Quantifying the relative importance of agricultural land use as a predictor of catchment nitrogen and phosphorus concentrations

Science of the Total Environment 954 (2024) 176589

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Quantifying the relative importance of agricultural land use as a predictor of catchment nitrogen and phosphorus concentrations

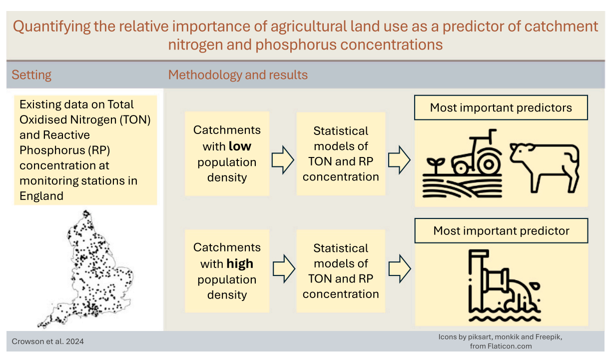
Merry Crowson [a,b,*], Nathalie Pettorelli [b], Nick J.B. Isaac [c], Ken Norris [d], Andrew J. Wade [a]

[a] *Department of Geography and Environmental Science, University of Reading, Reading RG6 6DW, UK*
[b] *Institute of Zoology, Zoological Society of London, Regent's Park, London NW1 4RY, UK*
[c] *UK Centre for Ecology & Hydrology, Wallingford OX10 8BB, UK*
[d] *Natural History Museum, Cromwell Road, London SW7 5BD, UK*

HIGHLIGHTS

- The relative importance of N and P sources in rivers was assessed nationally.
- A data-driven, statistical, approach focused on population density was used.
- Agricultural sources of N and P dominate in catchments with low population density.
- Waste water treatment works dominate in catchments with high population density.
- The findings inform spatially targeted management.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Ashantha Goonetilleke

Keywords:
Water quality
Statistical techniques
Waste water
Spatially targeted management
England

ABSTRACT

Agriculture is a major source of nitrogen (N) and phosphorus (P) in freshwater ecosystems, and different management strategies exist to reduce farmland nutrient losses and thus mitigate freshwater eutrophication. The importance of agricultural sources of N and P as drivers of water quality is known to vary spatially, but quantification of the *relative* importance of the nutrient sources shaping this variability remains challenging, especially with reference to inputs from waste water treatment works. Addressing this knowledge gap is key for targeting management strategies to where they are likely to have the greatest effect. To advance our understanding in this area, this study assesses the impact of population density as a driver of the relative importance of agricultural land use for predicting mean Total Oxidised Nitrogen (TON) and Reactive Phosphorus (RP) concentrations in rivers in England, using two different data-driven, statistical approaches: a generalised linear model and random forest. Our results show that agricultural N and P sources dominate in catchments with low population density, where stream water concentrations are lower and waste water treatment works are numerous, but smaller in terms of the population equivalent served. Agricultural N and P sources are not important predictors of N and P in catchments with high population density, where contributions from waste water treatment works dominate. These results require cautious interpretation, as model validation outcomes show that high TON and RP concentrations are consistently underpredicted. Altogether, our results suggest that the relative contribution of

* Corresponding author at: Department of Geography and Environmental Science, University of Reading, Reading RG6 6DW, UK.
E-mail address: m.r.d.crowson@pgr.reading.ac.uk (M. Crowson).

https://doi.org/10.1016/j.scitotenv.2024.176589
Received 7 June 2024; Received in revised form 26 September 2024; Accepted 26 September 2024
Available online 30 September 2024

agricultural sources may be overestimated in densely populated catchments, relative to point sources from waste water treatment works, and that management strategies to reduce the contribution of agriculture to N and P in rivers may be better targeted towards catchments with lower population density, as this is where agricultural land use is the primary source of N and P.

1. Introduction

Disruption of nitrogen (N) and phosphorus (P) cycling through organic and inorganic N and P fertiliser application, livestock rearing, and wastewater and industrial effluent release, have dramatically increased stream water N and P concentrations since World War II, leading to eutrophication (Howden et al., 2010). In freshwater ecosystems, eutrophication results in a shift to plant communities dominated by fast-growing competitive species (Mainstone and Parr, 2002; O'Hare et al., 2018), excess growth of aquatic weeds and phytoplankton, blooms of harmful algae and the associated negative impacts on invertebrates and fish (Smith and Schindler, 2009). This in turn adversely impacts on a range of water uses and societal benefits, including drinking water abstraction and treatment, livestock watering, water sports, angling, amenity value and tourism (Environment Agency, 2019).

Agriculture is known to be a major source of N and P, and nutrient runoff from agricultural practice is an underlying cause of eutrophication in many catchments (Carpenter et al., 2011; Moss, 2008). N and P reach streams through wash-off and leaching of nutrients from fertiliser and manure applications to arable landscapes, and through soil disturbance and sediment runoff due to land management practices and livestock grazing (Nisbet et al., 2022). Because of this, most studies focusing on diffuse agricultural sources of N and P in England consider arable and horticultural land cover (Davies and Neal, 2007; Bell et al., 2021), as well as cattle and sheep grazing (Johnes et al., 1996; Davison et al., 2008), as the main sources of agricultural N and P in rivers (Defra 2024a and b). A range of measures have been developed to reduce diffuse pollution from agriculture, including reduced fertiliser usage, reduced tillage, and crop rotation (Luna Juncal et al., 2023). There are concerns that these measures do not go far enough to reach water quality targets, which has led to the focus being increasingly put on land cover change, usually from crop to forest, peatland or wetland, that is, from land use that inputs N and P into river systems to one that can capture N and P (Nisbet et al., 2022). This type of natural habitat restoration often targets key areas around the sources, pathways, or receptors of N and P, whilst delivering many other benefits such as habitat creation, shade creation, carbon sequestration and increase access for recreation (Langhans et al., 2022).

However, most of the management strategies aimed at reducing diffuse agricultural sources of N and P are costly to implement and have implications in terms of reduced yield or added management effort for farmers. In some cases, farmers receive renumeration for carrying out management strategies on their land, through payment schemes funded in various ways, for example through taxes (e.g. the Environmental Land Management schemes in England) or water utility companies (Nisbet et al., 2022). Land cover change comes at both an economic and social cost, as taking agricultural areas out of production has implications for food security. Because of this, it is important for management strategies to be spatially targeted to the sites where the measures will have the biggest positive effect on improving water quality in rivers (Withers et al., 2014), through an understanding of the effect of key drivers in different contexts (Spake et al., 2019).

There is some evidence that in densely populated regions the contribution of agriculture to P concentrations in rivers may be less important than previously thought (Withers et al., 2014). A comparison of 10 countries in northwest Europe showed that mean P concentration in rivers were more strongly correlated with discharges associated with urban populations than with agricultural variables (Foy, 2007). In addition to this, a regional study of N concentration in an urban-dominated region showed that urban is the land characteristic which is most important in determining nitrate concentrations (Davies and Neal, 2004), but when the analysis was applied to landscapes across the UK, the area of arable land proved to be more important (Davies and Neal, 2007). This could mean that catchments with a low population density are a better choice for mitigation measures targeting agricultural sources of N and P, if they are shown to be the dominant cause of nutrient enrichment in these rivers, and thus more likely to respond to interventions with improved water quality. However, it is also possible that point sources from waste water treatment works (WWTWs) dominate in catchments with low population density, but with lower concentrations of N and P than catchments with higher population density.

To date, no studies have compared catchments with low population density to catchments with high population density explicitly and at a national scale, with respect to the relative contribution of agricultural sources to N and P concentration. This study aims to fill this gap by using statistical models to test a series of hypotheses, using England as a case study. Data-driven, statistical, approaches provide an interesting and useful contrast to other models that define the relative inputs or flux transfers from different nutrient sources at the outset, for example, export co-efficient modelling and similar (Johnes et al., 1996), since the statistical models determine the relationship between source and instream concentration through model fitting. England was chosen because of the availability of water quality and environmental data, and because catchments with a range of different population densities are available, including catchments with very high population densities. Based on previous work, our hypotheses are:

H1. We expected agricultural sources to be the most important predictor of N and P concentrations in catchments with low population density (Foy, 2007; Davies and Neal, 2007).

H2. We expected effluent from WWTWs to be the most important predictor of N and P concentrations in catchments with high population density, with agricultural sources being less important (Davies and Neal, 2004; Davies and Neal, 2007).

2. Materials and methods

2.1. Data

2.1.1. Dependent variables

We used data from the Water Quality Data Archive (Environment Agency, 2021) on concentrations of Total Oxidised Nitrogen (TON) (Total Oxidised as N in mg/l, representing the sum of nitrate and nitrite, determinand notation 116) and Reactive Phosphorus (RP) (Reactive Phosphorus as P in mg/l, Orthophosphate, determinand notation 180), filtered for measurements taken on a river or running surface water, and taken for monitoring purposes (as opposed to compliance). We chose these forms of N and P because they are much more commonly measured that total N and total P for monitoring purposes in England. For example, the 2019 dataset has 32,753 records for the determinand TON compared to 6064 for Total N, and 30,875 records for RP and none for Total P. We downloaded the data for the years 2015 to 2019 and filtered all available monitoring stations within England to those that had at least one measurement per season per year for this time period, providing us with 528 monitoring stations for TON and 507 for RP. We did this as there is likely to be substantial seasonal variation in the TON and RP concentrations (Shen et al., 2020), and we wanted to make sure that this is captured within the data for all monitoring stations included in the study. We then took the mean value for all the TON and RP concentration measurements for each monitoring station across the five years. We chose to use the mean value rather than the median, as the mean concentration of N and P is currently used in relation to standards for N and P in rivers in England within policy documents (e.g. Defra, 2014). We chose the period 2015–2019 after initial investigations showed that extending this period meant a drop in monitoring stations that met the criteria of having at least one measurement per season, particularly as during the COVID pandemic the number of measurements taken at some monitoring stations dropped considerably, leaving seasonal gaps.

2.1.2. Catchments

To create catchments for the monitoring stations included in the study we snapped the geolocation of each monitoring station to the Centre for Ecology and Hydrology (CEH) 1:50,000 Watercourse Network dataset (Moore et al., 1994) using the r.stream.snap function (Jasiewicz, 2021) in GRASS GIS (GRASS Development Team, 2022) with 2 km as the maximum distance tolerance. We then used the Watershed tool (Spatial Analysis) in ArcGIS Pro (ESRI, 2022) in batch mode, with the Integrated Hydrological Digital Terrain Model (IHDTM) Outflow Direction raster in its native 50 m resolution (Morris and Flavin, 1990, 1994) to automatically delineate a catchment for each monitoring station. The CEH Watercourse Network dataset is consistent with the IHDTM Cumulative Catchment area, so the step of snapping the monitoring station avoids spatial discrepancies between the monitoring stations and the IHDTM Cumulative Catchment Area that would lead to large mistakes in the catchment delineation step.

There were some instances in which the above process did not work, particularly in flat regions such as East Anglia. These cases were usually easy to spot as the resulting catchment were very small (< 0.05 km^2). In these cases, the catchments were created manually through visual inspection of the data and existing maps of catchment available through The National River Flow Archive (2023) and the Defra Catchment Explorer (2023).

There were a few instances in which catchments could not be reliably defined using the methods described above, and these were removed from the dataset. In addition, six catchments were removed because they fall mostly in Scotland and Wales, and thus have differences in data availability compared to England, in particular a lack of information on WWTWs, which are integral to the study. Finally, one catchment was removed because the monitoring station was immediately downstream from a fertiliser factory and had extremely high values for TON concentration. This process left a total of 515 monitoring stations to model concentrations of TON and 494 monitoring stations to model concentrations of RP. However, many of these catchments overlap, that is, they contain each other due to them being on the same river or branch of a river. The observations at monitoring stations that are downstream from each other are not independent from each other (Schreiber et al., 2022), as the water passing through them will be affected by the same conditions, processes and events, leading to pseudoreplication, which is an issue when interpreting the models used in this study (Mets et al., 2017). To avoid this bias, we grouped the catchments that overlap, and selected the catchments with the highest elevation monitoring station within each group. This means that there is a bias towards catchments with a greater ratio of upland to lowland land cover types, but it maximises the number of non-overlapping catchments. This is because, it was possible to keep various monitoring stations, and their associated catchments, on different tributaries by removing a monitoring station lower in the landscape (see Fig. A1, in Appendix A, for a sketch that illustrates this point). This process left a total of 404 monitoring stations to model concentrations of TON and 383 monitoring stations to model concentrations of RP (Fig. 1). For the TON dataset the median catchment size is 51 km^2 and the catchments cover a total area of approximately 31,500 km^2. For the RP dataset the median catchment size is 53 km^2, and the catchments cover a total area of approximately 30,800 km^2. As can be seen in Fig. 1, most of the monitoring stations are included in both datasets (379), with a few only included in the TON dataset (25) or the RP dataset (4).

2.1.3. Independent variables

Based on previous studies, we chose the independent variables for the models of TON and RP based on the environmental characteristics

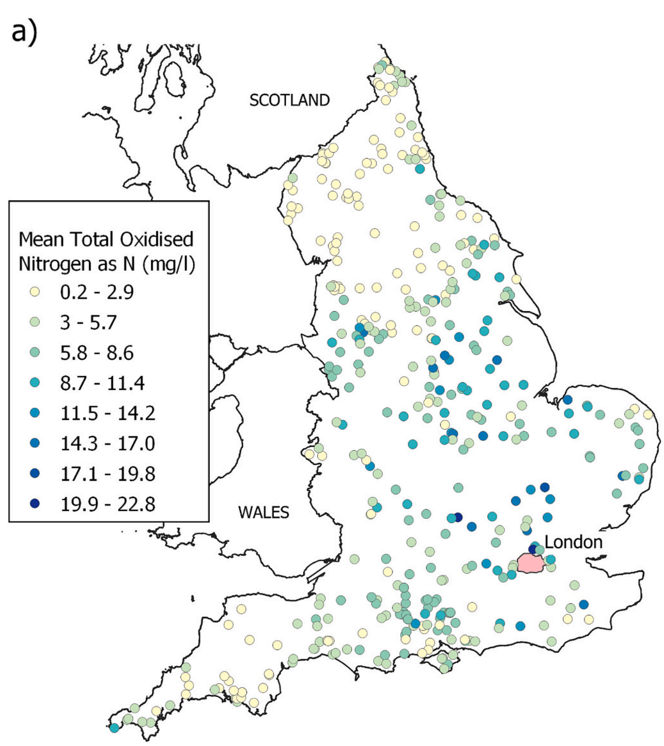


Fig. 1. Mean Total Oxidised Nitrogen (TON) (a) and mean Reactive Phosphorus (RP) (b) for the years 2015–2019 at the monitoring stations initially selected based on minimum data availability, and after ensuring that the monitoring stations' catchments did not overlap. For TON $n = 404$, for RP $n = 383$.

likely to impact river N and P concentrations: the proportion of the catchment with arable and horticultural land cover (Bell et al., 2021), the proportion of area covered by forest (Johnes and Heathwaite, 1997), mean population density in the catchment, catchment cattle and calf density, catchment sheep and lamb density (Davison et al., 2008), catchment maximum mean precipitation, mean slope (Shen et al., 2020), channel density, an estimate of the base flow index based on the Hydrology of Soil Types classification (BFIHOST) (Davison et al., 2008), the proportion of the catchment designated for conservation and/or recreation (Eastwood et al., 2016), catchment area (Virro et al., 2022) and the population equivalent of the WWTWs within the catchment (Redhead et al., 2018). Population equivalent is a parameter for characterizing the polluting potential of industrial wastewaters (in terms of biodegradable organic matter). For the models of TON, we also included the mean atmospheric deposition of N. For the model of RP, we assumed atmospheric P deposition is relatively low, occurring only from wind-blown dust, and is unlikely to show any systematic spatial variation at a small scale (Tipping et al., 2014). A summary of the independent variables included can be found in Table 1.

All the data preparation steps were carried out in R (R Core Team, 2022), unless otherwise stated. To calculate the proportion of each catchment with arable and horticultural land cover we used the CEH Land Cover Map of Great Britain for 2017 at 25 m resolution (Morton et al., 2020). We used the same dataset to calculate the area covered by forest in each catchment, considering the classes "Broadleaved woodland" and "Coniferous Woodland" together. We acquired a list of all WWTWs from the Environment Agency, which included the population equivalent for larger works covered by the Urban Waste Water Treatment Directive, specifically those works that serve population equivalents >2000 (Environment Agency, 2023). For the smaller WWTWs we assigned a value of 1000 for the population equivalent. We mapped the WWTWs based on the grid reference of the outlet and added together the population equivalent of all WWTWs that fall within each catchment.

Mean population density within each catchment was determined using the Output Areas from the 2011 Census for Population Density

Table 1
Independent variables included in the models for TON and RP. All variables are continuous. The independent variable marked with * was only used for the model of TON.

Variable type	Independent variable	Abbreviation
Land cover	Proportion arable and horticultural land cover	ArableHortProp
Land cover	Proportion forest land cover	ForestProp
Waste Water Treatment Works	Population equivalent of waste water treatment works in the catchment. Population equivalent is a parameter for characterizing the polluting potential of industrial wastewaters (in terms of biodegradable organic matter). It expresses the polluting load of a WWTW in terms of the population (number of people) that could produce the same polluting load.	PopEquiWWTW
Catchment size	Catchment area	CatchmentArea
Soil and geology	Estimate of the base flow index based on the Hydrology of Soil Types classification (BFIHOST)	HOSTBaseFlowIndex
Precipitation	Maximum mean annual precipitation for 2015–2019	MaxPrecipitation
Population	Population density	PopDensity
Slope	Mean slope in the catchment	MeanSlope
Atmospheric deposition*	Mean atmospheric deposition of N 2015–17*	AtmosDeposition
Channel density	Channel density	ChannelDensity
Land use	Cattle density	CattleDensity
Land use	Sheep density	SheepDensity
Land use	Proportion of catchment designated for conservation or recreation	DesignatedAreaProp

(Office for National Statistics, 2011). We calculated the mean of the Output Areas within the catchment, weighted by the area of each intersection between the Output Areas and the catchment. To estimate mean cattle density and mean sheep density within each catchment we used data from the England Agricultural Census, 2016 at 5 km resolution on the total number of cattle and calves, and the total number of sheep and lambs (England Agricultural Census, 2016). In each case, the total number was added across the catchment and divided by the area of the Agricultural Census grids that intersect with the catchment to estimate stocking densities.

To calculate the maximum mean annual precipitation for each catchment we used the HadUK-Grid rainfall data, averaged by year, on a 1 km grid over the UK (Met Office, 2020). We took the mean by grid for the years 2015 to 2019 and chose the maximum value that fell within each catchment.

Mean slope was computed for each catchment using the Slope tool (Spatial Analysis) in ArcGIS and the IHDTM Digital Elevation Model (Morris and Flavin, 1990, 1994). To determine channel density, we used the CEH 1:50,000 Watercourse Network dataset (Moore et al., 1994) to calculate the length of channels within each catchment and divided this by the catchment's total area (Rahman and Rahman, 2020). To account for the soil and geology we calculated a base flow index (BFIHOST) for each catchment. We used the Hydrology of Soil Types (HOST) dataset (Boorman et al., 1995; Griffin et al., 2019) as the basis for our calculations, following the area-weighting method in the Flood Estimation Handbook volume 5 (Bayliss, 1999; Griffin et al., 2019).

To calculate the proportion of each catchment designated for conservation or recreation, we acquired the shapefiles for terrestrial designated areas on mainland England, based on those described in Lawton et al. (2010). The designation types considered are National Parks, Areas of Outstanding Natural Beauty, Ramsar Sites, Special Areas of Conservation, Special Protection Areas, Local Nature Reserves, National Nature Reserves and Sites of Special Scientific Interest (SSSIs) ($n = 6349$).

For the models of TON concentration, we included the mean atmospheric deposition of N in the catchment, based on N deposition data at 1 km resolution, from the UK CEH Environmental Information Data Centre (Tomlinson et al., 2020). We used the period 2015–17, as this data was not available for after 2017. We took the mean value from all points within the catchment for each of the four forms of atmospheric deposition (dry deposition of reduced N, dry deposition of oxidised N, wet deposition of reduced N and wet deposition of oxidised N) and added them together to produce a single value.

2.2. Analysis

Our analysis uses catchments characteristics to explain the variation in N and P concentrations at monitoring stations at a national scale. The dependent variable used in the statistical models is either the mean TON or mean RP concentration at monitoring stations between 2015 and 2019 (as described previously in Section 2.1.1). Thus, each row in the dataset represents a monitoring station, and there is a single summary value of TON and/or a single summary value of RP for that site. The independent variables are a series of catchment characteristics summarised to a single value for each monitoring station's catchment (such as proportion of different land cover types, mean slope, etc., as described previously in Section 2.1.3). Agricultural land use is represented by three different variables: proportion of catchment with arable and horticultural land cover, cattle density and sheep density, representing the main agricultural sources of N and P in England (Defra, 2024a and b).

To test out first hypothesis, we selected the catchments from the TON dataset with a population density below the first quantile for population density (population density < 0.41 people/ha, $n = 101$, Fig. 2) to create a group of catchments to represent low population conditions. We did the same thing for the RP dataset (population density < 0.40 people/ha, $n = 96$, Fig. 2).

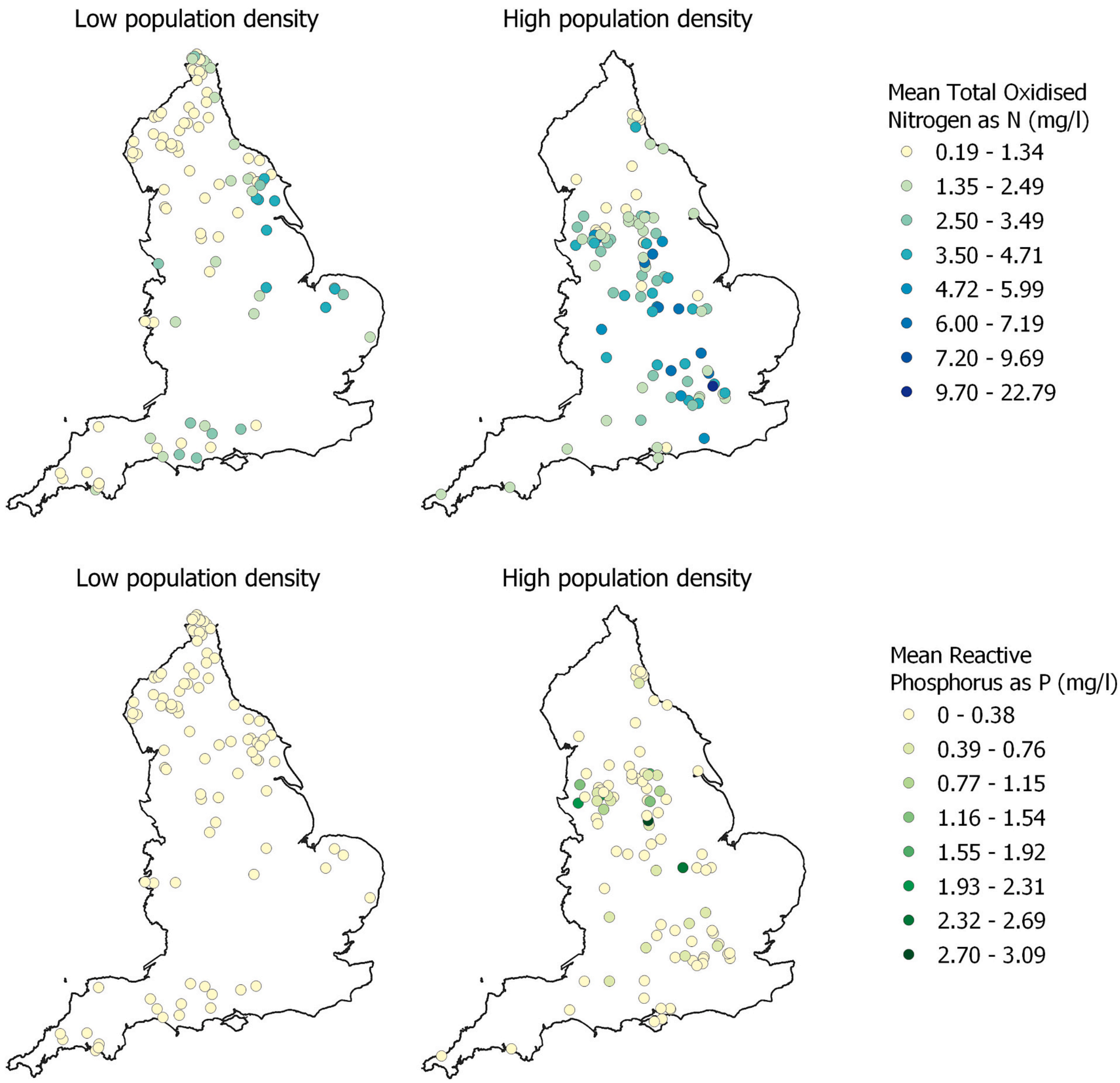


Fig. 2. Mean TON (top two maps) at monitoring stations with catchments with low population density and catchments with high population density for the years 2015–2019 (n = 101 in each case). Mean RP (bottom two maps) at monitoring stations with catchments with low population density and catchments with high population density for the years 2015–2019 (n = 96 in each case). Catchments with low population density are those below the first quantile for population density in the full dataset (population density < 0.41 people/ha for TON, population density < 0.40 people/ha for RP), and catchments with a high population density are those above the fourth quantile for population density in the full dataset (population density > 3.61 people/ha for TON, population density > 3.25 people/ha for RP).

To test our second hypothesis, we selected catchments from the TON dataset with population density above the fourth quantile for population density (population density > 3.61 people/ha, n = 101, Fig. 2) to create a group of catchments to represent high population conditions. We repeated the process with the RP dataset (population density > 3.25 people/ha, n = 96, Fig. 2). We chose the lower and upper quantile as cut off points for our two groups of catchments because this allows us to look for differences between two strongly contrasting groups.

Histograms of the distribution of the dependent variable for each of these four datasets can be seen in Appendix B, Fig. B1. We ensured that the independent variables were not strongly correlated in each case (Pearson's correlation coefficient < 0.75), as this is a requirement when interpreting the statistical methods used in this paper. The distribution of the independent variables for each of the four datasets can be found in Appendix B (Fig. B2 for TON and Fig. B3 for RP).

2.2.1. *Statistical models*

We used two methods to model each of the four datasets: negative binomial generalised linear model and random forest. We chose to use both a negative binomial generalised linear model and random forest model in each case to check for consistency of results across models, and to make use of the different strengths of the two approaches.

Generalised linear models were chosen for use in this study because they are highly interpretable, as the coefficients are a robust way to gain insight into the relationships between the independent and dependent variables, and the relative importance of the dependent variables (Zuur et al., 2009). They are a generalised form of linear regression (Crawley, 2007), that have previously been used in water quality research to study pond water quality in the United Kingdom (Spake et al., 2019) and the occurrence of macroinvertebrates in Guayas River Basin, Ecuador (Damanik-Ambarita et al., 2016).

Random forest is a widely used machine learning technique, including in environmental science (e.g. Cutler et al., 2007; Molnar, 2023; Ross et al., 2021), and was chosen for use in this study due to the algorithm's ability to deal with nonlinear interactions and excellent predictive capability (Yu et al., 2021). Random forest is a method based on an ensemble of decision trees, that use randomly selected predictor variables for each tree, as well as randomly selected training data subsets (Breiman, 2001a). Random forest has been successfully used in the past to model and predict N and P concentrations at a national scale in the USA (Shen et al., 2020), and annual total nitrogen and total phosphorus concentrations in Estonia (Virro et al., 2022). Here we make a first application to N and P concentrations measured across catchments in England. An advantage of the random forest approach (compared with GLM) is its ability to detect non-linear relationships. However, the results from random forest models are less easy to interpret and there is a risk of over-fitting (Saarela and Jauhiainen, 2021). Given these complementary strengths and weaknesses, the similarities and differences between the results from these two methods provide valuable insights for our study.

2.2.2. Model training and validation

We split each dataset into training and test data using a 70/30 split, and used the same training and test dataset for both the generalised linear model and random forest model in each case.

We used the negative binomial model with a log link function to model the concentrations of RP and TON, implemented using the function glm.nb in the MASS package in R (Venables and Ripley, 2002; R Core Team, 2022). We converted the concentrations to integer values (multiplication by 1000). Fixed covariates considered for the models of TON and RP were the proportion of arable and horticulture land cover within the catchment; log of the proportion of forest land cover; log of the population equivalent of the WWTWs within the catchment; population density; log of the density of cattle in the catchment; log of the density of sheep; log of the maximum mean precipitation; the mean slope in the catchment; channel density; the HOST base flow index; log of the proportion of the catchment that is designated for conservation and/or recreation; log of the total area of the catchment. The atmospheric deposition of N was also included as a fixed covariate for the model of TON.

Akaike's Information Criterion (AIC) was used as the selection criteria for independent variables to be included in our final best models. We used a stepwise approach, starting with a 'maximal' model including all the fixed covariates and conducting backward model selection (Zuur et al., 2009) using the function stepAIC in the MASS package (Venables and Ripley, 2002).

All the covariates were standardised so that the coefficients were comparable. Model assumptions were verified by plotting residuals versus fitted values and against each covariate.

To build random forest models for TON and RP, we trained the randomForest function in the R package randomForest (Liaw and Wiener, 2002). To optimise model performance, there are two parameters that need tuning: the number of features to select when splitting trees (mtry) and the number of trees to grow (ntree) (Liaw and Wiener, 2002). We used the function tuneRF, also in the randomForest package (Liaw and Wiener, 2002), to set the value for the parameter mtry (mtry = 3 for the model of TON for catchments with low population density, mtry = 4 for the model of TON for catchments with low population density, mtry = 3 for the model of RP for catchments with low population density, and mtry = 1 for catchments with high population density). We set the parameter ntree to the default of 500, based on various trial runs and recommendations in the literature (Belgiu and Drăgu, 2016). We trained the random forest using the same independent variables as were selected in the best negative binomial generalised linear model, however we did not log any of the variable or scale them, as random forest is invariant to such transformations of the independent variables, and they make model interpretation more difficult.

For all eight models we calculated the root mean square error (RMSE), plotted the test data against the concentrations predicted by the model, and calculated the strength of the correlation between these observed and predicted concentrations using Pearson's correlation coefficient. The variation explained by our generalised linear models and random forest models was calculated using the test data, following the method implemented in the randomForest package, using the formula: 1 – mse / Var(y).

We assessed the model's residuals for spatial autocorrelation by creating a map of the residuals for visual inspection, calculating Moran's I, and by plotting a distance-based semivariogram.

2.2.3. Effect size and variable importance

For the negative binomial generalised linear models, we created effect plots for the independent variables, with all other variables kept at their mean. For the random forest models, we extracted variable importance measures using the importance function in the randomForest package (Liaw and Wiener, 2002). There is little consensus in the machine learning literature on how to best calculate the relative importance of different independent variables (Yu et al., 2021), so we report two widely used methods to rank predictor variables associated with random forest: mean decrease in accuracy and mean decrease in node impurity. Mean decrease in accuracy is computed by permuting each independent variable in the random forest, comparing the prediction error using the out of bag data, and assessing the increase in error (mean square error) when each target variable is randomized (permuted) (Liaw and Wiener, 2002; Yu et al., 2021). Mean decrease in node impurity is the total decrease in node impurities (residual sum of squares) from splitting on the variable, averaged over all trees. Both are implemented within the randomForest package (Liaw and Wiener, 2002).

3. Results

3.1. Model validation

The Pearson correlations between predicted and observed values for the models of TON are in the range of 0.6–0.9 (df = 28, $p < 0.001$ in all cases) across the testing sets (Fig. 3), and for the RP datasets they are in the range of 0.39–0.84 (df = 28, p < 0.001 in all cases) across the testing sets (Fig. 4). The models for RP and TON underestimate the higher values in the dataset. The RMSE for the models are shown in Fig. 3 and Fig. 4.

The best negative binomial models for TON explain 41 % of the variation in the test data for catchments with low population density and 35 % in catchments with high population density. The best negative binomial models for RP explain 47 % of the variation in the test data in catchments with low population density and 61 % for the catchments with high population density. The variation in the test data explained by random forest is 77 % for the TON model of catchments with low population density, 44 % for the TON model of catchments with high population density, 5 % in the case of the RP model for catchments with low population density and 27 % for the RP model of catchments with high population density. Moran's I analyses on the model residuals shows no significant spatial autocorrelation in the residuals of any of the models (p-value >0.05) relevant to the scale of analysis.

3.2. Effect size and variable importance

As expected under (H1), the generalised linear model for TON shows that agricultural sources, namely arable and horticultural land cover, and cattle density, are significant positive predictors of TON in catchments with low population density, whilst the population equivalent of WWTWs is not a significant predictor of TON in these catchments (Table 2a). This is confirmed by the random forest models of TON in catchments with low population density, as it ranks arable and horticultural land use as one of the top two predictors of TON (with the other predictor being the HOST base flow index) (Fig. 5a).

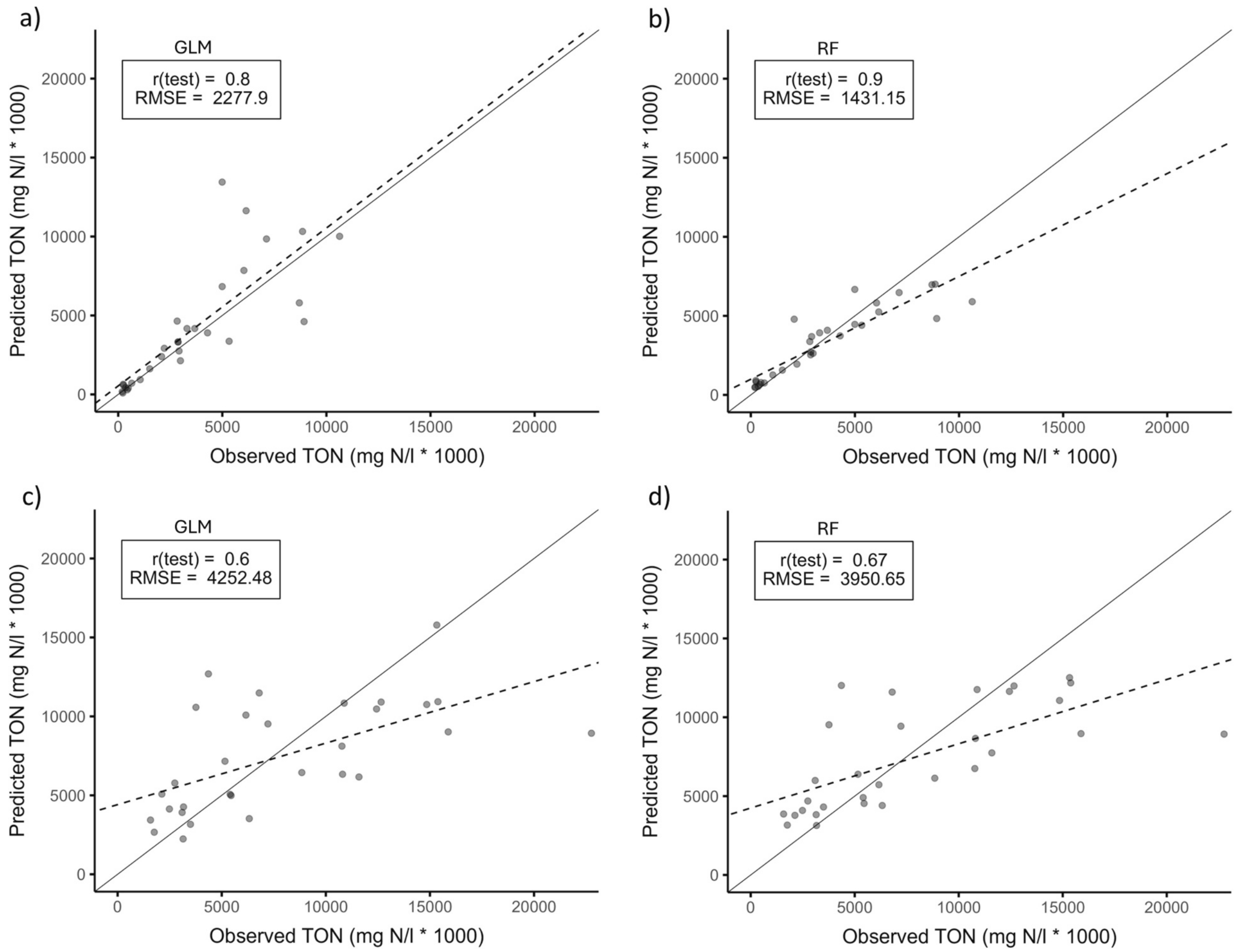


Fig. 3. Correlation plots for the out-of-bag test data (n = 30) for the negative binomial generalised linear model (a) and random forest model (b) of TON in catchments with low population density, and the negative binomial generalised linear model (c) and random forest model (d) of TON in catchments with high population density. Horizontal axes show the true values from the test data set, multiplied by 1000 to obtain an integer, whilst the vertical axes show the values predicted by the model. The dashed line shows the linear regression of the data points and the solid line represents the 1:1 relationship. The box in the upper left corner gives the Pearson coefficient and the value for RMSE for each model. The labels GLM and RF refer to generalised linear model and random forest respectively.

The results from the models of RP in catchments with low population density lend some support to (H1), as arable and horticultural land use, and cattle density, are both significant positive predictors of RP in the negative binomial generalised linear model of catchments with low population density (Table 2b), and have a bigger effect size than the population equivalent of WWTWs in the same model (Table 2b, and Fig. B4 in Appendix B). However, the population equivalent of WWTWs is still a significant predictor in catchments with low population density, and sheep density has a negative effect. The results from the random forest model of RP in catchments with low population density show that arable and horticultural land use and cattle density rank higher than the population equivalent from WWTWs (Fig. 5b), however the low R^2 for this random forest model, as well as the evidence of overfitting (Fig. 4b), means that these results should be interpreted with caution.

As expected under (H2), the generalised linear model for TON in catchments with high population density shows that arable and horticultural land cover and sheep density are not significant predictors of TON, whilst the population equivalent of WWTWs is a significant positive predictor of TON (Table 2c), with a large effect size (Table 2c, Fig. B5a in Appendix B). The random forest model for TON in catchments with high population density also supports (H2), with the population equivalent of WWTWs ranking as the top predictor of TON (Fig. 6a).

The results for RP in catchments with high population density also support (H2), as arable and horticultural land cover is a negative predictor of RP (Table 2d), whilst the population equivalent of WWTWs has a comparatively large positive effect on RP concentrations in these catchments (Table 2d, Fig. B5c in Appendix B). The random forest for RP in catchments with high population density confirmed this, with the population equivalent of WWTWs ranking as one of the top predictors (with the other being catchment area) (Fig. 6b).

Some of the other catchment characteristics are important to predict TON and RP in the models, beyond agricultural land use and the population equivalent from WWTWs. For example, the HOST baseflow index was found to be more important than both arable land cover and cattle density in the models of TON for catchments with low population density (Table 2a and Fig. 5a). Other independent variables that have comparatively high coefficients (positive or negative) in the generalised linear models and rank highly in the random forest variable important measures are the maximum average yearly precipitation, catchment area and mean slope (Table 2, Fig. 5 and Fig. 6).

4. Discussion

4.1. Summary of findings and implications for management

In this study we use data-driven, statistical techniques to model N and P concentrations nationally in England, comparing the results from catchments with low population density to catchments with high population density, and demonstrate how these techniques can be used to

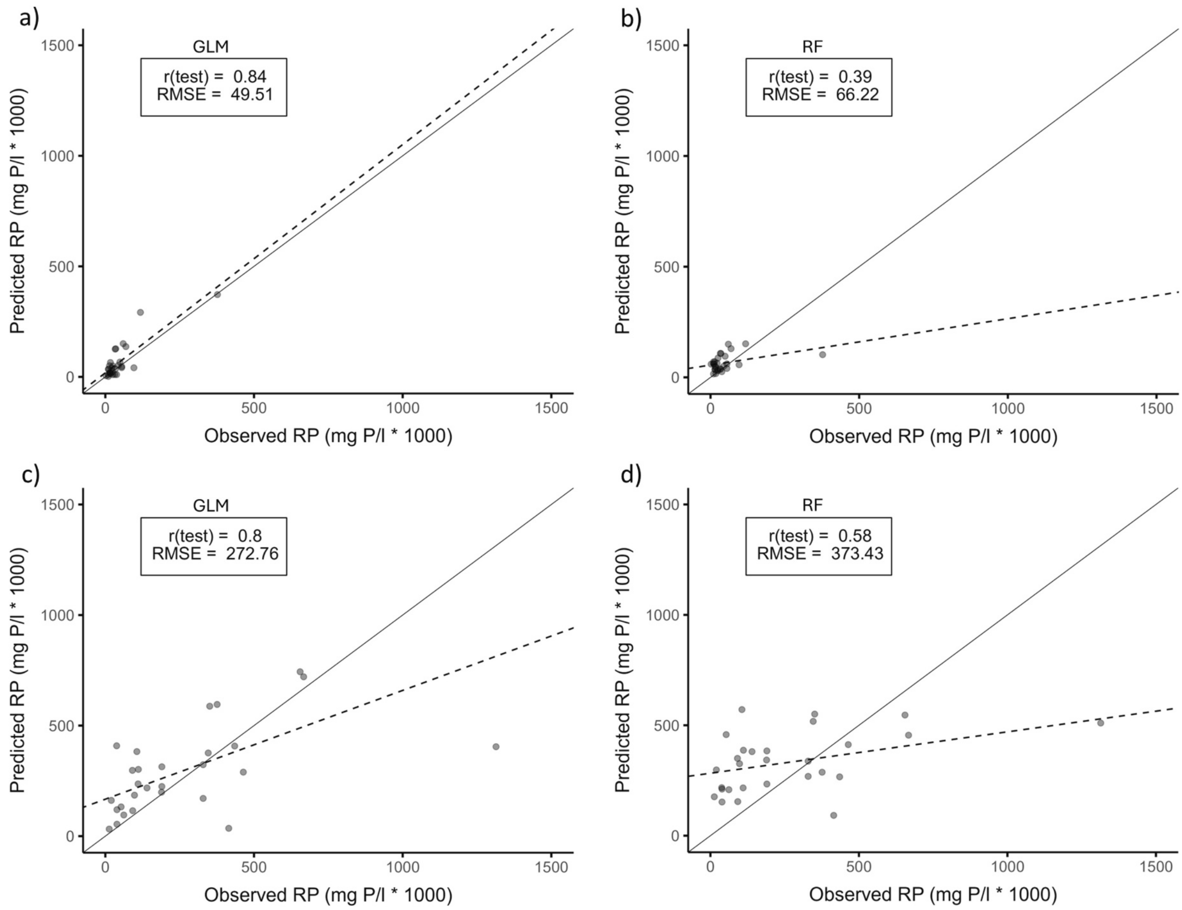


Fig. 4. Correlation plots for the test data ($n = 29$) for the negative binomial generalised linear model (a) and random forest model (b) of RP in catchments with low population density, and the negative binomial generalised linear model (c) and random forest model (d) of RP in catchments with high population density. Horizontal axes show the true values from the test data set, multiplied by 1000 to obtain an integer, whilst the vertical axes show the values predicted by the model. The dashed line shows the linear regression of the data points and the solid line represents the 1:1 relationship. The box in the upper left corner gives the Pearson coefficient and the value for RMSE for each model. The labels GLM and RF refer to generalised linear model and random forest respectively.

understand the relative importance of different sources of N and P in a way that is relevant to management and policy. Our models for N and P show satisfactory predictive ability for the most part, showing there is potential to use this approach more widely, although they consistently underestimate very high mean TON and RP concentrations (approx. > 12 mg N/l and approx. > 0.8 mg P/l), and the model validation results were poor for some of the models of RP. In terms of management and policy, our results suggest that action to reduce agricultural runoff in low population catchments is needed to mitigate nutrient impairment, as this is the dominant source of N (and to some extent P), although of course reductions of inputs from small WWTWs (and septic tanks) will also be beneficial. Management efforts in catchments with high population density should prioritise reducing inputs from WWTWs sources, as it is the dominant predictor of N and P concentrations in these catchments. These results lend support to previous suggestions (e.g. Withers et al., 2014) that the contribution of agricultural sources may be overestimated in catchments that are densely populated, relative to point sources from WWTWs.

This study has shown that agricultural sources of N and P are *comparatively* more important in catchments with lower population density. The mean concentrations of TON and RP are lower in the catchments with lower population density, so WWTWs could have been just as important in determining these concentrations as the WWTWs in the higher population density catchments (because the concentrations are lower and therefore, even though the WWTWs are smaller, their relative effect could have been the same or greater). It seems likely that as population increases so do other sources of N and P that were not directly included in the model – such as runoff from roads and urban areas, industrial effluent, illegal discharges and septic tanks – which could cumulatively become an important source of N and/or P, and thus makes agricultural sources comparatively less important as a predictor in these catchments. There is a slight correlation between population density and the population equivalent from WWTWs (Pearson's correlation <0.37 for all datasets), as would be expected, and this will go some way towards explaining why the population equivalent of WWTWs is an important predictor of N and P concentration in catchments with high population density. Previous studies have suggested that P enrichment is more likely to be the cause of nutrient impairment in lowland, high alkalinity rivers (Jarvie et al., 2018), and with this in mind, a continued emphasis on RP reduction at WWTWs in urban areas is likely to be the right approach to improving water quality in catchments with high population density.

Whilst this study focuses on the relative contribution of agricultural land use and WWTWs to N and P concentrations, our results also show that other catchment characteristics play an important role. The high importance of the variable representing geology in our models (HOST base flow index) is in line with a previous study in which the geological predictor (proportion of catchment located on limestone) was within the top ten for the feature importance ranking in national-scale random forest models of total nitrogen and total phosphorus in Estonia (Virro et al., 2022). This highlights the way that different catchment characteristics mediate the final effect that a particular source of N or P will have on nutrient concentrations in the stream. With this in mind, it is important to highlight that the relationships we found in our models are

Table 2
Formula, estimated regression parameters, standard errors, z-values and *p*-values for the minimum adequate negative binomial generalised linear models of catchments with low population density for the TON data (a) and RP data (b), and of catchments with high population density for the TON data (c) and RP data (d). Model R^2 are 0.41, 0.47, 0.35, and 0.61 respectively.

(a). $TON_{low\ pop}$ ~ ArableHortProp + MeanSlope + LogCatchmentArea + LogCattleDensity + PopDensity + LogMaxPrecipitation + HOSTBaseFlowIndex + LogDesignatedAreaProp

	Estimate	Std. error	z value	p-value
Intercept	7.55	0.04	180.35	< 0.001
ArableHortProp	0.35	0.08	4.52	< 0.001
MeanSlope	0.26	0.08	3.22	< 0.05
LogCatchmentArea	0.09	0.05	1.79	0.073
LogCattleDensity	0.30	0.05	6.17	< 0.001
PopDensity	0.22	0.05	4.07	< 0.001
LogMaxPrecipitation	−0.57	0.08	−6.94	< 0.001
HOSTBaseFlowIndex	0.36	0.05	7.08	< 0.001
LogDesignatedAreaProp	−0.15	0.06	−2.29	< 0.05

(b). $RP_{low\ pop}$ ~ LogForestProp + ArableHortProp + ChannelDensity + LogCatchmentArea + LogCattleDensity + LogSheepDensity + PopDensity + LogPopEquiWWTW

	Estimate	Std. error	z value	p-value
Intercept	3.68	0.07	51.11	< 0.001
LogForestProp	0.17	0.10	1.74	0.081
ArableHortProp	0.31	0.10	3.09	< 0.01
ChannelDensity	0.19	0.10	2.01	< 0.05
LogCatchmentArea	−0.49	0.11	−4.68	< 0.001
LogCattleDensity	0.45	0.11	4.30	< 0.001
LogSheepDensity	−0.33	0.14	−2.40	< 0.05
PopDensity	0.33	0.11	2.94	< 0.01
LogPopEquiWWTW	0.30	0.09	3.34	< 0.001

(c). $TON_{high\ pop}$ ~ MeanSlope + ChannelDensity + LogCatchmentArea + LogCattleDensity + HOSTBaseFlowIndex + LogPopEquiWWTW

	Estimate	Std. error	z value	p-value
Intercept	8.71	0.05	188.26	< 0.001
MeanSlope	−0.25	0.05	−4.77	< 0.001
ChannelDensity	−0.13	0.05	−2.50	< 0.05
LogCatchmentArea	−0.24	0.07	−3.55	< 0.001
LogCattleDensity	0.08	0.05	1.47	0.141
HOSTBaseFlowIndex	0.13	0.05	2.61	< 0.01
LogPopEquiWWTW	0.52	0.07	7.62	< 0.001

(d). $RP_{high\ pop}$ ~ ArableHortProp + LogCatchmentArea + LogMaxPrecipitation + HOSTBaseFlowIndex + LogPopEquiWWTW

	Estimate	Std. error	z value	p-value
Intercept	5.65	0.09	65.63	< 0.001
ArableHortProp	−0.25	0.11	−2.32	< 0.05
LogCatchmentArea	−0.57	0.12	−4.69	< 0.001
LogMaxPrecipitation	−0.31	0.11	−2.92	< 0.01
HOSTBaseFlowIndex	−0.35	0.09	−3.78	< 0.001
LogPopEquiWWTW	1.06	0.12	9.28	< 0.001

The independent variables are the proportion of arable and horticultural land cover; channel density; the HOST base flow index for the catchment; log of the catchment area; log of the cattle density; log of the proportion of the catchment designated for recreation or nature conservation; log of the proportion of forest land cover; log of maximum average yearly precipitation; log of the population equivalent for all the WWTWs in the catchment; log of the sheep density; mean slope; population density.

a generalisation based on a large-scale assessment, and regional differences in the drivers of water quality are likely to exist (Pharaoh et al., 2024). For local decision making many other aspects will be relevant, including the local environmental conditions, and social and economic aspects. Moreover, this approach is limited to making recommendations around large-scale land cover and land management (e.g. livestock density) changes. Quantification of the overall effectiveness of smaller scale measures, for example buffer strips, contour ploughing, at the national, or catchment, scale remain elusive.

4.2. Data and study limitations

There are several limitations to the data used in the study. Firstly, the geographical distribution of the catchments with low population and high population density used in this study are not identical (as seen previously in Fig. 2), which means we cannot completely rule out that the signal being picked up is due to some other variable that varies regionally and is not accounted for by the model, such as the main crop type in arable areas, or the distribution of industry. We also did not have access to the type of treatment applied to waste water at the different WWTWs, nor the exact population equivalent for the smaller WWTWs (those with population equivalent < 2000). Access to this information in a standardised way across England, Scotland and Wales would help provide a more nuanced picture of the contribution of WWTWs to N and P to freshwater ecosystems, and avoid the need to restrict the analysis to the administrative boundaries of England, as is currently the case. Finally, whilst the consistency between the two different models is reassuring (generalised linear models and random forest), more could be done in the future to assess the effect of data uncertainty, where alternative datasets exist or the approximate error in the data is known.

There are other considerations when interpreting the results. Firstly, the findings of this study cannot be applied to catchments with characteristics outside of the range available for inclusion in this study, such as upland catchments. Secondly, a limitation of this study is that the models are not giving information about particulate transport of N and P, and not separating between organic and inorganic forms. This means that the results are only able to present a partial picture of N and P retention, and this is likely to be particularly important for P, as the particulate transport pathways are known to be important (Reaney et al., 2011). However, the use of TON and RP means that the focus is on the predominant forms that affect plant growth, as they are readily available for uptake (Prasad and Chakraborty, 2019; Angus et al., 2013). There is also not currently enough data available on Total N and Total P concentrations, or organic N and P, at monitoring stations in England to use the approach presented in this study on these deteminands. Thirdly, there is the question of spatial configuration of the catchment characteristics. The models give insight into the importance of various catchment characteristics which have been summarised at a catchment level, in generally large catchments, but the situation may be very different at a local scale. Certain catchments characteristics, such as the proportion of area covered by forest, were not found to be important in predicting TON and RP in this study, but they may or may not play a role more locally in patches or as buffer zones along a river. Finally, defining the catchment for each of the monitoring stations in an automated way was challenging, and although we carried out a large quality-control effort through visual inspection and comparison with other available datasets, it is possible that some mistakes remain in catchment definition, which would then affect all independent variables for that monitoring station.

4.3. Future research directions

Much progress has been made to better understand the sources and dynamics of natural and anthropogenic inputs of N and P into rivers (e.g. Jarvie et al., 2018; Johnes et al., 2022). However, modelling N and P in rivers at large scales remains challenging, and different approaches have emerged to tackle the problem, including empirical models, such as the export coefficient models (e.g. Redhead et al., 2018; Johnes et al., 1996), as well as processed based models (e.g. the LTLS Freshwater Model described in Bell et al., 2021). The increasing availability of large and often publicly available datasets with water quality measurements and other environmental data has led to an increase in the use of statistical

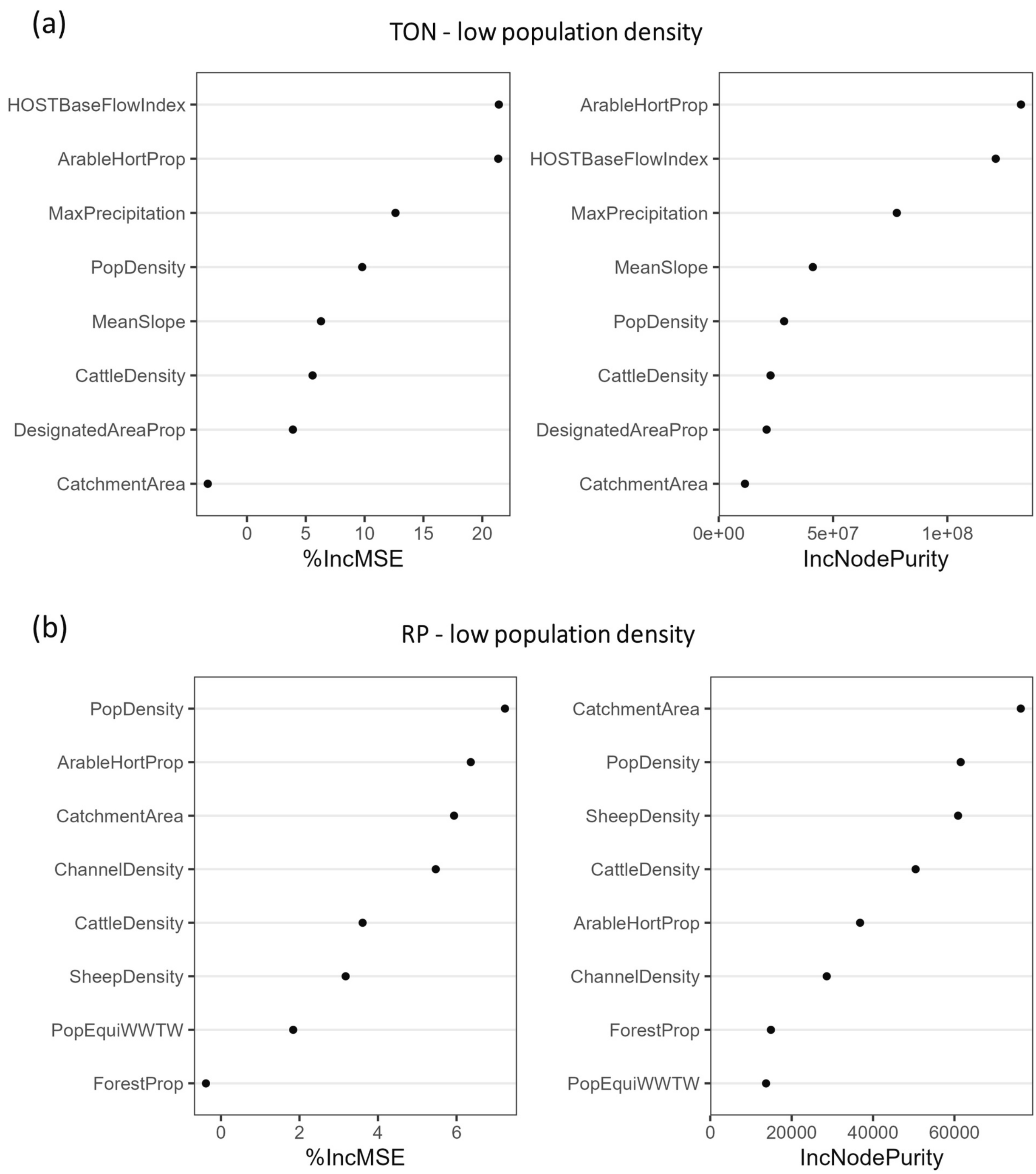


Fig. 5. Variable importance for the random forest for TON in catchments with low population density (a), RP in catchments with low population density (b). In each case the left hand panel shows the permutation importance and the right hand panel the Gini importance.

techniques to study water quality (Schreiber et al., 2022; Spake et al., 2019; Moorhouse et al., 2018; Tate et al., 2003), as we have done in this study. These techniques are in the spirit of a wider body of work that aims to develop data science and artificial intelligence techniques for the natural environment (Blair, 2021; Scowen et al., 2021; Breiman, 2001b; Lucas, 2020), in the hope that environmental science and ecology can reap the benefit of the increasing quantity and diversity of data available to researchers. However, the approach we used has limitations, including inconsistent results between different measures of variable importance in some cases, and results that may not make sense from a process perspective, such as the negative effect of arable and horticultural land use on RP concentrations in the generalised linear model for catchments with high population density. Future work in this area could explore the use of other statistical models, including other machine learning models. Overall, however, the results from the generalised linear models and random forest models were fairly consistent with each other for each dataset, which is reassuring. The model validation results do, however, highlight that outliers have a strong effect on the models and that particularly random forest tended to overfit in these situations.

An interesting avenue going forward would be to use the models in this study to make predictions, that would allow a more nuanced discussion of the decrease in nutrient concentration that could be expected to result if land cover change were to take place in different contexts. It is important to interpret any output from the models, whether it be the measures of variable importance or predictions, in the context of broader temporal and spatial change, for example the effect of national

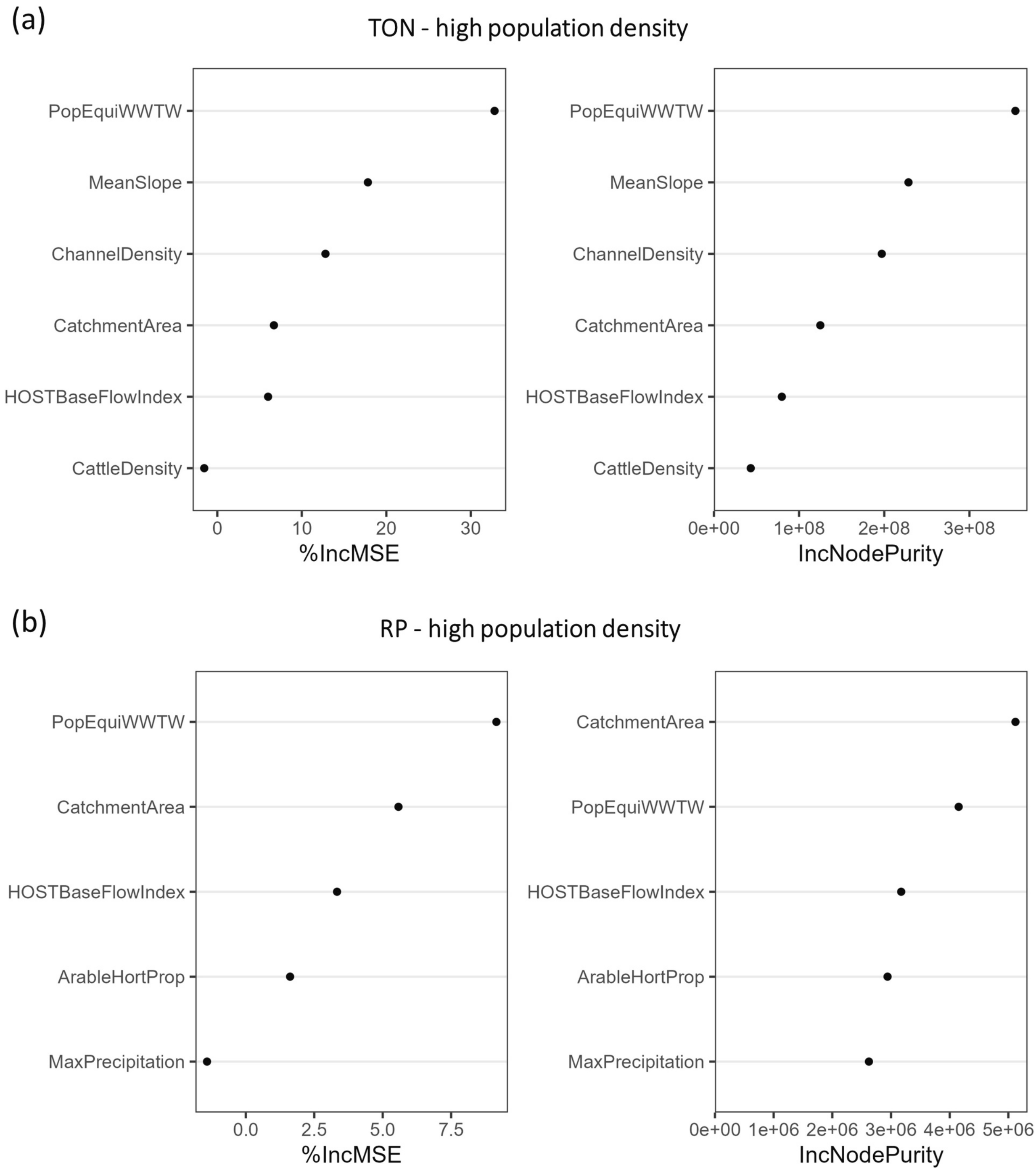


Fig. 6. Variable importance for the random forest for TON in catchments with high population density (a) and RP in catchments with high population density (b). In each case the left hand panel shows the permutation importance and the right hand panel the Gini importance.

policy change or climate change, which could have important impacts on N and P concentrations in the mid to long term. In addition to this, adapting the approach to take seasonality into account in some way is probably important to be able to discuss to what extent the findings of this study are ecologically meaningful, as variable importance may well vary seasonally and the effect of changes to catchment characteristics may also be sensitive to seasonality. For example, it would help to consider ecologically sensitive periods in spring and summer when rooted aquatic plants and algae grow (Jarvie et al., 2006). Finally, conducting this type of study in other countries or regions would help understand to what extent the findings of this study are specific to England or represent more general patterns due to the nature of N and P transport and retention.

5. Conclusion

The sources and dynamics of natural and anthropogenic inputs of N and P into rivers remains a complex problem, and despite substantial domain knowledge about these processes it remains challenging to model. However, understanding the relative importance of diffuse sources from agriculture, and how this varies in different contexts, is important because it allows us to spatially target management strategies to the places where they are likely to have the strongest positive effect. The results of this study need to be interpreted with some caution, but they do provide some insight and recommendations. Firstly, our results suggest that management strategies aimed at reducing N and P from agricultural sources might be better suited to catchments with low (ca. < 0.4 people/ha) population density. This is based on our finding that the predictors relating to agricultural sources were more important than the population equivalent of WWTWs in these catchments, as they were found to have a larger effect size in the generalised linear models, and ranked higher within the variable importance measures in the random forest models for these catchments. Secondly, they suggest that to reduce the concentration of TON and RP in catchments in England with high (ca. > 3.6 people/ha) population density, a continued focus on WWTWs as point sources should be a priority, as the population equivalent of WWTWs was shown to be the most important variable in all of the models for these catchments. This is a generalised suggestion based on a national-scale assessment, local factors would most likely also be important in any decision-making process. Going forward, more could be done to make detailed data on WWTWs available, including their population equivalent and the type of treatments applied, which would make it easier to include this independent variable in all types of models.

Climate change is likely to increase pressure on river systems and the ecosystem services they support, which will increase the need to target management strategies to preserve the benefits we receive from nature. The debate about the relative contribution of diffuse agricultural sources and point sources from WWTWs to N and P concentrations in rivers will only become more relevant, as these two sources are affected by climate change in different ways (Wade et al., 2022). One way of furthering our understanding of these processes is through harnessing the opportunities brought about by the increasing availability of diverse environmental datasets (Lavallin and Downs, 2021; Blair and Henrys, 2023), and the development of methods and approaches to use these data to gain insight (Yu et al., 2021). This study explored a particular approach to this, using well established methods and a broad range of environmental data, highlighting some of the opportunities and challenges in the approach.

Funding sources

This research was supported by the QMEE CDT, funded by NERC grant number NE/R012229/1. NP is funded by Research England.

CRediT authorship contribution statement

Merry Crowson: Writing – original draft, Visualization, Formal analysis, Data curation, Conceptualization. **Nathalie Pettorelli:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition. **Nick J.B. Isaac:** Writing – review & editing, Supervision, Methodology. **Ken Norris:** Writing – review & editing, Supervision. **Andrew J. Wade:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The code is available at https://github.com/merrycrowson/water_quality.git, and the data is available on the University of Reading Research Data Archive.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2024.176589.

References

Angus, J., Baldock, J., Farrel, M., Norton, R., 2013. Plant-available nitrogen fact sheet. Retrieved from. https://grdc.com.au/_data/assets/pdf_file/0026/126494/grdc-fs-plantavailablenitrogen-pdf.pdf.

Bayliss, A.C., 1999. Catchment descriptors. In: Volume 5 of the Flood Estimation Handbook. Centre for Ecology & Hydrology.

Belgiu, M., Drăgu, L., 2016. Random forest in remote sensing: a review of applications and future directions. ISPRS J. Photogramm. Remote Sens. 114, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011.

Bell, V.A., Naden, P.S., Tipping, E., Davies, H.N., Carnell, E., Davies, J.A.C., Wu, L., 2021. Long term simulations of macronutrients (C, N and P) in UK freshwaters. Sci. Total Environ. 776. https://doi.org/10.1016/j.scitotenv.2021.145813.

Blair, G.S., 2021. Digital twins of the natural environment. Patterns 2 (10), 100359. https://doi.org/10.1016/j.patter.2021.100359.

Blair, G.S., Henrys, P.A., 2023. The role of data science in environmental digital twins: in praise of the arrows. Environmetrics 34 (2), 1–6. https://doi.org/10.1002/env.2789.

Boorman, D., Hollis, J.M., Lilly, A., 1995. Hydrology of Soil Types: A Hydrologically-based Classification of the Soils of United Kingdom. Institute of Hydrology, Wallingford. https://doi.org/10.1029/98GL02804.

Breiman, L., 2001a. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Breiman, L., 2001b. Statistical modeling: the two cultures. Stat. Sci. 16 (3), 199–215. https://doi.org/10.1214/ss/1009213726.

Carpenter, S.R., Stanley, E.H., Vander Zanden, M.J., 2011. State of the world's freshwater ecosystems: physical, chemical, and biological changes. Annu. Rev. Env. Resour. 36, 75–99. https://doi.org/10.1146/annurev-environ-021810-094524.

Crawley, M.J., 2007. Generalized Linear Models. In The R Book, pp. 511–526. https://doi.org/10.1002/9780470515075.ch13.

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

Damanik-Ambarita, M.N., Everaert, G., Forio, M.A.E., Nguyen, T.H.T., Lock, K., Musonge, P.L.S., Goethals, P.L.M., 2016. Generalized linear models to identify key hydromorphological and chemical variables determining the occurrence of macroinvertebrates in the Guayas River Basin (Ecuador). Water (Switzerland) 8 (7). https://doi.org/10.3390/W8070297.

Davies, H., Neal, C., 2004. GIS-based methodologies for assessing nitrate, nitrite and ammonium distributions across a major UK basin, the Humber. Hydrol. Earth Syst. Sci. 8 (4), 823–833. https://doi.org/10.5194/hess-8-823-2004.

Davies, H., Neal, C., 2007. Estimating nutrient concentrations from catchment characteristics across the UK. Hydrol. Earth Syst. Sci. 11 (1), 550–558. https://doi.org/10.5194/hess-11-550-2007.

Davison, P.S., Withers, P.J.A., Lord, E.I., Betson, M.J., Strömqvist, J., 2008. PSYCHIC - a process-based model of phosphorus and sediment mobilisation and delivery within agricultural catchments. Part 1: model description and parameterisation. J. Hydrol. 350 (3–4), 290–302. https://doi.org/10.1016/j.jhydrol.2007.10.036.

Defra, 2014. Water Framework Directive implementation in England and Wales: new and updated standards to protect the water environment. Retrieved from. https://assets.publishing.service.gov.uk/media/5a7f00e740f0b62305b84861/river-basin-planning-standards.pdf.

Defra, 2024a. Chapter 7: crops, agriculture in the United Kingdom. https://www.gov.uk/government/statistics/agriculture-in-the-united-kingdom-2021/chapter-7-crops.

Defra, 2024b. Chapter 8: livestock. https://www.gov.uk/government/statistics/agriculture-in-the-united-kingdom-2021/chapter-8-livestock.

Defra Catchment Explorer, 2023. https://environment.data.gov.uk/catchment-planning/, 19 August 2023.

Eastwood, A., Brooker, R., Irvine, R.J., Artz, R.R.E., Norton, L.R., Bullock, J.M., Pakeman, R.J., 2016. Does nature conservation enhance ecosystem services delivery? Ecosyst. Serv. 17 (2016), 152–162. https://doi.org/10.1016/j.ecoser.2015.12.001.

England Agricultural Census, 2016. 5km [SHAPE geospatial data], scale 1:10000, tiles: GB, updated: 30 June 2016, DEFRA, using: EDINA agcensus digimap service. https://digimap.edina.ac.uk. Downloaded: 2022-12-14 14:01:24.586.

Environment Agency, 2019. Phosphorus and freshwater eutrophication pressure narrative. Retrieved from. https://consult.environment-agency.gov.uk/environment-and-business/challenges-and-choices/user_uploads/phosphorus-pressure-rbmp-2021.pdf.

Environment Agency, 2021. Environment agency water quality data from the water quality archive. https://environment.data.gov.uk/water-quality/view/landing.

Environment Agency, 2023. Urban waste water treatment directive treatment plants dataset. available at: https://www.data.gov.uk/dataset/0f76a1c3-1368-476b-a4df-7ef32bfd9a8b/urban-waste-water-treatment-directive-treatment-plants.

ESRI, 2022. ArcGIS pro 3.1.0. Environmental Systems Research Institute, Redlands, CA.

Foy, R.H., 2007. Variation in the reactive phosphorus concentrations in rivers of northwest Europe with respect to their potential to cause eutrophication. Soil Use Manage. 23 (Suppl. 1), 195–204. https://doi.org/10.1111/j.1475-2743.2007.00111.x.

GRASS Development Team, 2022. Geographic resources analysis support system (GRASS) software, version 8.2.0. Open source geospatial foundation. http://grass.osgeo.org.

Griffin, A., Stewart, L., Young, A., 2019. Revising the BFIHOST catchment descriptor to improve UK flood frequency estimates. Hydrol. Res. 50 (6), 1508–1519. https://doi.org/10.2166/nh.2019.166.

Howden, N.J.K., Burt, T.P., Worrall, F., Whelan, M.J., Bieroza, M., 2010. Nitrate concentrations and fluxes in the river Thames over 140 years (1868-2008): are increases irreversible? Hydrol. Process. 24 (18), 2657–2662. https://doi.org/10.1002/hyp.7835.

Jarvie, H.P., Neal, C., Withers, P.J., 2006. Sewage-effluent phosphorus: a greater risk to river eutrophication than agricultural phosphorus? Sci. Total Environ. 360 (1–3), 246–253. https://doi.org/10.1016/j.scitotenv.2005.08.038. May 1.

Jarvie, H.P., Smith, D.R., Norton, L.R., Edwards, F.K., Bowes, M.J., King, S.M., Bachiller-Jareno, N., 2018. Phosphorus and nitrogen limitation and impairment of headwater streams relative to rivers in Great Britain: a national perspective on eutrophication. Sci. Total Environ. 621 (December), 849–862. https://doi.org/10.1016/j.scitotenv.2017.11.128.

Jasiewicz, J., 2021. r.stream.snap - snap point to modelled stream network. https://grass.osgeo.org/grass82/manuals/addons/r.stream.snap.html.

Johnes, P.J., Heathwaite, A.L., 1997. Modelling the impact of land use change on water quality in agricultural catchments. Hydrol. Process. 11 (3), 269–286. https://doi.org/10.1002/(SICI)1099-1085(19970315)11:3<269::AID-HYP442>3.0.CO;2-K.

Johnes, P., Moss, B., Phillips, G., 1996. The determination of total nitrogen and total phosphorus concentrations in freshwaters from land use, stock headage and population data: testing of a model for use in conservation and water quality management. Freshw. Biol. 36 (2), 451–473. https://doi.org/10.1046/j.1365-2427.1996.00099.x.

Johnes, P.J., Heathwaite, A.L., Spears, B.M., Brownlie, W.J., Elser, J.J., Haygarth, P.M., Withers, P.J.A., 2022. Phospehorus and water quality. Our Phosphorus Future May. https://doi.org/10.13140/RG.2.2.14950.50246.

Langhans, K.E., Schmitt, R.J.P., Chaplin-Kramer, R., Anderson, C.B., Vargas Bolaños, C., Vargas Cabezas, F., Daily, G.C., 2022. Modeling multiple ecosystem services and beneficiaries of riparian reforestation in Costa Rica. Ecosyst. Serv. 57 (October 2021). https://doi.org/10.1016/j.ecoser.2022.101470.

Lavallin, A., Downs, J.A., 2021. Machine learning in geography–past, present, and future. Geogr. Compass 15 (5). https://doi.org/10.1111/gec3.12563.

Lawton, J.H., Brotherton, P.N.M., Brown, V.K., Elphick, C., Fitter, A.H., Forshaw, J., Wynne, G.R., 2010. Making space for nature: a review of England's wildlife sites and ecological network. Report to Defra 1–107. Retrieved from. https://webarchive.nationalarchives.gov.uk/20130402170324/http:/archive.defra.gov.uk/environment/biodiversity/documents/201009space-for-nature.pdf.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2 (3), 18–22. https://CRAN.R-project.org/package=randomForest.

Lucas, T.C.D., 2020. A translucent box: interpretable machine learning in ecology. Ecological monographs 90 (4), 1–17. https://doi.org/10.1002/ecm.1422.

Luna Juncal, M.J., Masino, P., Bertone, E., Stewart, R.A., 2023. Towards nutrient neutrality: a review of agricultural runoff mitigation strategies and the development of a decision-making framework. Sci. Total Environ. 874 (February), 162408. https://doi.org/10.1016/j.scitotenv.2023.162408.

Mainstone, C.P., Parr, W., 2002. Phosphorus in rivers - ecology and management. Sci. Total Environ. 282–283, 25–47. https://doi.org/10.1016/S0048-9697(01)00937-8.

Met Office, 2020. HadUK-Grid Gridded Climate Observations on a 1km grid over the UK, v1.0.2.1 (1862–2019). Centre for Environmental Data Analysis, 21 October 2020. https://doi.org/10.5285/89908dfcb97b4a28976df806b4818639.

Mets, K.D., Armenteras, D., Dávalos, L.M., 2017. Spatial autocorrelation reduces model precision and predictive power in deforestation analyses. Ecosphere 8 (5). https://doi.org/10.1002/ecs2.1824.

Molnar, C., 2023. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd Edition. https://christophm.github.io/interpretable-ml-book/.

[dataset] Moore, R.V., Morris, D.G., Flavin, R.W., 1994. Sub-set of UK Digital 1:50,000 Scale River Centre-Line Network. NERC, Institute of Hydrology, Wallingford.

Moorhouse, H.L., McGowan, S., Taranu, Z.E., Gregory-Eaves, I., Leavitt, P.R., Jones, M. D., Brayshaw, S.A., 2018. Regional versus local drivers of water quality in the Windermere catchment, Lake District, United Kingdom: the dominant influence of wastewater pollution over the past 200 years. Glob. Chang. Biol. 24 (9), 4009–4022. https://doi.org/10.1111/gcb.14299.

Morris, D.G., Flavin, R.W., 1990. A digital terrain model for hydrology. In: Proc 4th International Symposium on Spatial Data Handling. Vol 1 Jul 23-27 Zurich, pp. 250–262.

[dataset] Morris, D.G., Flavin, R.W., 1994. Sub-Set of UK 50 M by 50 M Hydrological Digital Terrain Model Grids. NERC, Institute of Hydrology, Wallingford.

Morton, R.D., Marston, C.G., O'Neil, A.W., Rowland, C.S., 2020. Land Cover Map 2017 (25m Rasterised Land Parcels, GB). NERC Environmental Information Data Centre. https://doi.org/10.5285/499212cd-d64a-43ba-b801-95402e4d4098.

Moss, B., 2008. Water pollution by agriculture. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 363, 659–666, 2008 Feb 12. 10.1098/rstb.2007.2176. PMID: 17666391; PMCID: PMC2610176. https://pubmed.ncbi.nlm.nih.gov/17666391/.

Nisbet, T.R., Andreucci, M.-B., De Vreese, R., Högbom, L., Kay, S., Kelly-Quinn, M., Valatin, G., 2022. Forest Green Infrastructure to Protect Water Quality: A Step-by-Step Guide for Payment Schemes, pp. 105–131. https://doi.org/10.1007/978-981-16-6791-6_8.

Office for National Statistics, 2011. Population density [GeoPackage geospatial data], Scale 1:2500, Tiles: GB, Updated: 27 March 2011, ONS, Using: EDINA Society Digimap Service. https://digimap.edina.ac.uk. Downloaded: 2021-06-24 16:03:58.872.

O'Hare, M.T., Baattrup-Pedersen, A., Baumgarte, I., Freeman, A., Gunn, I.D.M., Lázár, A. N., Sinclair, R., Wade, A.J., Bowes, M.J., 2018. Responses of aquatic plants to eutrophication in rivers: a revised conceptual model. Front. Plant Sci. 9 (April), 1–13. https://doi.org/10.3389/fpls.2018.00451.

Pharaoh, E., Diamond, M., Jarvie, H.P., Ormerod, S.J., Rutt, G., Vaughan, P., 2024. Science of the total environment potential drivers of changing ecological conditions in english and welsh rivers since 1990. Sci. Total Environ. 946 (May), 174369. https://doi.org/10.1016/j.scitotenv.2024.174369.

Prasad, R., Chakraborty, D., 2019. Phosphorus basics: understanding phosphorus forms and their cycling in the soil. Crop production, Alabama A&M & Auburn Universities Extension, ANR 2535, 2535. Retrieved from. https://www.aces.edu/blog/topics/crop-production/understanding-phosphorus-forms-and-their-cycling-in-the-soil/%0Ahttp://files/852/understanding-phosphorus-forms-and-their-cycling-in-the-soil.html.

R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rahman, A.S., Rahman, A., 2020. Application of principal component analysis and cluster analysis in regional flood frequency analysis: a case study in new South Wales, Australia. Water (Switzerland) 12 (3), 1–26. https://doi.org/10.3390/w12030781.

Reaney, S.M., Lane, S.N., Heathwaite, A.L., Dugdale, L.J., 2011. Risk-based modelling of diffuse land use impacts from rural landscapes upon salmonid fry abundance. Ecol. Model. 222 (4), 1016–1029. https://doi.org/10.1016/j.ecolmodel.2010.08.022.

Redhead, J.W., May, L., Oliver, T.H., Hamel, P., Sharp, R., Bullock, J.M., 2018. National scale evaluation of the InVEST nutrient retention model in the United Kingdom. Sci. Total Environ. 610–611, 666–677. https://doi.org/10.1016/j.scitotenv.2017.08.092.

Ross, C.W., Hanan, N.P., Prihodko, L., Anchang, J., Ji, W., Yu, Q., 2021. Woody-biomass projections and drivers of change in sub-Saharan Africa. Nat. Clim. Chang. 11 (5), 449–455. https://doi.org/10.1038/s41558-021-01034-5.

Saarela, M., Jauhiainen, S., 2021. Comparison of feature importance measures as explanations for classification models. SN Appl. Sci. 3 (2), 1–12. https://doi.org/10.1007/s42452-021-04148-9.

Schreiber, S.G., Schreiber, S., Tanna, R.N., Roberts, D.R., Arciszewski, T.J., 2022. Statistical tools for water quality assessment and monitoring in river ecosystems – a scoping review and recommendations for data analysis. Water Qual. Res. J. 57 (1), 40–57. https://doi.org/10.2166/wqrj.2022.028.

Scowen, M., Athanasiadis, I.N., Bullock, J.M., Eigenbrod, F., Willcock, S., 2021. The current and future uses of machine learning in ecosystem service research. Sci. Total Environ. 799, 149263. https://doi.org/10.1016/j.scitotenv.2021.149263.

Shen, L.Q., Amatulli, G., Sethi, T., Raymond, P., Domisch, S., 2020. Estimating nitrogen and phosphorus concentrations in streams and rivers, within a machine learning framework. Sci. Data 7 (1), 1–11. https://doi.org/10.1038/s41597-020-0478-7.

Smith, V.H., Schindler, D.W., 2009. Eutrophication science: where do we go from here? Trends Ecol. Evol. 24 (4), 201–207. https://doi.org/10.1016/j.tree.2008.11.009.

Spake, R., Bellamy, C., Graham, L., Watts, K., Wilson, T., Norton, L., Eigenbrod, F., 2019. A new analytical framework for spatially targeting the management of natural capital. Nat. Sustain. 2 (February). https://doi.org/10.1038/s41893-019-0223-4.

Tate, K.W., Atwill, E.R., Mcdougald, N.K., George, M.R., Tate, K.W., Atwill, E.R., George, M.R., 2003. Spatial and temporal patterns of cattle feces deposition on rangeland. J. Range Manage. 56 (5), 432–438.

The National River Flow Archive, 2023. https://nrfa.ceh.ac.uk/data/search, 21 August 2023.

Tipping, E., Benham, S., Boyle, J.F., Crow, P., Davies, J., Fischer, U., Toberman, H., 2014. Atmospheric deposition of phosphorus to land and freshwater. Environ. Sci. Process. Impacts 16 (7), 1608–1617. https://doi.org/10.1039/c3em00641g.

Tomlinson, S.J., Carnell, E.J., Dore, A.J., Dragosits, U., 2020. Nitrogen Deposition in the UK at 1km Resolution, 1990–2017 NERC Environmental Information Data Centre. https://doi.org/10.5285/9b203324-6b37-4e91-b028-e073b197fb9f.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S-PLUS, Fourth edition. Springer.

Virro, H., Kmoch, A., Vainu, M., Uuemaa, E., 2022. Random forest-based modeling of stream nutrients at national level in a data-scarce region. Sci. Total Environ. 840 (May), 156613. https://doi.org/10.1016/j.scitotenv.2022.156613.

Wade, A.J., Skeffington, R.A., Couture, R.M., Lampa, M.E., Groot, S., Halliday, S.J., Skuras, D., 2022. Land use change to reduce freshwater nitrogen and phosphorus will be effective even with projected climate change. Water (Switzerland) 14 (5). https://doi.org/10.3390/w14050829.

Withers, P.J.A., Neal, C., Jarvie, H.P., Doody, D.G., 2014. Agriculture and eutrophication: where do we go from here? Sustainability (Switzerland) 6 (9), 5853–5875. https://doi.org/10.3390/su6095853.

Yu, Q., Ji, W., Prihodko, L., Ross, C.W., Anchang, J.Y., Hanan, N.P., 2021. Study becomes insight: ecological learning from machine learning. Methods Ecol. Evol. 12 (11), 2117–2128. https://doi.org/10.1111/2041-210X.13686.

Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A., Smith, G.M., 2009. Mixed effects models and extensions in ecology with R. In: Statistics for Biology and Health. Springer, New York, NY, USA, pp. 1–574.