An investigation into the impact of nine catchment characteristics on the accuracy of two phosphorus load apportionment models

Stevenson, J. L. (ORCID: 0000-0003-2042-9130)[a], O'Riordain, S. (ORCID: 0000-0002-1266-0845)[b], Harris, W. E. (ORCID: 0000-0002-9038-8656)[a], Crockford, L. (ORCID: 0000-0001-8336-4149)[a*],

[a] Agriculture and Environment Dept., Harper Adams University, Edgmond, Shropshire, UK

[b] School of Computer Science and Statistics, Trinity College Dublin, Ireland

*Corresponding author

Agriculture and Environment
Harper Adams University
Edgmond
Shropshire
TF10 8NB
UK
+44 (0)1952 815476
lcrockford@harper-adams.ac.uk

Keywords

phosphorus, load apportionment model, modelling, certainty, catchment

Declarations

Funding No funding was provided for this study aside from supervisory support through Harper Adams University.
Conflicts of interest/Competing interests NA
Availability of data and material Data were sourced from the National River Flow Archive and the EA historical river water quality. These were obtained under licence and are unavailable for distribution without the express permission of NRFA or the EA.
Code availability Packages in R were used for data analyses, namely phoslam, randomForest':randomForest, hydroGOF':*pbias,* and base packages. The code for this is available on request.
Authors' contributions The primary author completed the majority of analyses of data, interpretation of findings and production of manuscript. Second author provided expertise in model and method development. Third author provided expertise in data analysis and development of methods. Fourth author provided supervisory support, aid in interpretation of data, and final manuscript preparation for submission.

Abstract

Phosphorus (P) load apportionment models (LAMs), requiring only spatially and temporally paired P and flow (Q) measurements, provide outputs of variable accuracy using long-term monthly datasets. Using a novel approach to investigate the impact of catchment characteristics on accuracy variation, 91 watercourses Q-P datasets were applied to two LAMs, BM and GM, and bootstrapped to ascertain standard errors (SEs). Random forest and regression analysis on data pertaining to catchments' land use, steepness, size, base flow and sinuosity were used to identify the individual relative importance of a variable on SE. For BM, increasing urban cover was influential on raising SEs, accounting for c.19% of observed variation, whilst analysis for GM found no individually important catchment characteristic. Assessment of model fit evidenced BM consistently outperformed GM, modelling P values to ± 10% of actual P values in 85.7% of datasets, as opposed to 17.6% by GM. Further catchment characteristics are needed to account for SE variation within both models, whilst interaction between variables may also be present. Future research should focus on quantifying these possible interactions and should expand catchment characteristics included within the random forest. Both LAMs must also be tested on a wide range of high temporal resolution datasets to ascertain if they can adequately model storm events in catchments with diverse characteristics.

Introduction

The trophic status and risk of eutrophication within watercourses is heavily influenced by phosphorus (P) concentrations (Sharpley, 2016; Omari et al., 2019). So severe is the threat posed by the nutrient that excessive presence is the most common reason for failure to achieve Good Ecological Status, as defined by the Water Framework Directive (2000), in UK waterbodies (Leaf, 2018). To effectively target resources at reducing P loads, accurate identification of the nutrient's origin is required (Bowes et al*.,* 2014), with alternative load apportionment models (LAMs) proposed by Bowes et al*.* (2008) and Greene et al*.* (2011) to undertake this task; henceforth referred to as BM and GM respectively. Both models require spatially and temporally matched P and flow (Q) measurements, meaning they offer a cost- and labour-efficient tool compared to export coefficient and geographical information systems-based approaches (Bowes et al*.,* 2008; Greene et al*.,* 2011). The models exploit an, ostensibly, fundamental difference in the observed Q-P relationship when P is derived from point sources, such as wastewater treatment plants, or diffuse sources, such as agricultural fertiliser. The former is largely independent of river flow, as P does not usually require transport to the watercourse via rainfall, whereas the latter is dependent on mobilisation via precipitation. Therefore, in point source dominated rivers P concentration should decrease as a function of Q, due to dilution, whereas the opposite would be true for diffuse pollution. Details of model functions and dissimilarities are available in Crockford et al. (2017).

Despite initial studies asserting their accuracy (Bowes et al., 2008; Bowes et al., 2009; Bowes et al., 2010; Greene et al*.,* 2011), Crockford et al. (2017) found both LAMs (BM and GM) are prone to substantial errors by calculating certainty statistics for each model under varying sampling temporal frequencies. The authors concluded this having used high frequency data from a river in Ireland and the statistical method of bootstrapping (Efron, 1979) to enable the calculation of standard errors (SEs) when the LAMs were applied to Q-P datasets. Crockford et al. (2017) went on to make the recommendation of using bootstrapping to ascertain accuracy levels of further datasets to understand the applicability and reliability of these LAMs. By doing so in a diverse range of catchments, statistical analysis of catchment characteristics could infer their influence on LAM accuracy, and may provide further insight into where the models would be best utilised or avoided. Validating the accuracy of these modelling methods is extremely important, as they continue to be used to apportion P load in rivers, e.g. BM has recently been used to forecast the impact of climate change influences on P loadings, realising the possible application of these models in varied catchments (Charlton et al., 2018).

To address this knowledge gap, secondary Q and P data from 136 watercourses (Figure 1) throughout Britain were used to calculate point source apportionment according to both BM and GM, with results bootstrapped (N=2000) and applied to high frequency Q data to provide SE estimates for each method. The data used here comprised all that were available from the Environment Agency and the National River Flow Archive (NRFA) constrained by proximity as explained in the methodology. Therefore, these datasets are typical of those used by local authorities to apportion P load in a river catchment. Land use, base flow index, holistic catchment steepness, watercourse sinuosity and catchment size data were then obtained, or calculated, for each catchment to facilitate investigation into the importance of these variables on SE of model outputs. This provides a novel method for evaluating model output variability and a framework for elucidating the drivers for model error in future studies.

Material and methods

Selection of catchment metrics

Catchment characteristics (Land Use; Baseflow Index; Catchment Steepness; Catchment Sinuosity; Catchment Size) were selected given evidence their variability may impact observed Q-P mechanisms, which in turn could affect assumptions of the algorithms behind each LAM. For instance, land use causes alteration in Q flow paths, the level, dominant form and source of P (MacDonald et al., 2012; Daryanto et al., 2017; Rogger et al., 2017; Lou et al., 2018). Baseflow index is representative of catchment geology and soil type (Yaeger et al., 2012), the properties of which will influence P retention (Antoniadis et al., 2016) and Q dynamics such as residence times (Maxwell et al., 2016). Catchment steepness can cause an increase in soil erosion (Bridge and Demicco, 2008) and consequently the transport of soil adsorbed P to watercourses, whilst increased sinuosity encourages sedimentation (He et al., 2018), that also facilitates P adsorption. The release of this adsorbed P can occur at high flows, indicating diffuse sources regardless of actual point source contributions (Jarvie et al., 2012). Finally, catchment size increases can enable observation of Q variations over a longer period post rainfall event in comparison to smaller catchments (Crochemore et al., 2018).

Acquisition of secondary phosphorus (P) and river flow (Q) data

Water quality datasets from 2010 to 2019 were obtained from the Environment Agency website (EA, not dated) providing data for England only. Datasets were combined, filtered to remove information pertaining to other water quality measures, and grouped according to their co-ordinates. Locations with fewer than 50 data points were identified using Microsoft Excel COUNTIF function and removed, leaving 3358 potentially eligible datasets (dependant on Q data availability). The threshold of fewer than 50 data points was arbitrary, defined to ensure sufficient data points for the process described in “Data preparation for Load Apportionment Modelling”, where data point removal was anticipated, leaving sufficient numbers of data points remaining for statistical robustness.

The NRFA provided coordinates of all UK river flow gauging stations (NRFAa, 2019). These were plotted in ArcMap 10.5.1 (ESRI, 2019) and overlaid with P sampling locations and a shapefile containing UK rivers (OS, 2019) to facilitate visual identification of gauging stations located on the same watercourses as P data locations. As P and Q data are collected by different agencies in the UK there were few locations where these data were spatially matched. Therefore, data for Q (15 minute interval) and P (collected monthly) were obtained from locations on the same stem of a river, with no watercourse entering or exiting in-between for the period 2010 to 2019. This yielded 136 eligible datasets for analysis.

Data preparation for Load Apportionment Modelling

R (R Core Team, 2019) was used to pair P data points to Q data points of the closest temporal proximity, and to calculate the mean of Q data within a one hour around this point. Creating an hourly average standardised the matching process, as simply pairing P points to the closest Q points facilitated time difference variation between paired data points. If requisite Q data points were absent then the respective P point was removed. Where this reduced dataset sample size to fewer than 30, which occurred in 29 cases, the dataset was excluded. This threshold was implemented in an effort to maintain representation of real-life data availability and a high number of datasets for analysis, whilst not using datasets with such low levels of data that they were unsuitable for analysis.

Determining point apportionment according to load apportionment models

Point source apportionment for each watercourse was calculated using algorithms extracted from Bowes et al. (2008) and Greene et al. (2011), equations 1 and 2 respectively. For BM, the B variable was constrained to 0 (following Bowes et al., 2010 and Charlton et al., 2018). Bootstrapping (N=2000) using high frequency Q data was then undertaken to calculate output SE using the phoslam package in R (O'Riordain and Crockford, 2014). Due to error messages from model fit a further sixteen datasets were incompatible and were discounted.

(Equation 1)

$$P = A.Q^{B-1} + C.Q^{D-1}$$

where P is phosphorus concentration, Q is flow, A, B (=0), C and D (≥1) are time-invariable coefficients.

(Equation 2)

$$P = aQ^{-1} + bQ + cQ^{2}$$

where P is phosphorus concentration, Q is flow and a, b and c are time-invariable coefficients.

Acquisition and calculation of catchment metrics

For remaining datasets shapefiles detailing catchment boundaries and size for each Q sampling point were sourced from the NRFA (NRFAb, 2019) along with statistics on land-use, baseflow index and holistic steepness of catchments available from NRFAb (2019; detailed in Table 1). Finally, a sinuosity index score for each watercourse was calculated using equation 3, as employed by Yu (2017).

(Equation 3)

$$S = \frac{L}{Lv}$$

where *S* is sinuosity, *L* is the length of the river following all curves and *Lv* is the length between these points following a direct path.

To obtain metrics for equation 3, a UK river shapefile (OS, 2019) was overlaid with each catchment boundary in ArcMap 10.5.1. Using the *clip* function the river layer was reduced so only watercourses within individual catchments were present, with the resultant attribute table containing the length of these watercourse polygons which could be appropriately selected and totalled, whilst the *measure* function was utilised to provide *Lv* measurements. In total, nine catchment metrics (Table 1) were provided as explanatory variables to observed SE variation.

Statistical analysis methodology

All data was combined into one dataset (Appendix 1), and analysed in R using a range of packages and functions; denoted in text by 'Package':*function*. If not specified, functions were present in the base package.

Summary statistics, normality testing, data transformation and model SE correlation

The mean, standard deviation, median and quartile statistics were calculated for each variable. To test normality, histograms were plotted and the Anderson-Darling p statistic calculated, using 'nortest':*ad.test* (Ligges, 2015). Those variables evidencing non-normal distribution were logarithmically transformed to coerce data into normal, or closer to normal, distribution. Where this resulted in negative numbers each dataset value was increased by one (Fletcher et al., 2005). Spearman's correlation analysis was also undertaken between BM and GM to ascertain if any association was present.

Random Forest Analysis

To identify relative importance of individual explanatory variables on SE obtained from BM and GM, random forest analysis was undertaken, using ‘randomForest’:*randomForest* (Liaw, 2018). The analysis, based on the algorithm by Breiman (2001), created a series of decision trees (questions with multiple answers regarding the explanatory variables) by randomly sub-sampling the dataset (Ekstrøm, 2016). Thus, machine-learning was employed to identify the relative importance of explanatory variables in correctly predicting the response variable category (Cutler et al., 2007), measured by Mean Decrease of Accuracy (MDA) and the Gini Index. Specifically, the MDA value provided a measure of loss in predictive performance when a variable was removed or permutated (San Diego University, 2017). The Gini Index measures node purity after each split (question) in the decision tree. Node purity refers to homogeneity of data categories contained within a child node after a split in the decision tree. The Gini coefficient for all nodes were summed and normalised for each variable individually to provide a ranking (San Diego University, 2017). Out of Bag Error (OOB) statistics were also calculated, which detail overall prediction error rate for the model built by the random forest, whilst error rates for the prediction of individual response variable categories were also provided in the output.

For the random forest analysis, the continuous response variable was converted to categorical data, with a similar number of data points within categories to minimise bias in correctly predicting an individual category. Thus, SE was split into three categories (low, medium, high) with 30, 30 and 31 points respectively; reflecting the interest in change of SE across datasets in general as opposed to SE beyond a given threshold.

Three forests were grown for each dataset (BM SE or GM SE) to enable comparison of model outputs for the individual datasets and so ensure outputs were consistent when different start points of random data selection were specified via the *set.seed* function. The number of trees grown within each forest was 500 to ensure each dataset row (individual catchments) would be predicted more than once but not oversampled. Numerical results of explanatory variable importance were scaled to the variable with the largest score.

Correlation and regression analysis of variables identified as most important in Random Forest testing

Where Random Forest analysis evidenced individual explanatory variables were important in predicting response variables, further testing was undertaken to quantify the strength of potential univariate relationships. Correlative tests were first employed to determine if relationships were present ($p<0.05$) with linear regression undertaken to generate R^2 statistics where true. Post-hoc tests (Anderson-Darling p statistic and Residuals vs Fitted, Normal Q-Q and Scale-Location plots) were performed to ensure model errors had a normal distribution, which evinces statistical assumptions of linear regression are being met (Li et al., 2012).

Where post-hoc testing suggested assumptions were violated, plots were visually examined to ascertain if individual data points had disproportionate leverage, as linear regression is sensitive to outliers which distort true data patterns (Fox, 2015). Where found, the linear regression and post-hoc tests were re-run with the data points removed to evaluate their impact on model assumption violation (following Osbourne et al., 2004) and, if errors were then normally distributed, to recalculate R^2 statistics.

Moreover, to understand if a relationship was present when considering the full dataset, a quantile regression, which negates the need for normal error distribution, was undertaken using ‘quantreg’:*rq* (Koenker, 2019). Model fit was compared, via AIC(k=2), to a null model created using the interaction term ‘~1’; if the null model had a better fit it evidenced the perceived relationship could be reproduced in a simple model which did not incorporate the explanatory variable of interest (Gotelli, 2001). Quantile regression could not indicate the strength of relationship, as pseudo R^2 cannot be interpreted as the proportion of response variability explained by the explanatory variable (Fox, 2015).

Assessing Load Apportionment Model(s)' fit

For each catchment, AIC values were calculated to quantify BM and GM model fit and so provide information on which model provided the better fit. As the base package AIC function was incompatible with phoslam, calculation of the value was undertaken in Microsoft Excel using equation 4 as set out by Zhou et al. (2013):

(Equation 4)

$$AIC = \frac{2k}{n} + \log(RSS/n)$$

where *k* represents number of model parameters (Bowes=4 and Greene=3), *n* represents number of data points and *RSS* sum of squared residuals.

To calculate the RSS, modelled P values from observed Q values were produced in Excel using the BM and GM algorithms. The required parameter values for BM and GM were sourced using *phoslam*, entered into the spreadsheet and linked via cell coding to the algorithm. Furthermore, the tendency of modelled values to be greater or smaller than observed values, indicating bias, was calculated using 'hydroGOF':*pbias* (Zambrano-Bigiarini, 2017). The function returns a percentage value representing the datasets average difference between modelled and actual values; negative values indicating underestimation and positive values indicating overestimation.

Results

Summary statistics, normality testing, data transformation and model SE correlation

Summary statistics of all variables are contained in Table 2. Inspection of histograms and the results of Anderson-Darling tests evinced that all variables were considered to have non-normal data distribution and were therefore logarithmically transformed. All data except those for Catchment Size, Slope and Grassland were increased by one prior to transformation to remove negative datapoints post transformation. Spearman's correlation analysis revealed a 'strong' (Pallant, 2016) positive correlation between BM SE and GM SE ($r=0.83$, $n=91$, $p<.001$).

Random Forest Analysis

Prediction error rates

The mean OOB for the three BM forests created was 52.75% (SE: 0.64), whilst the same statistic was 61.90% (SE: 2.03) for GM forests. Mean prediction error for individual response variable categories within the BM forests was 37.63% for 'high', 62.22% for 'medium' and 58.88% for 'low' (SE: 0.01 for all). Regarding GM forests, mean prediction error was 54.83%, 73.33% and 57.78% when predicting 'high', 'medium' and 'low' categories (SE: 0.01, 0.01 and 0.02 respectively).

Variable importance

Scaled importance of variables, using both the MDA and Gini Index, are presented within Figure 2. Relative importance, and order, of explanatory variables in effecting response variables was notably different between BM and GM forest outputs. Furthermore, divergence in variable order was present between MDA and Gini ratings *within* models, BM or GM forest respectively; though this pattern only applied to the order *after* the variable considered of the greatest importance, which remained constant between the two measures *within* models, though not *between* models.

Correlation and regression analysis of variables identified as most important in Random Forest

Spearman's correlation testing between GM SE and Catchment Size and GM SE and Slope returned non-significant results (p=.16 and p=.23 respectively). The same test for BM SE to Urban did evidence a relationship (p<.001), so a linear regression was undertaken (t=4.72, d.f.=89, p<.001, R^2=0.20).

Post-hoc testing of the linear regression revealed model errors were not normally distributed (p<.001), with two outlying data point residuals (catchments 15 and 89) potentially disproportionally impacting the linear regression result. These points were removed and the test re-run, with a notable benefit to error normality (p=.29), though less of a change noted in model output (t=4.566, d.f.=87, p<.001 and R^2=0.19); Figure 3.

As per the methodology, a quantile regression was then undertaken on the full dataset and compared for fit, using AIC(k=2), with a null model. The quantile regression had the better fit, evidencing that the perceived relationship between BM SE and Urban was not able to be reproduced when no explanatory variable was included.

Assessment of Load Apportionment Model(s) fit

AIC values evidenced the BM algorithm provided a better modelled fit to observed data in 84 of the 91 catchments. For all catchments the GM algorithm provided a higher estimate of point load apportionment compared to BM, ranging from 1.02 to 14.66 times greater (mean: 2.15, SD: 2.18). Percentage bias statistics evidenced model bias varied hugely (-99% to >200% and -100% to >1000% for BM and GM respectively). Overall BM had a more consistent, lower, bias (mean: 3.3%, SD: 32%) than GM (mean: >500% SD: >1000%), with the BM modelling P values to ± 10% of actual P values in 85.7% of datasets, opposed to GMs 17.6%.

Discussion

Relationship between catchment characteristics and the GM

Relative homogeneity of the aggregated GM random forests output, especially in relation to the Gini Index (Figure 2), evidences catchment characteristics are not individually influential in determining GM SE, as re-iterated by correlation analysis, which could suggest variables may be interacting together. It may also infer that a parameter not included within the study is having a disproportionate impact. The high OOB strengthens this theory as it demonstrates the random forest model is having low success in predicting SE class from included variables, which would be illogical if the variables are interacting and responsible for the majority of SE variation. In reality, a combination of theories is likely to be more accurate in that variables are interacting to cause variation, though further parameters are necessary to fully account for SE alteration. If the range in SEs has been produced through chance with no real catchment characteristic influence then this could infer that the model could be applied in any catchment. However, as the model was relatively low for accuracy of modelled outputs there are remaining challenges for the use of GM in catchment management.

Relationship between catchment characteristics and the BM

Conversely, the BM random forest and proceeding regression analysis identified one variable, Urban, as being responsible for c.19% of SE variation. Although this figure is derived from post data point removal, a contentious although often necessary procedure (Osborne and Overbay, 2004), confidence in its validity is provided through the quantile regression results and how exclusion of data points caused only a minor alteration in the R^2 value.

The LAM relies upon the relationship between Q and P altering in response to the predominant contribution source and should anything facilitate a deviation from the assumptions of this relationship then model output variability will be observed, as is the case with BM SE and Urban. Urbanisation fundamentally alters hydrological mechanisms and pathways, which consequently impacts the level and timing of runoff (Hung, 2018). This is predominantly manifested by a reduction in pervious surfaces and an increase in flow velocity (Trudeau and Richardson, 2016; Pumo et al., 2017) caused by diversion of flow. Changes in surface permeability and increased water velocity can all cause a 'flashy' hydrograph of reduced flow periods and increased peak discharges (Neave and Rayburg, 2016). This characteristic, combined with low frequency sampling, is a likely cause of model variability and loss of output robustness as the dataset will not represent the full range of storm events within the catchments and so cannot accurately model diffuse P contributions (Bowes et al., 2008). Additionally, urbanisation also impacts processes such as evapotranspiration (Locatelli et al., 2017) and the geomorphological dimensions of a watercourse, due to increased water velocity (Jacobson, 2011).

The impact of 'flashy' hydrographs and low sampling frequency on nutrient load estimation uncertainty has long been proposed (Johnes, 2007), with it still being highlighted as a barrier to robust models and reliable outputs in contemporary studies (Hollaway et al., 2018; Jung et al., 2020).This reduction in high Q data will be a further likely source of model uncertainty as true levels of diffuse contributions are masked (Johnes, 2007; Bowes et al., 2008).

Stormwater infrastructure can also cause higher levels of in-stream sedimentation through either transfer of stored sediment, or the increase of bankside erosion from elevated flow rates if water diversion is the utilised management method (Ruhlman et al., 2016). Within a watercourse, sedimentation further complicates Q-P patterns as adsorbed sediment may be released during higher flows. This behaviour will mean that true point source apportionment levels are masked as the rise in Q and P would be attributed to diffuse source by the LAM assumptions (Jarvie et al., 2012), whilst increasing levels of P retention reduce the BM applicability. Furthermore, climate, chemical state and river geomorphological characteristics will impact the variability of retention rates and observed patterns (McDowell et al., 2017; Omari et al., 2019; Xiao et al., 2019). This may further conspire to cause model output variability as the Q-P relationships that the LAM rely upon are being complicated.

Despite these issues, it remains that the defined relationship between BM SE and urban does not account for the majority of SE variation. Given there are complex interlinked processes that govern hydrological processes and P transfer (Holloway et al., 2018) it is feasible, as hypothesised with the GM, that the variables are interacting to cause the variation. It is also feasible that variables included in this study do not fully account for observed BM variation and other factors should be considered to estimate variation in BM and GM analyses. This sentiment becomes evident when considering catchment 89, which provided the highest SE for the BM and GM, although the quantified catchment characteristics were not obviously divergent or extreme from other datasets, so indicating that further factors are required to account for the SE variation.

Applicability of LAMs

Although the GM did not, holistically, provide an accurate representation of observed data points, the BM yielded results which demonstrate the algorithm generally performs well on datasets of the type analysed within this study. However, a challenge remains that these datasets are unlikely, given sampling frequency, to capture the full range of Q-P variation that occur within watercourses as recently shown by Jung et al. (2020). Only by using high frequency Q-P data can true patterns be identified (Bieroza and Heathwaite, 2015; Williams et al., 2015; Elwan et al., 2018) and thereby increase the accuracy of BM P apportionment. Moreover, P models are known to have a reduced ability to model P at high Q (Cassidy and Jordan, 2011; Chen et al., 2013; Crockford et al., 2017). When these issues are coupled with original model designers highlighting the need for high Q data to increase model robustness (Bowes et al., 2008) then interpreting BM outputs calculated from low temporal resolution datasets as representative of true trends appears unwise. Such issues will also conspire to undermine the model's usefulness for future application on low frequency datasets, given that more frequent storm events are forecast due to climate change (GOV.UK, 2018). Not only does capturing the full range of storm events enable accurate outputs from these models, but the change in storm frequency and vigour has the capability to alter pathways and intensity of diffuse P transfer (Forber et al., 2018), which could further facilitate deviation from the Q-P relationships on which the LAM rely upon.

It must also be noted that though the BM has a high success rate at predicting observed data points, not utilising methods other than LAM to explain these data points could result in misinterpretation. For example, those catchments which consist predominantly of dynamic land-use, such as arable, or over a longer time period forestry, could instigate biased outputs if Q-P monitoring is over too long a period or too short a period. In the example of forestry, if monitoring was centred around a felling period then diffuse contributions would be weighted highly. However, if the monitoring period was either between felling or over many years, then this diffuse loss could be missed or diluted. Only by investigating data trends and comparing these to catchment characteristics can effective, accurate mitigation measures be designed.

Future research

Load apportionment modelling

Given concerns about the effect of low frequency data use on output accuracy it would be beneficial to undertake a study, spanning a wider range of datasets as possible, looking at how BM and GM point apportionment and SE are impacted by the inclusion of high frequency data. This would also then facilitate re-analysis of the effect of catchment characteristics on SE, which would test the conclusions of this study. Moreover, it would be valuable to expand the catchment characteristics incorporated within the random forest analysis as the results indicate SE variation is not fully explained by those included. This may include the prevalence of known point sources which may not be adequately represented by degree of urbanisation. Quantifying specific soil types and their distribution would also be an obvious choice given soil type is known to be influential in P dynamics (Bergström et al., 2015). Although base flow index is heavily influenced by soil type and so may be considered a proxy for this, it does not provide the in-depth understanding of soil type and distribution that may be contributing to the SE variation not accounted for within this study. Regarding interactions between variables being potentially responsible for SE variation, especially in the case of GM, further statistical analysis of the dataset (Appendix 1) would enable interactions between variables to be explicitly identified and quantified. This may be important when considering the role that catchment area plays in the magnitude of export of P in a river.

It would also be highly useful to quantify the impact on the LAMs output and SE of using Q-P data which was not temporally and spatially matched at the point of collection. While every effort was made to ameliorate this concern, it represents a methodological deviation from that set out by Bowes et al. (2008) and Greene et al. (2011). Moreover, if it was found to be a significant issue then it could further question the applicability of LAMs as a tool for quickly analysing a range of watercourses, as the issue itself was borne from current data availability.

Finally, it would be advantageous to comprehend if the use of LAMs models on low frequency datasets could be incorporated into a wider framework for accurately assessing P apportionment. This study has shown that the BM is capable of providing a relatively accurate model of widely available low frequency datasets, whilst the models themselves facilitate reduced time and labour requirements when assessing P apportionment. If accuracy is not greatly compromised by the use of high frequency data, though this seems probable, the BM could be utilised in catchments where the outputs (SE) are found to be most consistent and avoided where model error is known to be exacerbated, such as heavily urbanised catchments. Therefore, where limited resources are available, efforts to comprehend P apportionment using other methods with increased labour requirements could be targeted towards those catchments where the BM is considered less accurate and more variable.

Using catchment characteristics to evaluate models

Across the 91 catchments investigated, catchment characteristics displayed diversity in their respective measurements, therefore providing a good basis for this study's investigation into their role in LAM variation. Furthermore, that BM and GM evidence linearity in their SE outputs suggests that environmental variables, not accounted for is this study, are influencing model variation which a simple numerical model is compromised to reflect. Using catchment characteristics to evaluate the causation of standard error in models has been largely inconclusive in this study except for the suggestion that BM is influenced by percentage urban cover. Using catchment characteristics to evaluate model error remains, however, a novel method of identifying the influences on standard error as simple numerical models continue to be used in catchment management (e.g. Ascott et al., 2018). Previous use of catchment descriptors with model outputs have allowed predictions in other scenarios with fewer data available, such as Deckers et al. (2010) or determined the impact of changing a catchment characteristic such as catchment size in Andrianaki et al. (2019). Catchment characteristics have been cited as possible explanatory influences on the variation in hydrological simulation across 979 catchments in the US and UK with geology and baseflow contributions particularly identified (Seibert et al., 2018), thus confirming that investigating the causation of error may make the applicability of models more robust in the future.

Conclusion

This study has been the first to calculate certainty statistics when applying the BM and GM to a wide range of river catchment datasets. In doing so, it has been evidenced that the BM output variability increases as levels of urban cover rise, whilst the GM SE is less influenced by individual variables. It is hypothesised that further variables beyond those included within this study are impacting the SE of both models, whilst interactions between studied variables may also be present.

Further investigation into these hypotheses is required, though more pressing is the need to ascertain if the outputs, even where there is low SE, represent true patterns of the Q-P relationship. Such research using high temporal frequency data could provide justification of the continued use of each LAM to accurately model P changes as a function of Q on low frequency datasets. Moreover, this may yield differing results regarding the importance of catchment characteristics on model variation than has been shown within this study.

Finally, this study has demonstrated a method for using catchment descriptors to identify the drivers for SE variability across modelled river catchments. By identifying the descriptors that models are highly sensitive to, more appropriate use of simple numerical models, such as LAMs, may be developed.

References

Andrianaki, M., Shrestha, J., Kobierska, F., Nikolaidis, N. P., & Bernasconi, S. M. (2019). Assessment of SWAT spatial and temporal transferability for a high-altitude glacierized catchment. *Hydrol Earth Syst Sc*, 23(8), 3219-3232. doi: 10.5194/hess-23-3219-2019

Antoniadis, V., Koliniati, R., Efstratiou, E., Golia, E. and Petropoulos, S. 2016. Effect of soils with varying degree of weathering and pH values on phosphorus sorption. *CATENA,* 139, 214-219. doi: 10.1016/j.catena.2016.01.008

Bergström, L., Kirchmann, H., Djodjic, F., Kyllmar, K., Ulen, B., Liu, J., Andersson, H., Aronsson, H., Börjesson, G., Kynkäänniemi, P., Svanbäck, A. and Villa, A. 2015. Turnover and losses of phosphorus in Swedish agricultural soils: long-term changes, leaching trends, and mitigation measures. *J Env Qual,* 44(2), 512-523. doi: 10.2134/jeq2014.04.0165

Bieroza, M.Z. and Heathwaite, A.L. 2015. Seasonal variation in phosphorus concentration-discharge hysteresis inferred from high frequency *in situ* monitoring. *J Hydrol,* 524, 333-347. doi: 10.1016/j.jhydrol.2015.02.036

Bong, C.H.J., Lau, T.L. and Ghani, A.A. 2016. Potential of tipping flush gate for sedimentation management in open stormwater sewer. *Urban Water J,* 13(5), 486-498. doi: 10.1080/1573062X.2014.994002

Bowes. M.J., Smith, J.T., Jarvie, H.P and Neal, C. 2008. Modelling of phosphorus inputs to rivers and diffuse point sources. *Sci Total Environ,* 395 (2-3), 125-138. doi: 10.1016/j.scitotenv.2008.01.054

Bowes, M.J., Smith, J.T., Jarvie, H.P., Neal, C. and Barden, R. 2009. Changes in point and diffuse source phosphorus inputs to the River Frome (Dorest, UK) from 1966 to 2006). *Sci Total Environ,* 407, 1954-1966. doi: 10.1016/j.scitotenv.2008.11.026

Bowes, M.J., Neal, C., Jarvie, H.P., Smith, J.T. and Davies, H.N. 2010. Predicting phosphorus concentrations in British rivers resulting from the introduction of improved phosphorus removal from sewage effluent. *Sci Total Env,* 408(19), 4239-4250. doi: 10.1016/j.scitotenv.2010.05.016

Bowes, M.J., Jarvie, H.P., Naden, P.S., Old, G.H., Scarlett, P.M., Roberts, C., Armstrong, L.K., Harman, S.A., Wickham, H.D. and Collins, A.L. 2014. Identifying priorities for nutrient mitigation using river concentration-flow relationships: The Thames basin, UK. *J Hydrol,* 517, 01-12. doi: 10.1016/j.jhydrol.2014.03.063

Breiman, L. 2001. Random forests. *Machine Learning,* 45, 05-32. doi: 10.1023/A:1010933404324

Bridge, J.S. and Demicco, R.V. 2008. *Earth surface processes, landforms and sediment deposits.* New York: Cambridge University Press.

Cassidy, R. and Jordan, P. 2011. Limitations of instantaneous water quality sampling in surface-water catchments: Comparison with near-continuous phosphorus time-series data. *J Hydrol* 405(1-2), 182-193. doi: 10.1016/j.jhydrol.2011.05.020

Charlton, M. B., Bowes, M. J., Hutchins, M. G., Orr, H. G., Soley, R., & Davison, P. (2018). Mapping eutrophication risk from climate change: Future phosphorus concentrations in English rivers. *Sci Total Environ, 613*, 1510-1526. 10.1016/j.scitotenv.2017.07.218

Chen, D., Dahlgren, R.A. and Lu, J. 2013. A modified load apportionment model for identifying point and diffuse source nutrient inputs to rivers from stream monitoring data. *J Hydrol.* 501, 25-34. doi: 10.1016/j.jhydrol.2013.07.034

Crochmore, L., Rafael, P., Luis P., Abdulghani, H., Ilias, P., Kristina, I., Jafet, A. and Berit, A. 2018. *Understanding and evaluating catchment memory from a global hydrological model: paper presented at the 20th EGU general assembly conference 04-13 April 2018 Vienna, Austria.* Germany: European Geosciences Union.

Crockford, L., O'Riordain, O., Taylor, D., Melland, A.R., Shortle, G. and Jordan P. 2017. The application of high temporal resolution data in river catchment modelling and management strategies. *Environ Mon Assess,* 189(9), doi: 10.1007/s10661-017-6174-1

Cutler, D.R., Edwards, T.C., Beard, K.H, Cutler, A., Hess, K.T., Gibson, J. and Lawler, J.J. 2007. Random forests for classification in ecology. *Ecology,* 88(11), 2783-2792. doi: 10.1890/07-0539.1

Daryanto, S., Wang, L. and Jacinthe, P.A. 2017. Meta-analysis of phosphorus loss from no-till soils. *J Env Quality,* 46(5), 1028-1037. doi:10.2134/jeq2017.03.0121

Deckers, D., Booij, M. J., Rientjes, T. M., & Krol, M. S. (2010). Catchment Variability and Parameter Estimation in Multi-Objective Regionalisation of a Rainfall-Runoff Model. *Water Res Manage*, 24(14), 3961-3985. doi: 10.1007/s11269-010-9642-8

EA (Environment Agency). not dated. *Download open water quality archive datasets.* environment.data.gov.uk/water-quality/view/download

Efron, B. (1979). Bootstrap Methods: Another look at the Jacknife. *Ann Statis,* 1, 01-26. doi: 10.1007/978-1-4612-4380-9_41

Ekstrøm, C.T. 2016. *The R primer.* Boca Raton: CRC Press.

Elwan, A., Singh, R., Patterson, M., Roygard, J., Horne, D., Clothier, B. and Jones, G. 2018. Influence of sampling frequency and load calculation methods on quantification of annual river nutrient and suspended solids loads. *Environ Mon Assess,* 190(2). doi: 10.1007/s10661-017-6444-y

ESRI (Environmental Systems Research Institute). 2019. *ArcMap.* desktop.arcgis.com/en/arcmap/

Fletcher, D., MacKenzie, D., Villouta, E., 2005. Modelling skewed data with many zeros: A simple approach combining ordinary and logistic regression. *Environ Ecol Stat*, 12, 45–54. doi: 10.1007/s10651-005-6817-1

Forber, K.J., Withers, P.J.A., Ockenden, M.C. and Haygarth, P.M. 2018. The phosphorus transfer continuum: A framework for exploring effects of climate change. *Ag Environ Let,* 3. doi: 10.2134/ael2018.06.0036

Fox, J. (2015). *Applied regression analysis and generalized linear models* (Third ed.). Thousand Oaks: SAGE Publications, Inc.

Gotelli, N.J. 2001. Research frontiers in null model analysis. *Global Ecol Biogeogr,* 10, 337-343. 10.1046/j.1466-822X.2001.00249.x

GOV.UK. 2018. *Climate change means more frequent flooding, warns Environment Agency.* www.gov.uk/government/news/climate-change-means-more-frequent-flooding-warns-environment-agency

Greene, S., Taylor, D., McElarney, Y.R. and Jordan, P. 2011. An evaluation of catchment-scale phosphorus mitigation using load apportionment modelling. *Sci Total Environ,* 409 (11), 2211-2221. doi: 10.1016/j.scitotenv.2011.02.016

He, S., Wang, D., Chang, S., Fang, Y. and Lan, H. 2018. Effects of morphology of sediment-transporting channels on the erosion and deposition of debris flows. *Environ Earth Sci,* 77(14). doi: 10.1007/s12665-018-7721-y

Holloway, M.J., Beven, K.J., Benskin, C.McW.H., Cllins, A.L., Evans, R., Falloon, P.D., Forber, K.J., Hiscock, K.M., Kahana, R., Macleod, C.J.A., Ockenden, M.C., Villamizar, M.L., Wearing, C., Withers, P.J.A., Zhou, J.G., Barber, N.J. and Haygarth, P.M. 2018. The challenges of modelling phosphorus in a headwater catchment: Applying a 'limits of acceptability' uncertainty framework to a water quality model. *J Hydrol,* 558, 607-624. doi: 10.1016/j.jhydrol.2018.01.063

Hung, C.J. 2018. *Catchment hydrology in the Anthropocene: Impacts of land-use and climate change on stormwater runoff.* South Carolina: University of South Carolina.

Jacobson, C.R. 2011. Identification and quantification of the hydrological impacts of imperviousness in urban catchments: A review. *J Environ Manage,* 6, 1438-1448. doi: 10.1016/j.jenvman.2011.01.018

Jarvie, H.P., Sharpley, A.N., Scott, J.T., Haggard, B.E., Bowes, M.J., Massey, L.B. 2012. Within-river phosphorus retention: accounting for a missing piece in the watershed phosphorus puzzle. *Environ Sci Technol,* 46(24), 13284-13292. doi: 10.1021/es303562y

Johnes, P.J. 2007. Uncertainties in annual riverine phosphorus load estimation: impact of load estimation methodology, sampling frequency, baseflow index and catchment population density. *J Hydrol,* 332, 241-258. doi: 10.1016/j.jhydrol.2006.07.006

Jung, H., Senf, C., Jordan, P., and Krueger, T. 2020. Benchmarking inference methods for water quality monitoring and status classification. *Env Monit Assess,* 192, 261. doi: 10.1007/s10061-020-8223-4

Koenker, R. 2019. *Quantreg: Quantile Regression. R package version 5.40.* CRAN.R-project.org/package=quantreg

Leaf, S. 2018. Taking the P out of pollution: an English perspective on phosphorus stewardship and the Water Framework Directive. *Water Environ J,* 32, 04-08. doi: 10.1111/wej.12268

Li, X., Wong, W., Lamoureux, E.L. and Wong, T.Y. 2012. Are linear regression techniques appropriate for analysis when the dependent (outcome) variable is not normally distributed? *In Opth Vis. Sci,* 53, 3082-3083. doi: 10.1167/iovs.12-9967

Li, Z., Tang, H., Xiao, Y., Zhao, H., Li, Q. and Ji, F. 2016. Factors influencing phosphorus adsorption onto sediment in a dynamic environment. *J Hydro-Environ Res,* 10, 01-11. doi: 10.1016/j.jher.2015.06.002

Liaw, A. 2018. *randomForest v4.6-14.* cran.r-project.org/web/packages/randomForest/index.html

Ligges, U. 2015. *nortest function.* cran.r-project.org/web/packages/nortest/index.html

Locatelli, L., Mark, O., Mikkelsen, P.S., Arnbjerg,-Nielsen, K., Deletic, A., Roldin, M. and Binning, P.J. 2017. Hydrologic impact of urbanization with extensive stormwater infiltration. *J Hydrol,* 544, 524-537. doi: 10.1016/j.jhydrol.2016.11.030

Lou, H., Zhao, C., yang, S., Shi, L., Wang, L., Ren, X. and Bai, J. 2018. Quantitative evaluation of legacy phosphorus and its spatial distribution. *J Environ Manage,* 211, 296-305. doi: 10.1016/j.jenvman.2018.01.062

MacDonald, G.K., Bennet, E.M. and Taranu, Z.E. 2012. The influence of time, soil characteristics, and land-use history on soil phosphorus legacies: a global meta-analysis. *Global Change Biol,* 18(6), 1904-1917. doi: 10.1111/j.1365-2486.2012.02653.x

Maxwell, R.M., Condon, I.E., Kollet, S.J., Maher, K., Haggerty, R. and Forrester, M.M. 2016. The imprint of climate and geology on the residence times of groundwater. *Geophys Res Lett,* 43, 701-708. doi: 10.1002/2015GL066916

McDowell, R.W., Elkin, K.R and Kleinman, P.J.A. 2017. Temperature and Nitrogen effects on Phosphorus uptake by agricultural stream- bed sediments. *J Environ Qual,* 46, 295-301. doi: 10.2134/jeq2016.09.0352

Neave, M. and Rayburg, S. 2016. Designing urban rivers to maximise their geomorphic and ecologic diversity. *Geotec, Const Mat & Env*, 11(25), 2468-2473. doi: http://www.geomatejournal.com/sites/default/files/articles/2468-2473-5164-Neave-Sept-2016-c1.pdf

NRFAa (National River Flow Archive), 2019. *Derived flow statistics.* https://nrfa.ceh.ac.uk/derived-flow-statistics

NRFAb (National River Flow Archive), 2019. *FEH catchment statistics*. https://nrfa.ceh.ac.uk/feh-catchment-descriptors

Omari, H., Dehbi, A., Lammini, A. and Abdallaoui, A. 2019. Study of phosphorus adsorption on the sediments. *J Chem* doi: 10.1155/2019/2760204

O'Riordain, S. and Crockford, L. 2014. *Phoslam package in R.* https://github.com/seanpor/phoslam

OS (Ordnance Survey). 2019. *OS open rivers shapefile download.* https://www.ordnancesurvey.co.uk/business-and-government/products/os-open-rivers.html

Osbourne, J.W. and Overbay, A. 2004. The power of outliers (and why researchers should *always* check for them). *Prac Assess Res Eval,* (6), 01-12. scholarworks.umass.edu/pare/vol9/iss1/6/

Pallant, J. 2016. *SPSS survival manual.* 6th ed. Berkshire: Open University Press.

Pumo, D., Arnone, E., Francipane, A., Caracciolo, D. and Noto, L.V. 2017. Potential implication of climate change and urbanization on watershed hydrology. *J Hydrol,* 554, 80-99. doi: 10.1016/j.jhydrol.2017.09.002

R Core Team, (2019). *R, a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Rogger, M., Agnoletti, M., Alaoui, A., Bathurst, J.C., Bodner, G., Borga, M., Chaplot, V., gallart, F., Glatzel, G., Hall, J., Holden, J., Holko, L., Horn, R., Kiss, A., Kohnova, S., Leitinger, G., Lennartz, B., parajka, J., Perdigao, R., Peth, S., Plavcova, L., Quinton, J.N., Robinson, M., Salinas, J.L., Santoro, A., Szolgay, J., Tron, S., Akker, J.J.H, Viglione, A. and Bloschl, G. 2017. Land use change impacts on floods at the catchment scale: Challenges and opportunities for future research. *Water Resour Res,* 53, 5209-5219. doi: 10.1002/2017WR020723

Ruhlman, M., Vandelay, A. and Roper, C. 2016. *Cooperative planning for source water protection: Targeting sediment in the upper Saluda river watershed. Presented at the South Carolina Water Resources Conference, 17-18 October 2016, South Carolina.*

San Diego University. 2017. *Random Forests.* https://dinsdalelab.sdsu.edu/metag.stats/code/randomforest.html

Seibert, J., Vis, M. J. P., Lewis, E., & van Meerveld, H. J. (2018). Upper and lower benchmarks in hydrological modelling. *Hydrol Process*, 32(8), 1120-1125. doi: 10.1002/hyp.11476

Sharpley, A. 2016. Managing agricultural phosphorus to minimize water quality impacts. *Sci Agri,* 73, 01-08. doi: 10.1590/0103-9016-2015-0107

Trudeau, M.P. and Richardson, M. 2016. Empirical assessment of effects of urbanization on event flow hydrology in watersheds of Canada's Great lakes-St Lawrence basin. *J. Hydrol.* 541, 1456-1474. doi: 10.1016/j.jhydrol.2016.08.051

Williams, M.R., King, K.W., Macrae, M.L., Ford, W., Esbroeck, C., Brunke, R.I., English, M.C. and Schiff, S.L. 2015. Uncertainty in nutrient loads from tile-drained landscapes: Effect of sampling frequency, calculation algorithm, and compositing strategy. *J Hydrol.* 530, 306-316. doi: 10.1016/j.jhydrol.2015.09.060

Xiao, C., Chen, J., Chen, D. and Chen, R. 2019. Effects of river sinuosity on the self-purification capacity of the Shiwuli River, China. *Water Supply,* 19(4), 1152-1159. doi: 10.2166/ws.2018.166

Yaeger, M., Coopersmith, E., Ye, S., Cheng, L., Viglione, A., & Sivapalan, M. (2012). Exploring the physical controls of regional patterns of flow duration curves - Part 4: A synthesis of empirical analysis, process modeling and catchment classification. *Hydrol Earth Syst Sc, 16*(11), 4483-4498. doi: 10.5194/hess-16-4483-2012

Yu, P.W.C. 2017. *Submarine landslides, canyons, and morphological evolution of the East Australian Continental Margin: A thesis submitted for the degree of Doctor of Philosophy.* Sydney: The University of Sydney.

Zambrano-Bigiarini, M. 2017. *HydroGoF function.* cran.r-project.org/web/packages/hydroGOF/index.html

Zhou, J., Zhao, X. and Sun, L. 2013. A new inference approach for joint models of longitudinal data with informative observation and censoring times. *Stat Sin,* 23, 571-593. https://www.jstor.org/stable/24310353

Table 1 Study variables and description

Variable Name	Description
BM P Apportionment	The mean percentage of a river's phosphorus load apportioned to point sources according to the bootstrapped BM (Bowes et al., 2008); equation 1.
BM SE	Standard error of the bootstrapped BM P Apportionment
GM P Apportionment	The mean percentage of a rivers phosphorus load apportioned to point sources according to the bootstrapped GM (Greene et al., 2011); equation 2.
GM SE	Standard error of the bootstrapped GM P Apportionment.
Catchment Size	The catchment size in km^2 of the Q data collection point; as defined by NRFAb (2019).
Slope	The holistic steepness of a catchment varying from <25 in the flattest areas of the country to >300 in mountainous regions (NRFAb, 2019).
Base Flow	Baseflow index score derived from the Hydrology of Soil Types classification system which provides calculated runoff responses for individual soil types. These scores are aggregated across the catchment (NRFAb, 2019).
Sinuosity	Sinuosity index score, calculated as detailed in Section 3.5.
Woodland	Percentage of catchment classified as 'woodland' by NRFAb (2019).
Arable	Percentage of catchment classified as 'arable or horticultural' by NRFAb (2019).
Grassland	Percentage of catchment classified as 'grassland' by NRFAb (2019).
Urban	Percentage of catchment classified as 'urban' by NRFAb (2019).
Heath	Percentage of catchment classified as 'mountain, heath or bog' by NRFAb (2019).

Table 2 Summary statistics of variables. *Note: BM and GM P Apportionment were not included in statistical analysis given this study's principal focus (SE), although they are included here to detail variation in P point apportionment across datasets*

	Min.	1st Qu.	Median	Mean	SD	3rd Qu.	Max.	Anderson-Darling p statistic of log transformation
BM P Apportionment	1.0900	11.1500	22.4000	25.7205	18.4040	38.8000	69.3000	n/a
GM P Apportionment	4.6200	20.5000	36.1000	37.3716	19.5268	53.6500	79.8000	n/a
BM SE	0.0295	0.4560	0.6710	0.7478	0.4701	0.9810	3.0900	.002
GM SE	0.0087	0.4405	0.5460	0.5927	0.3325	0.7090	2.2200	<.001
Catchment Size	9.000	63.250	128.000	336.411	543.231	269.700	3315.000	.055
Slope	11.5000	29.8000	55.9000	65.8121	48.6541	92.4000	330.7000	.010
Base Flow	0.2200	0.4100	0.5100	0.5341	0.1663	0.6050	0.9700	.024
Sinuosity	0.9700	1.1950	1.2900	1.3256	0.1913	1.3950	2.2100	<.001
Woodland	1.2300	6.5050	9.3600	11.0327	7.4910	12.8150	45.7800	.069
Arable	0.1400	15.9600	36.3700	37.9219	24.4800	54.3900	82.9500	<.001
Grassland	9.9500	22.2800	34.8000	38.5304	19.2820	52.7500	80.9900	.009
Urban	0.0000	3.0150	5.3100	8.6696	10.3945	9.8250	70.4600	.447
Other	0.0000	0.0000	0.0800	3.1884	6.8373	2.7800	40.7500	<.001

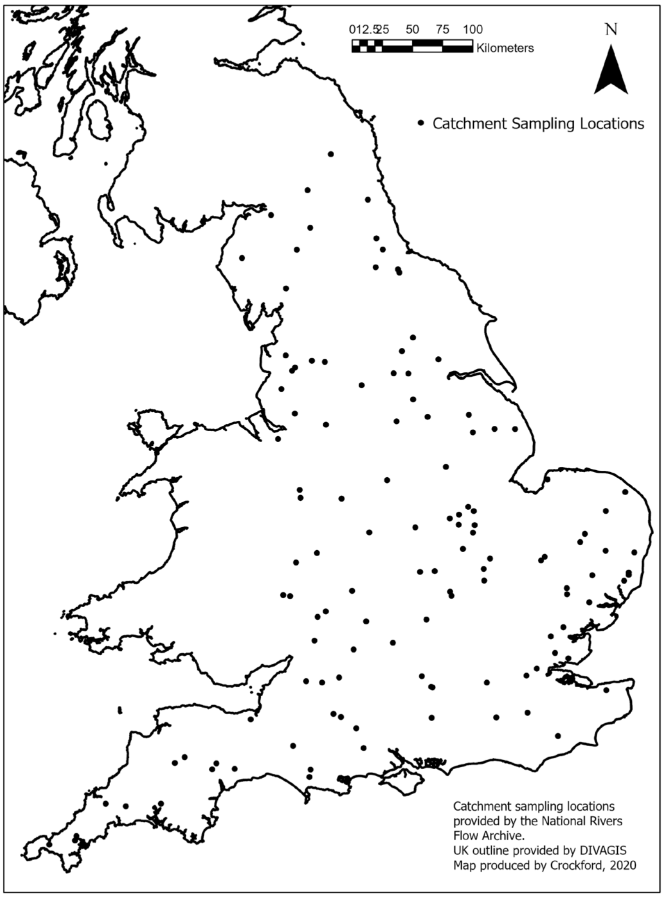

Figure 1 Location of original 136 sampling locations used in this study. Please note that due to thresholds set for dataset size and model fit challenges, the final number analysed was 91

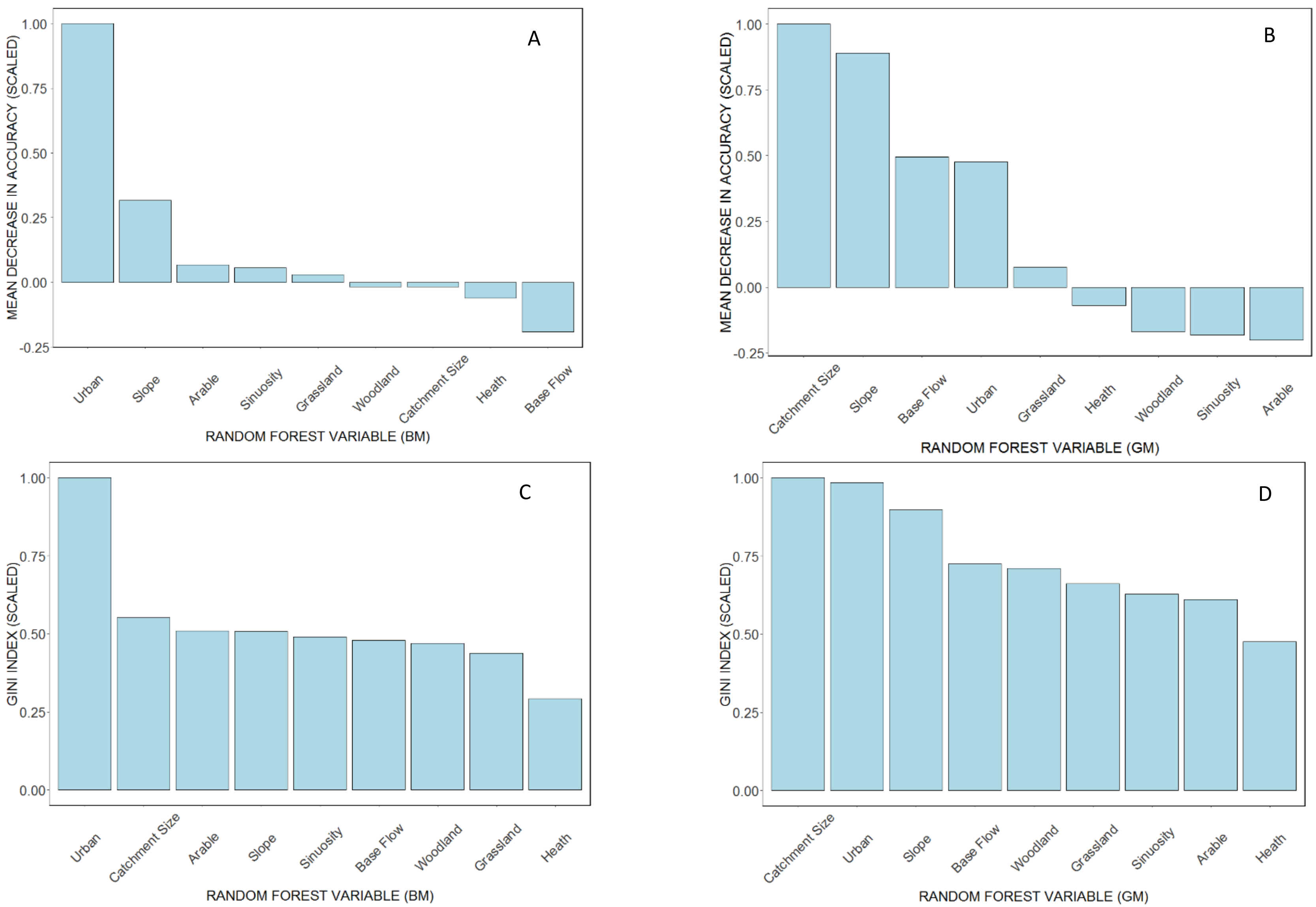

Figure 2 A) Mean Decrease of Accuracy (MDA) of BM forests, B) MDA of GM forests, C) Gini Index of BM forests, D) Gini Index of GM forests
Note: Higher the scaled value, greater the variable importance

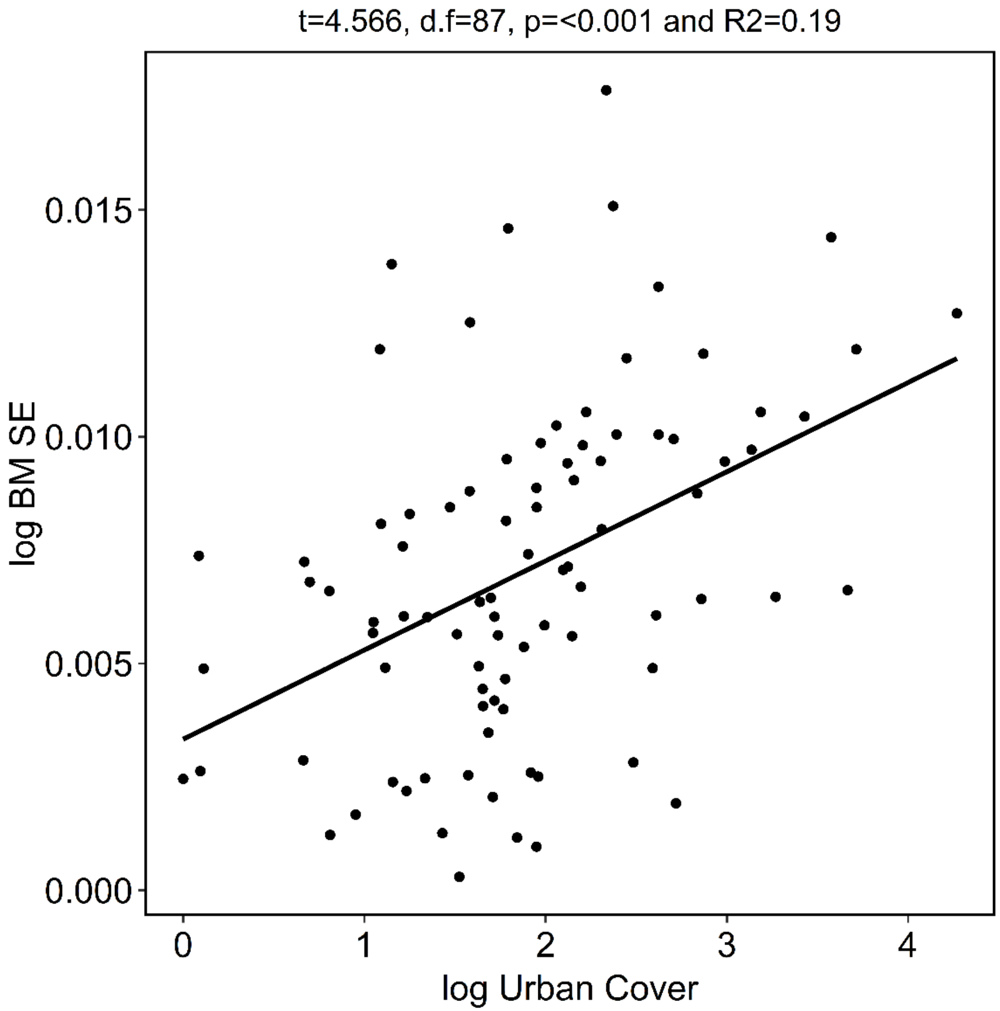

Figure 3. Regression of standard errors (SEs) in BM against measure of urban cover.

Note: post removal of data points with outlying residuals, with both variables increased by 1 to avoid negative numbers and logarithmically transformed.