlsson, C.; Stamm, C. Significance of Urban and Agricultural Land Use for Biocide and Pesticide Dynamics in Surface Waters. *Water Res.* **2010**, *44* (9), 2850–2862. https://doi.org/10.1016/j.watres.2010.01.030.

(12) Schulz, R.; Liess, M. Toxicity of Fenvalerate to Caddisfly Larvae: Chronic Effects of 1- vs 10-h Pulse-Exposure with Constant Doses. *Chemosphere* **2000**, *41* (10), 1511–1517. https://doi.org/10.1016/S0045-6535(00)00107-7.

(13) Liess, M.; Liebmann, L.; Vormeier, P.; Weisner, O.; Altenburger, R.; Borchardt, D.; Brack, W.; Chatzinotas, A.; Escher, B.; Foit, K.; Gunold, R.; Henz, S.; Hitzfeld, K. L.; Schmitt-Jansen, M.; Kamjunke, N.; Kaske, O.; Knillmann, S.; Krauss, M.; Küster, E.; Link, M.; Lück, M.; Möder, M.; Müller, A.; Paschke, A.; Schäfer, R. B.; Schneeweiss, A.; Schreiner, V. C.; Schulze, T.; Schüürmann, G.; Von Tümpling, W.; Weitere, M.; Wogram, J.; Reemtsma, T. Pesticides Are the Dominant Stressors for Vulnerable Insects in Lowland Streams. *Water Res.* **2021**, *201*, 117262. https://doi.org/10.1016/j.watres.2021.117262.

(14) Moore, D. R. J.; Rathjens, H. Are Pesticides the Dominant Stressors in German Lowland Streams? *Integr. Environ. Assess. Manag.* **2025**, *21* (4), 739–744. https://doi.org/10.1093/inteam/vjaf038.

(15) Schäfer, R. B.; Gerner, N.; Kefford, B. J.; Rasmussen, J. J.; Beketov, M. A.; De Zwart, D.; Liess, M.; Von Der Ohe, P. C. How to Characterize Chemical Exposure to Predict Ecologic Effects on Aquatic Communities? *Environ. Sci. Technol.* **2013**, *47* (14), 7996–8004. https://doi.org/10.1021/es4014954.

(16) Schriever, C.; Jene, B.; Resseler, H.; Spatz, R.; Sur, R.; Weyers, A.; Winter, M. The European Regulatory System for Plant Protection Products—Cause of a “Silent Spring” or Highly Advanced and Protective? *Integr. Environ. Assess. Manag.* **2025**, *21* (1), 3–19. https://doi.org/10.1093/inteam/vjae007.

(17) Heinze, G.; Dunkler, D. Five Myths about Variable Selection. *Transpl. Int.* **2017**, *30* (1), 6–10. https://doi.org/10.1111/tri.12895.

(18) Burgess, B. J.; Jackson, M. C.; Murrell, D. J. Are Experiment Sample Sizes Adequate to Detect Biologically Important Interactions between Multiple Stressors? *Ecol. Evol.* **2022**, *12* (9), e9289. https://doi.org/10.1002/ece3.9289.

(19) White, J. W.; Boote, K. J.; Kimball, B. A.; Porter, C.; Salmeron, M.; Shelia, V.; Thorp, K. R.; Hoogenboom, G. From Field to Analysis: Strengthening Reproducibility and Confirmation in Research for Sustainable Agriculture. *Npj Sustain. Agric.* **2025**, *3* (1), 27. https://doi.org/10.1038/s44264-025-00067-z.

(20) Lima, E.; Hyde, R.; Green, M. Model Selection for Inferential Models with High Dimensional Data: Synthesis and Graphical Representation of Multiple Techniques. *Sci. Rep.* **2021**, *11* (1), 412. https://doi.org/10.1038/s41598-020-79317-8.

(21) Meinshausen, N.; Bühlmann, P. Stability Selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2010**, *72* (4), 417–473. https://doi.org/10.1111/j.1467-9868.2010.00740.x.

(22) Liess, M.; Liebmann, L.; Vormeier, P.; Weisner, O.; Altenburger, R.; Borchardt, D.; Brack, W.; Chatzinotas, A.; Escher, B.; Foit, K.; Gunold, R.; Henz, S.; Hitzfeld, K. L.; Schmitt-Jansen, M.; Kamjunke, N.; Kaske, O.; Knillmann, S.; Krauss, M.; Küster, E.; Link, M.; Lück, M.; Möder, M.; Müller, A.; Paschke, A.; Schäfer, R. B.; Schneeweiss, A.; Schreiner, V. C.; Schulze, T.; Schüürmann, G.; von Tümpling, W.; Weitere, M.; Wogram, J.; Reemtsma, T. The Lowland Stream Monitoring Dataset (KgM, Kleingewässer-Monitoring) 2018, 2019, 2021, 52.3 MBytes. https://doi.org/10.1594/PANGAEA.931673.

(23) Halbach, K.; Möder, M.; Schrader, S.; Liebmann, L.; Schäfer, R. B.; Schneeweiss, A.; Schreiner, V. C.; Vormeier, P.; Weisner, O.; Liess, M.; Reemtsma, T. Small Streams–Large Concentrations? Pesticide Monitoring in Small Agricultural Streams in Germany during Dry Weather and Rainfall. *Water Res.* **2021**, *203*, 117535. https://doi.org/10.1016/j.watres.2021.117535.

(24) Liess, M.; Ohe, P. C. V. D. Analyzing Effects of Pesticides on Invertebrate Communities in Streams. *Environ. Toxicol. Chem.* **2005**, *24* (4), 954–965. https://doi.org/10.1897/03-652.1.

(25) Kolkwitz, R.; Marsson, M. Ökologie der tierischen Saprobien. Beiträge zur Lehre von der biologischen Gewässerbeurteilung. *Int. Rev. Gesamten Hydrobiol. Hydrogr.* **1909**, *2* (1–2), 126–152. https://doi.org/10.1002/iroh.19090020108.

(26) Lenat, D. R. Water Quality Assessment of Streams Using a Qualitative Collection Method for Benthic Macroinvertebrates. *J. North Am. Benthol. Soc.* **1988**, *7* (3), 222–233. https://doi.org/10.2307/1467422.

(27) Helsel, D. R. Fabricating Data: How Substituting Values for Nondetects Can Ruin Results, and What Can Be Done about It. *Chemosphere* **2006**, *65* (11), 2434–2439. https://doi.org/10.1016/j.chemosphere.2006.04.051.

(28) Grömping, U. Relative Importance for Linear Regression in *R*: The Package **Relaimpo**. *J. Stat. Softw.* **2006**, *17* (1). https://doi.org/10.18637/jss.v017.i01.

(29) Venables, W. N.; Ripley, B. D. *Modern Applied Statistics with S*; Chambers, J., Eddy, W., Härdle, W., Sheather, S., Tierney, L., Series Eds.; Statistics and Computing; Springer New York: New York, NY, 2002. https://doi.org/10.1007/978-0-387-21706-2.

(30) Kuhn, M.; Johnson, K. *Feature Engineering and Selection: A Practical Approach for Predictive Models*, 1st ed.; Chapman and Hall/CRC, 2019. https://doi.org/10.1201/9781315108230.

(31) Bivand, R. S.; Wong, D. W. S. Comparing Implementations of Global and Local Indicators of Spatial Association. *TEST* **2018**, *27* (3), 716–748. https://doi.org/10.1007/s11749-018-0599-x.

(32) Zou, H.; Hastie, T. Regularization and Variable Selection Via the Elastic Net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2005**, *67* (2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

(33) Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1996**, *58* (1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

(34) Zhang, C.-H. Nearly Unbiased Variable Selection under Minimax Concave Penalty. *Ann. Stat.* **2010**, *38* (2). https://doi.org/10.1214/09-AOS729.

(35) R Core Team. R: A Language and Environment for Statistical Computing, 2024. https://www.R-project.org/.

(36) Wickham, H. *Ggplot2*; Use R!; Springer International Publishing: Cham, 2016. https://doi.org/10.1007/978-3-319-24277-4.

(37) Gustavsson, M.; Kreuger, J.; Bundschuh, M.; Backhaus, T. Pesticide Mixtures in the Swedish Streams: Environmental Risks, Contributions of Individual Compounds and Consequences of Single-Substance Oriented Risk Mitigation. *Sci. Total Environ.* **2017**, *598*, 973–983. https://doi.org/10.1016/j.scitotenv.2017.04.122.

(38) Schreiner, V. C.; Link, M.; Kunz, S.; Szöcs, E.; Scharmüller, A.; Vogler, B.; Beck, B.; Battes, K. P.; Cimpean, M.; Singer, H. P.; Hollender, J.; Schäfer, R. B. Paradise Lost? Pesticide Pollution in a European Region with Considerable Amount of Traditional Agriculture. *Water Res.* **2021**, *188*, 116528. https://doi.org/10.1016/j.watres.2020.116528.

(39) Spycher, S.; Kalf, D.; Lahr, J.; Gönczi, M.; Lindström, B.; Pace, E.; Botta, F.; Bougon, N.; Staub, P.-F.; Hitzfeld, K. L.; Weisner, O.; Junghans, M.; Kroll, A. Linking Chemical Surface Water Monitoring and Pesticide Regulation in Selected European Countries. *Environ. Sci. Pollut. Res.* **2024**, *31* (30), 43432–43450. https://doi.org/10.1007/s11356-024-33865-y.

(40) Betz-Koch, S.; Jacobs, B.; Oehlmann, J.; Ratz, D.; Reutter, C.; Wick, A.; Oetken, M. Pesticide Dynamics in Three Small Agricultural Creeks in Hesse, Germany. *PeerJ* **2023**, *11*, e15650. https://doi.org/10.7717/peerj.15650.

(41) Bighiu, M. A.; Höss, S.; Traunspurger, W.; Kahlert, M.; Goedkoop, W. Limited Effects of Pesticides on Stream Macroinvertebrates, Biofilm Nematodes, and Algae in Intensive Agricultural Landscapes in Sweden. *Water Res.* **2020**, *174*, 115640. https://doi.org/10.1016/j.watres.2020.115640.

(42) Schäfer, R. B.; Pettigrove, V.; Rose, G.; Allinson, G.; Wightwick, A.; Von Der Ohe, P. C.; Shimeta, J.; Kühne, R.; Kefford, B. J. Effects of Pesticides Monitored with Three Sampling Methods in 24 Sites on Macroinvertebrates and Microorganisms. *Environ. Sci. Technol.* **2011**, *45* (4), 1665–1672. https://doi.org/10.1021/es103227q.

(43) Betz-Koch, S.; Oehlmann, J.; Oetken, M. Timing Matters: Impact of Different Frequencies of Low Pesticide Pulses on Aquatic Invertebrates. *Environ. Sci. Eur.* **2025**, *37* (1), 55. https://doi.org/10.1186/s12302-025-01091-z.

(44) Ashauer, R.; Caravatti, I.; Hintermeister, A.; Escher, B. I. Bioaccumulation Kinetics of Organic Xenobiotic Pollutants in the Freshwater Invertebrate *Gammarus Pulex* Modeled with Prediction Intervals. *Environ. Toxicol. Chem.* **2010**, *29* (7), 1625–1636. https://doi.org/10.1002/etc.175.

(45) Wiberg-Larsen, P.; Nørum, U.; Rasmussen, J. J. Repeated Insecticide Pulses Increase Harmful Effects on Stream Macroinvertebrate Biodiversity and Function. *Environ. Pollut.* **2021**, *273*, 116404. https://doi.org/10.1016/j.envpol.2020.116404.

(46) Nowell, L. H.; Moran, P. W.; Waite, I. R.; Schmidt, T. S.; Bradley, P. M.; Mahler, B. J.; Van Metre, P. C. Multiple Lines of Evidence Point to Pesticides as Stressors Affecting Invertebrate Communities in Small Streams in Five United States Regions. *Sci. Total Environ.* **2024**, *915*, 169634. https://doi.org/10.1016/j.scitotenv.2023.169634.

(47) Link, M.; Schreiner, V. C.; Graf, N.; Szöcs, E.; Bundschuh, M.; Battes, K. P.; Cîmpean, M.; Sures, B.; Grabner, D.; Buse, J.; Schäfer, R. B. Pesticide Effects on Macroinvertebrates and Leaf Litter Decomposition in Areas with Traditional Agriculture. *Sci. Total Environ.* **2022**, *828*, 154549. https://doi.org/10.1016/j.scitotenv.2022.154549.

(48) Birk, S.; Bonne, W.; Borja, A.; Brucet, S.; Courrat, A.; Poikane, S.; Solimini, A.; Van De Bund, W.; Zampoukas, N.; Hering, D. Three Hundred Ways to Assess Europe's Surface Waters: An Almost Complete Overview of Biological Methods to Implement the Water Framework Directive. *Ecol. Indic.* **2012**, *18*, 31–41. https://doi.org/10.1016/j.ecolind.2011.10.009.

(49) Beketov, M. A.; Foit, K.; Schäfer, R. B.; Schriever, C. A.; Sacchi, A.; Capri, E.; Biggs, J.; Wells, C.; Liess, M. SPEAR Indicates Pesticide Effects in Streams – Comparative Use of Species- and Family-Level Biomonitoring Data. *Environ. Pollut.* **2009**, *157* (6), 1841–1848. https://doi.org/10.1016/j.envpol.2009.01.021.

(50) Agyekum, M. K.; Pathak, D.; Kindinger, A.; Kumar, R.; Borchardt, D.; Weitere, M.; Frank, K.; Schmitt-Jansen, M.; Büttner, O.; Brauns, M.; Fink, P.; Scharfenberger, U. A

Hydrologically Informed Agricultural Land Use Intensity Index for Assessing Ecological Impacts on Streams and Rivers. *Commun. Earth Environ.* **2025**, *6* (1), 991. https://doi.org/10.1038/s43247-025-02933-7.

(51) Cornejo, A.; Tonin, A. M.; Checa, B.; Tuñon, A. R.; Pérez, D.; Coronado, E.; González, S.; Ríos, T.; Macchi, P.; Correa-Araneda, F.; Boyero, L. Effects of Multiple Stressors Associated with Agriculture on Stream Macroinvertebrate Communities in a Tropical Catchment. *PLOS ONE* **2019**, *14* (8), e0220528. https://doi.org/10.1371/journal.pone.0220528.

(52) Liebmann, L.; Schreiner, V. C.; Vormeier, P.; Weisner, O.; Liess, M. Combined Effects of Herbicides and Insecticides Reduce Biomass of Sensitive Aquatic Invertebrates. *Sci. Total Environ.* **2024**, *946*, 174343. https://doi.org/10.1016/j.scitotenv.2024.174343.

(53) Rasmussen, J. J.; Wiberg-Larsen, P.; Baattrup-Pedersen, A.; Friberg, N.; Kronvang, B. Stream Habitat Structure Influences Macroinvertebrate Response to Pesticides. *Environ. Pollut.* **2012**, *164*, 142–149. https://doi.org/10.1016/j.envpol.2012.01.007.

(54) Munafò, M. R.; Nosek, B. A.; Bishop, D. V. M.; Button, K. S.; Chambers, C. D.; Percie Du Sert, N.; Simonsohn, U.; Wagenmakers, E.-J.; Ware, J. J.; Ioannidis, J. P. A. A Manifesto for Reproducible Science. *Nat. Hum. Behav.* **2017**, *1* (1), 0021. https://doi.org/10.1038/s41562-016-0021.

(55) Alston, J. M.; Rick, J. A. A Beginner's Guide to Conducting Reproducible Research. *Bull. Ecol. Soc. Am.* **2021**, *102* (2), e01801. https://doi.org/10.1002/bes2.1801.

Supplementary Information

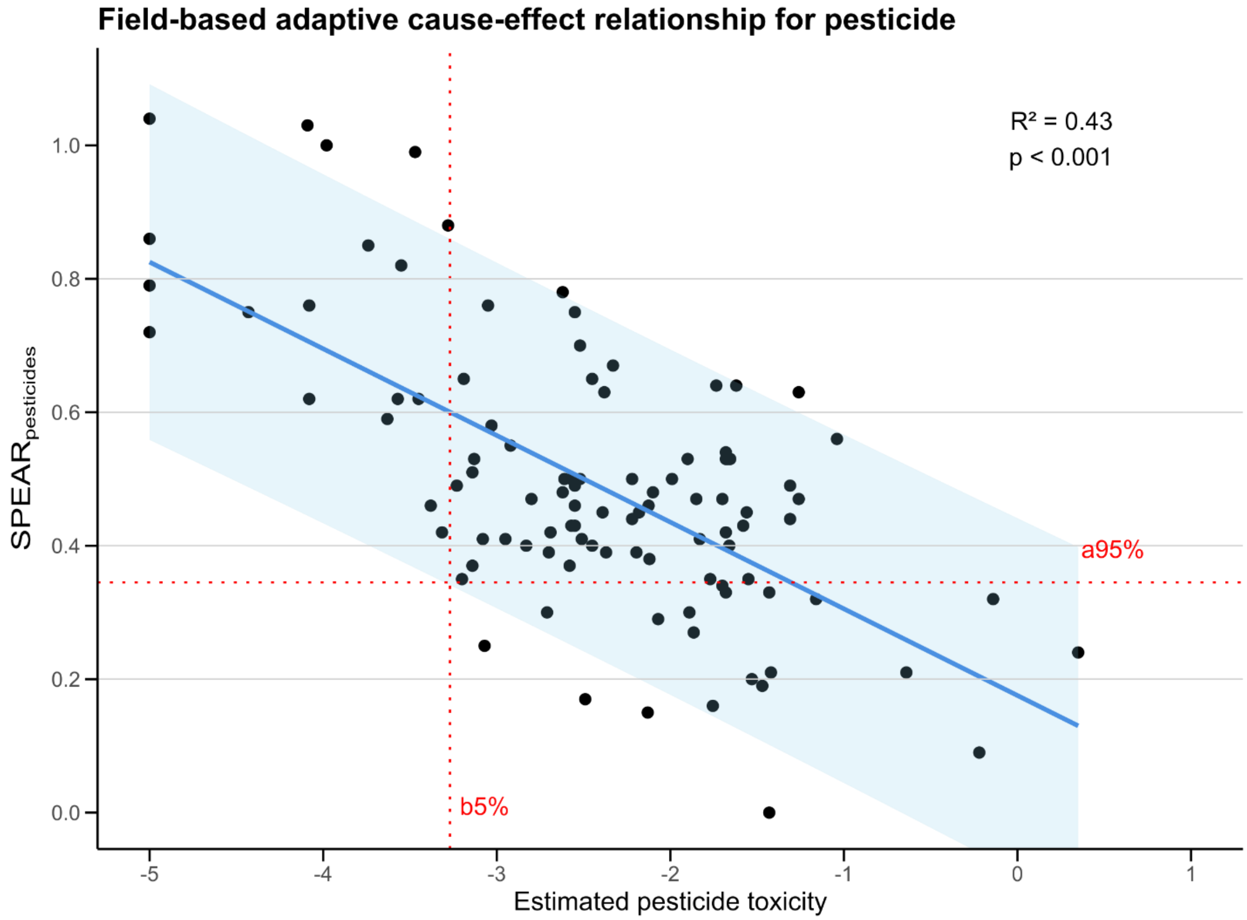

Figure A1. Reproduced negative association between maximum toxic unit (TUmax) and Species at risk from pesticides ($SPEAR_{pesticide}$) metrics based on the German Kleingewässermonitoring (KGM) dataset (Liess et al., 2021b). a95% - the line indicates the $SPEAR_{pesticides}$ benchmark for detecting unacceptable pesticide-related ecological effects at a 95% confidence level, derived by reducing the "good"–"moderate" classification boundary by 1.645 standard deviations of the linear regression. b5% - the line corresponds to a log TUmax threshold of −3.27, below which 5% of sampled streams exhibit an unacceptable ecological status based on $SPEAR_{pesticides}$, at a 95% confidence level. The detailed reproduction of TUmax calculation and correlation analysis with $SPEAR_{pesticide}$ are provided in the computer code.

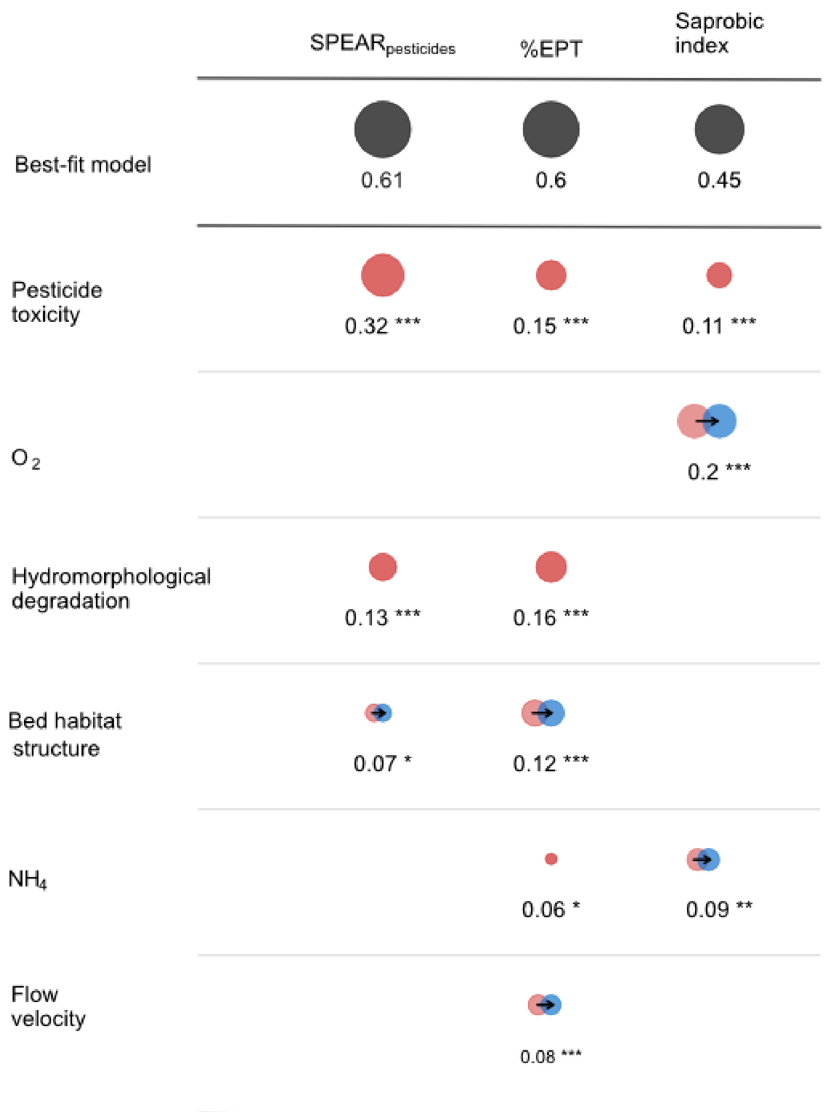

Figure A2. Reproduced relative importance of multiple stressors for three ecological metrics (SPEAR$_{pesticides}$, % EPT, saprobic index) from Liess et al. (2021a), which are based on full time series (i.e., including abiotic sampling after stream macroinvertebrate sampling) of thirteen abiotic variables (see 2.5.1.) and using the model structure with two-way interactions. Dots represent explained variances (numbers below dots), with red and blue indicating negative and positive estimates as shown in model outputs. Black arrows indicate relationships where estimate signs were corrected, accompanied by a change in colour. Significance: *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. SPEAR$_{pesticides}$ = Species at risk from pesticides. %EPT = Percentage of taxa belonging to Ephemeroptera, Plecoptera and Trichoptera relative to the total number of taxa in the benthic macroinvertebrate community. NH_4 = Ammonium. The detailed reproduction of multiple linear regressions with TUmax and other stressors are provided in the computer code.

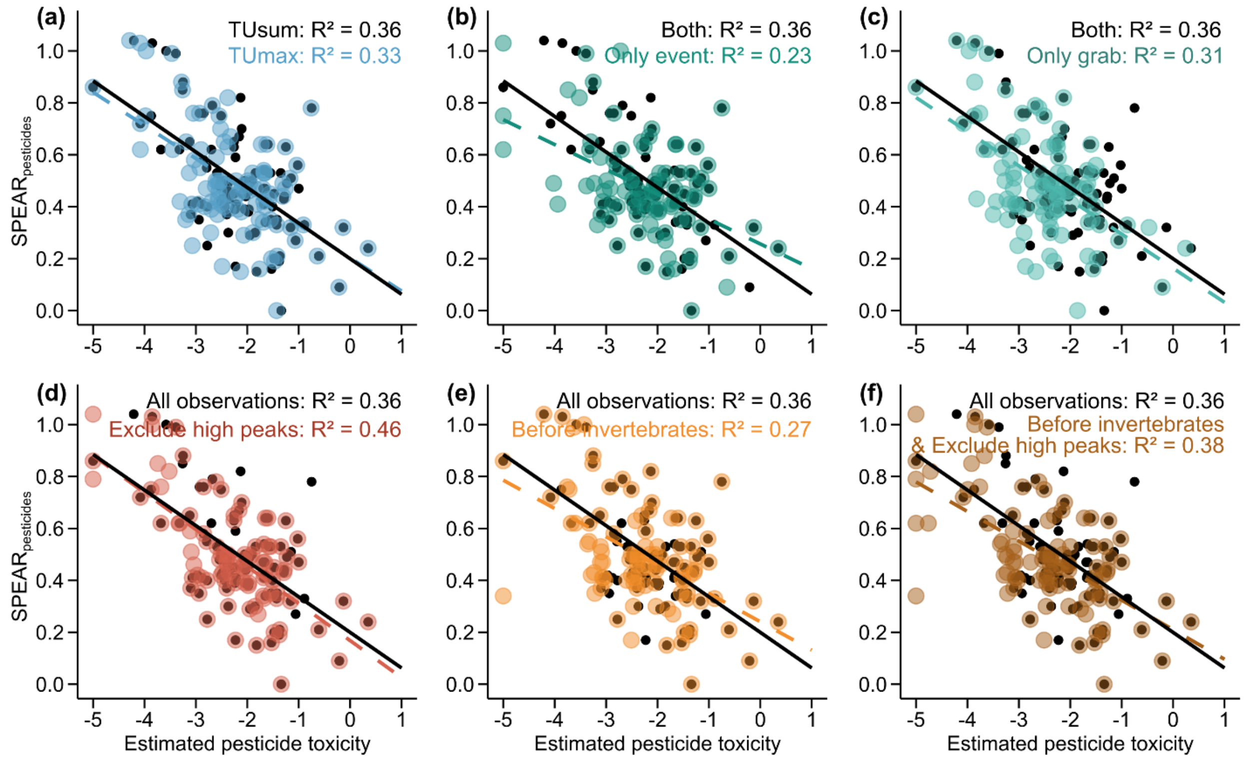

Figure A3. Association between different pesticide toxicity estimates and Species at risk from pesticides (SPEAR$_{pesticides}$), expressed as explained variance (R^2) from linear regression. Pesticide toxicity estimates compared to sum toxic unit (labelled as TUsum; Both; All observations): (a) Max toxic unit (TUmax); (b) event (only event); (c) grab (only grab); (d) Excluding high peaks; (e) Restricted to before macroinvertebrate sampling (before macroinvertebrates); and (f) both (before macroinvertebrates and exclude high peaks).

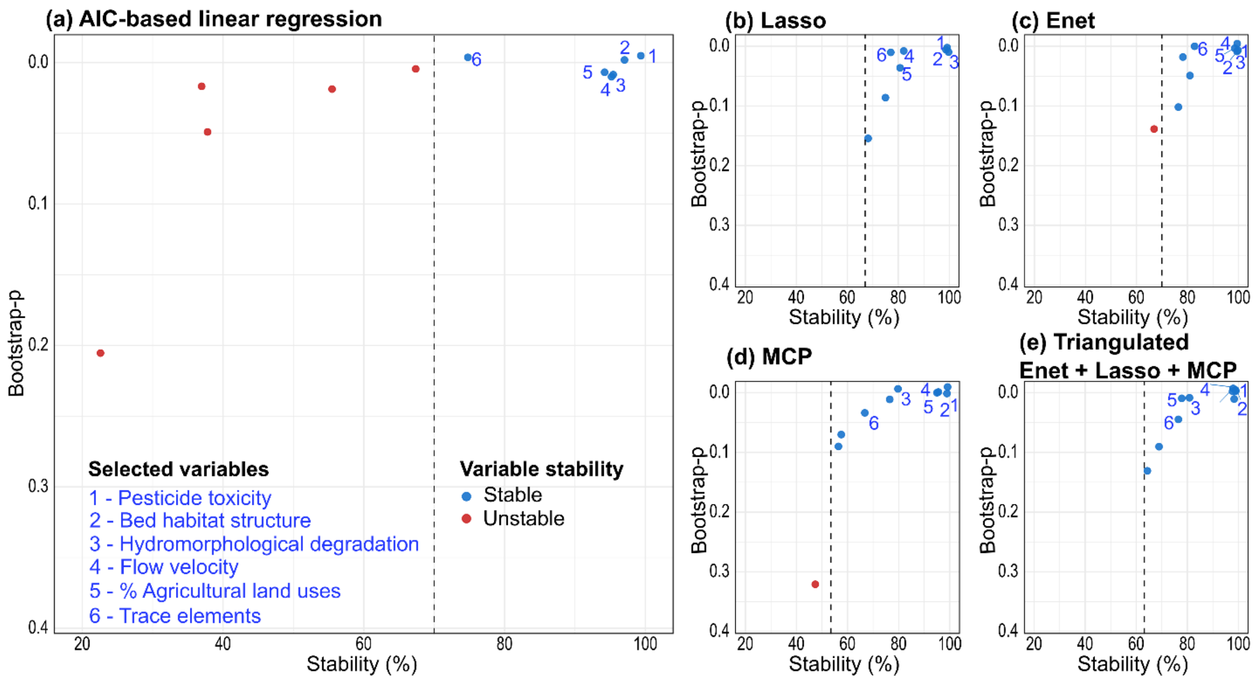

Figure A4. Stability of variable selection across five statistical models, with the case of %EPT and TUsum before macroinvertebrate sampling and excluding high peaks: AIC-based linear regression – Stepwise linear regression model using the Akaike Information Criterion, Lasso – least absolute shrinkage and selection operator regression, Enet – elastic net regression, MCP – minimax convex penalty, and triangulated method. The dashed vertical line is the threshold defining stable versus unstable variables. %EPT = Percentage of taxa belonging to Ephemeroptera, Plecoptera and Trichoptera relative to the total number of taxa in the benthic macroinvertebrate community. TUsum = sum toxic unit.

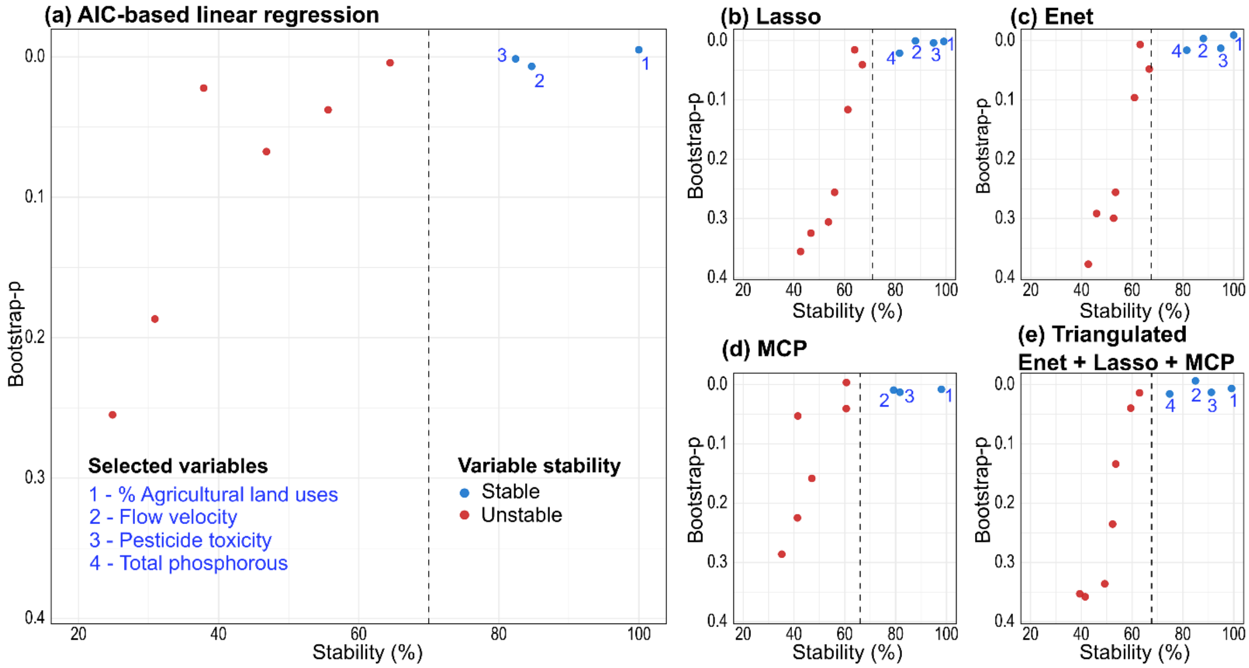

Figure A5. Stability of variable selection across five statistical models, with the case of the saprobic index and TUsum before macroinvertebrate sampling and exclude high peaks: AIC-based linear regression – Stepwise linear regression model using the Akaike Information Criterion, Lasso – least absolute shrinkage and selection operator regression, Enet – elastic net regression, MCP – minimax convex penalty, and triangulated method. The dashed vertical line is the threshold defining stable versus unstable variables. TUsum = sum toxic unit.