



Towards estimating the sources of river water pollution

IGI ENVIRONMENT TECHNICAL NOTE

Tom Bryden^{1*}, Lisa Neale¹, Aimee Whatling¹, Chris Posser¹, Dan Cornford¹, Paul Harris², Bruce Griffith²

1 IGI Ltd, The Clock House, The Old Stables Business Park, Bideford, Devon, EX39 3QW, UK

2 Rothamsted Research, North Wyke, Okehampton, Devon, EX20 2SB, UK

** Corresponding author: tom@igiltd.com*

Published on 26 October 2026

This study investigates the potential of continuously monitored water quality data and machine learning to identify land use signatures from agricultural fields. High frequency (15-minute) data from the North Wyke Farm Platform were used to characterise the hydrological and water quality responses to rainfall events. Each water quality sensor on the platform is located so that it captures runoff from fields under a single land use only, enabling each event in the dataset to be attributed to either livestock or arable farming. Following feature selection and parameter tuning, a neural network classifier was trained to distinguish between livestock and arable land use and evaluated on independent validation catchments. The model achieved a balanced accuracy of 76% and a ROC AUC of 0.84, demonstrating good discrimination between land use types. These findings show that continuous water quality data contain detectable agricultural runoff signatures and could support future estimation of river pollution sources.

1 INTRODUCTION

River water pollution can harm aquatic ecosystems and limit the recreational and agricultural use of freshwater resources. A wide range of factors can contribute to river water quality, including agricultural runoff, wastewater and sewage overflows, diffuse urban pollution, and natural catchment processes [1]. While river water pollution can often be detected through routine monitoring, identifying the specific sources of pollution remains a major challenge [2]. Traditional pollution source identification methods, such as isotope tracing [3] and targeted investigations [4], can provide robust evidence of contaminant origins, but these approaches are often expensive, labour intensive, and time consuming.

The long-term aim of this work is to support a new method for source identification of river water pollution using continuous water quality sensor data. The method will combine IGI's machine learning based unmixing technology [5] with river catchment modelling approaches. By integrating data driven pattern recognition with process-based environmental understanding, this method has the potential to improve both the speed and accuracy of identifying the sources of river water pollution. The development of this method is at an early stage, and this article focuses on characterising the water quality signatures produced by one potential source of river water quality issues: agricultural runoff. In future applications, it is envisaged that these identified signatures could be used to support the detection and estimation of the contribution of agricultural runoff directly from continuous river sensors.

Agricultural processes are well researched and many modelling and field experiment studies have quantified how changes in land management influence productivity, carbon and nutrient losses [6-8]. Alongside these process-based approaches, machine learning and other data driven methods are

increasingly used to interpret environmental monitoring data, detecting complex, non-linear relationships that are difficult to capture with traditional methods. At Rothamsted Research's North Wyke site, using the open systems data collections from the North Wyke Farm Platform (NWFP) [9], machine learning has similarly seen an increase in use, including studies to predict soil moisture from remotely sensed data [10] and to improve process-based simulations of hydrological extremes [11]. Building on this data driven direction, the present study applies machine learning to the platform's continuous water quality data to characterise agricultural runoff signatures.

2 DATA DESCRIPTION

The NWFP [9], seen in Figure 1, is in southwest England ($50^{\circ}46'10''$ N, $3^{\circ}54'05''$ W) and was established in 2010. It consists of three 21-ha small farms, each of which periodically transition to different farming practices to enable comparisons of sustainable and resilient farming practices. All farms at baseline were under permanent pasture, grazing beef cattle and sheep. Following baseline, the first system-change period started across 2013 to 2015, and the second system-change period started in 2019, while the third is due in 2026. For the first system-change period, 'red' and 'blue' farms were created each with improved pastures ('red' with high sugar grass and 'blue' with a high sugar grass white clover mix). For the second system change period, the 'red' farm transitioned to an arable farm with rotations of wheat, oats and beans. Throughout, the 'green' farm remains under permanent pasture and will continue to do so as a long-term control.

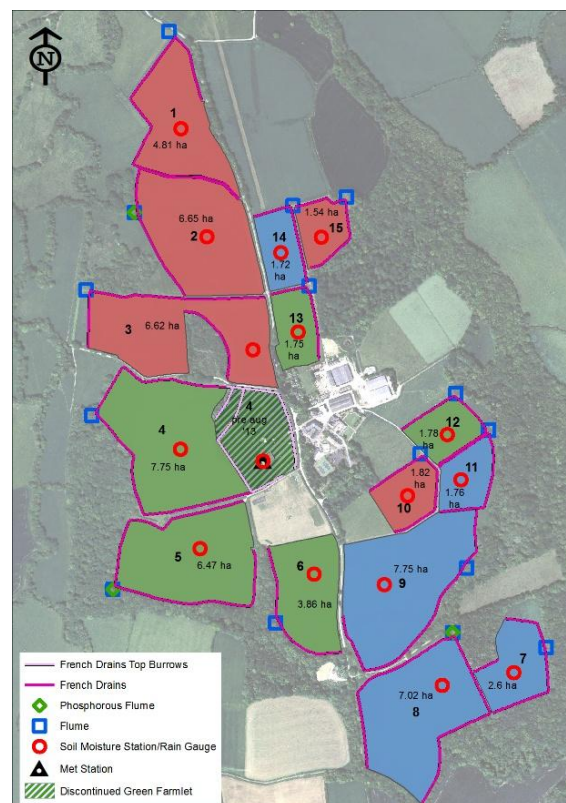


Figure 1. North Wyke Farm Platform with its three small farms (red, blue and green) and sensor locations across 15 catchments. Some catchments are split into two fields [9]

The platform is situated on a ridge at 120-180 m above sea level, where its fields slope to the west to the River Taw and to the east to the Cocktree stream (a tributary of the River Taw). From 1982 to 2023, the average annual precipitation at North Wyke was 1,032 mm together with average minimum and maximum daily temperatures of 6.8 and 13.5 °C, respectively. The platform's soil comprises predominantly two similar

series, Hallsworth (Dystric Gleysol) and Halstow (Gleyic Cambisol) with about 36% clay on the topsoil. Each farm consists of five hydrologically isolated catchments ranging in size from 1.62 to 8.08 ha, where some catchments consist of multiple fields. All farms are heavily monitored and provide high-value systems-scale data on soils, crops, livestock, biodiversity and water/gaseous emissions, that are each contextualised by farm management and weather data. For this study, we used all 15 catchments across all three farms, with 7 years of data from 2017-2023.

To investigate the agricultural signatures, we used continuous (15-minute frequency) water emission (quality) data for the parameters shown in Figure 2. The water being measured is runoff from very specific locations, having a catchment area of just one or two fields, as shown in Figure 1. This allows attribution of the water quality data to whatever was happening in the field(s) at the time. Figure 2 shows data for fields NW003-4 (NWFP catchment 3), which at the times shown on the graph were being exclusively used for arable farmland. There are therefore 105 (15*7) plots like Figure 2 available.

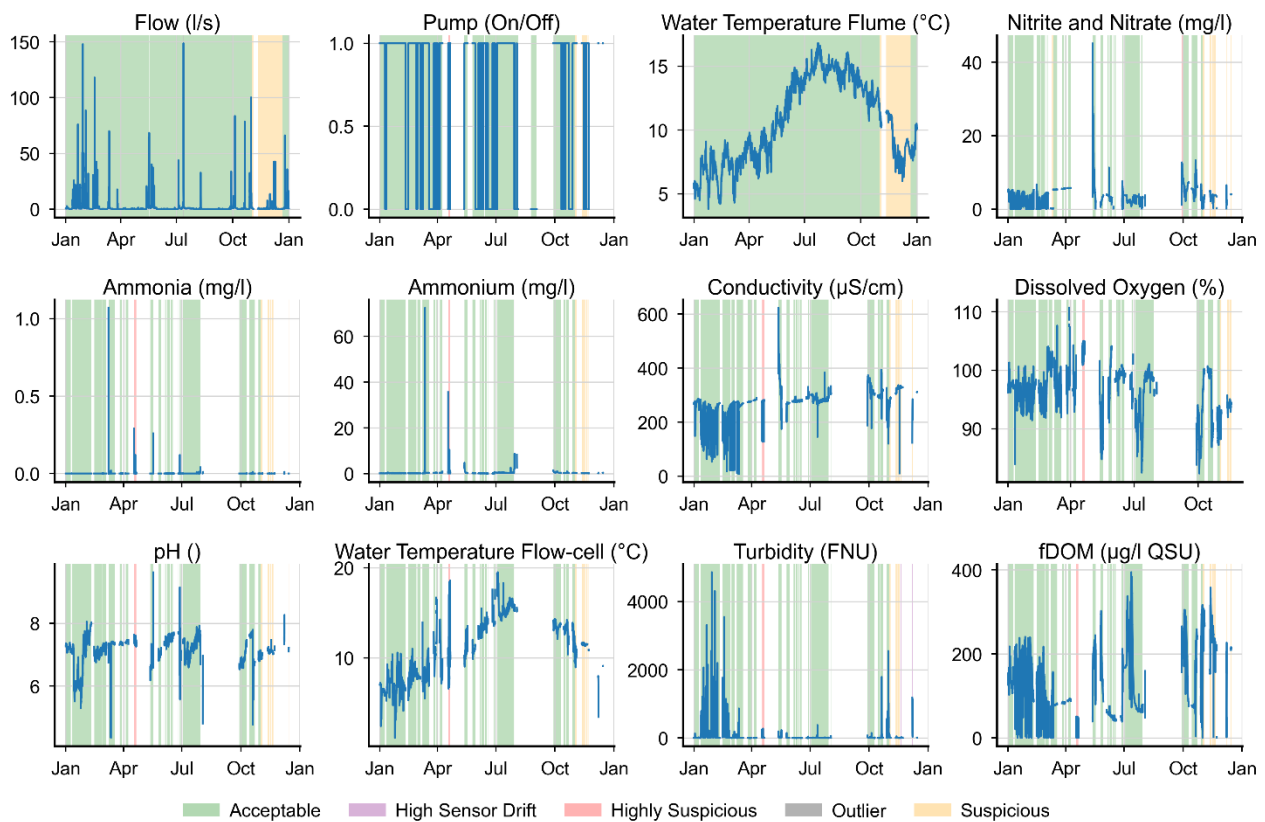


Figure 2. Water quality data for fields NW003-4 (catchment 3 of the red farm) for 2021

As seen in Figure 2, the data does have many temporal gaps, which are associated with periods of limited rainfall. As the data is runoff from only one or two fields, flow can cease entirely during dry conditions, resulting in intermittent records. The NWFP datasets are quality controlled, as seen in Figure 2, with green indicating good data quality and other colours often indicating data quality issues. In this study, only data marked “Acceptable” was used.

3 METHOD

To identify water quality signatures within the study dataset, the problem is formulated as a machine learning classification task. The classifier uses the water quality data to predict the corresponding field condition or activity at a given time. For example, using the water quality data, the classifier may predict

whether a field is being managed for livestock or arable production. Characteristic patterns within the water quality data can then be identified by comparing conditions across these classes. For instance, a particular turbidity response may consistently occur during livestock use but not during arable management, suggesting a potential agricultural runoff signature associated with livestock activity. An overview of the method is shown in Figure 3.

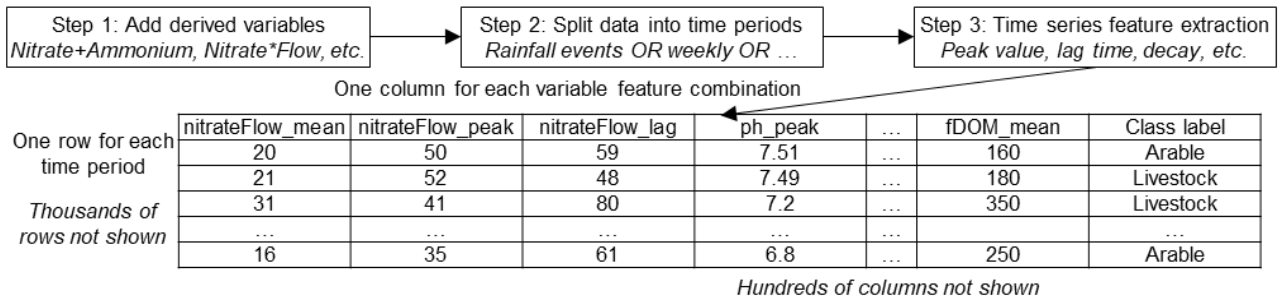


Figure 3. Method for generating the input feature matrix.

3.1 Step 1: Add derived variables

For step 1, there are many derived variables that could be added and this study did not attempt an exhaustive evaluation of all possible combinations. For this work, along with the measured water quality parameters (Figure 2), we multiplied the concentration variables by flow, producing flow weighted variables intended to better represent pollutant transport during runoff events, named [Flux] variables. The application of step 1 resulted in 8 variables:

- Water Temperature Flume (°C)
- Conductivity (µS/cm)
- Dissolved Oxygen (%)
- pH
- Turbidity (FNU)
- [Flux] Nitrite and Nitrate Flux (mg/s)
- [Flux] Ammonium Flux (mg/s)
- [Flux] Fluorescent Dissolved Organic Matter (fDOM) Flux (µg/s QSU)

3.2 Step 2: Split data into time periods

Step 2 involved creating each row of the input feature matrix, with each row representing a single time period. A time period could simply be a fixed duration, such as one week. For this analysis, rainfall events were used as the time period, as these often produce clear responses in the water quality parameters. A rainfall event can be seen in blue in Figure 4 and was defined as:

- Event start: >0.2 mm rain (in 15-minute period)
- Event end: <0.2 mm rain for 6 consecutive hours

These thresholds could be adjusted to be more or less stringent. Reducing the rainfall threshold or dry period duration would increase the number of identified events, while increasing them would reduce the number of events but generally produce more distinct water quality responses. The selected thresholds were chosen to provide a suitable balance between event prominence and the total number of events available for analysis. Using these criteria, 1,184 rainfall events were identified across the 7 years, resulting in a maximum of 17,760 (15 × 1,184) samples (rows) across the 15 farm catchments.

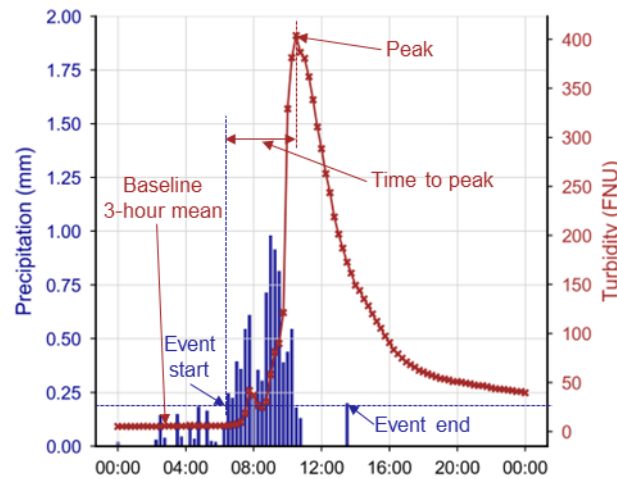


Figure 4. Example rainfall event and turbidity response.

3.3 Step 3: Time series feature extraction

Step 3 involved creating the columns of the input feature matrix. A typical turbidity response to a rainfall event is shown in red in Figure 4. The columns of the input feature matrix consisted of features extracted from these event response curves, examples include: the peak response magnitude, the time from rainfall onset to the turbidity peak, etc. These extracted features summarise the dynamics of each rainfall response in a form suitable for machine learning classification. To ensure accurate feature extraction, all instances where more than 20% of the expected observations were missing from Figure 4 were excluded from the analysis.

As with the derived variables from step 1, this study did not attempt an exhaustive exploration of all possible event features. A very large number of additional features could potentially be extracted, including interactions between different features, such as the time difference between turbidity and nitrate peaks. Instead, this work focused on a smaller set of physically interpretable features considered most relevant to agricultural runoff responses. The extracted feature groups are summarised in Table 1, resulting in a total of 31 extracted feature columns.

Table 1. Features extracted from each rainfall event

4 rainfall (x 1 = 4 columns)	3 parameter (x 8 = 24 columns)	1 flux parameter (x 3 = 3 columns)
• Total in previous 10 days • Total during event • Peak intensity • Event duration	• Baseline mean, 3 h before event • Peak (or trough) • Lag from event start to peak	• Flow normalised event load (area under response curve / total flow during event)

4 ARABLE / LIVESTOCK MACHINE LEARNING CLASSIFICATION

The input feature matrix therefore consisted of 17,760 rows and 31 columns. The next step was to assign a class label column to the input matrix. In this initial experiment, field use [arable or livestock] was used as the class label. Five of the catchments (1-3, 10 and 15) were converted from livestock to arable in 2019 (red farm). For these 5 catchments, the arable class started on 2 October 2019 and continued until the end of the dataset in 2023. The livestock class for the red farm covered the period from the start of the dataset in 2017 until the last day livestock was seen in each of the catchments (various days in 2019). The other 10 catchments were livestock the entire time (green and blue farms).

For this analysis, only events with no missing features were used. After removing these events and any events when the field use is not known, there remained 3,621 rows in the input feature matrix. Before conducting any further work, 2 of the 5 (40%) arable catchments and 4 of the 10 (40%) livestock catchments were removed from the data for future model validation. As validation catchments lie within the NWFP, they test the model's ability to generalise to unseen catchments within the same set of farms rather than to an entirely separate site.

- Learning catchments (farms): 1 (red), 4 (green), 5 (green), 8 (blue), 10 (red), 11 (blue), 13 (green), 14 (blue), 15 (red)
(Livestock : Arable = 1644 : 355 = 82% : 18%)
- Validation catchments (farms): 2 (red), 3 (red), 6 (green), 7 (blue), 9 (blue), 12 (green)
(Livestock : Arable = 1334 : 288 = 82% : 18%)

Using only the learning catchments, backward sequential feature selection was performed to identify the most informative features for predicting field usage, with feature subsets evaluated using the area under the receiver operating characteristic curve (ROC AUC) for the classification models. Two classifiers were evaluated: logistic regression and a multilayer perceptron, implemented using the python library scikit-learn [12]. Default hyperparameters were used for logistic regression. The multilayer perceptron used a single hidden layer with 12 neurons, logistic activation, an L2 regularisation parameter of 1, a maximum of 1000 iterations, and a fixed random seed of 42. The selection of the model architecture was based on 5-fold cross validation on the learning catchments. The results are shown in Table 2.

The columns to take forward are selected based on the neural network results for a tolerance of -0.001. The tolerance defines how much 'information' can be lost by removing the feature from the learning set, the lower the tolerance the more features are suggested to be retained. These 7 features are:

- Rainfall, total in previous 10 days
- Conductivity, baseline
- Turbidity, baseline
- Dissolved Oxygen, baseline
- fDOM flux, peak (or trough)
- Turbidity, peak (or trough)
- Nitrite and Nitrate Flux, flow normalised event load

Table 2. Sequential feature selection results for logistic regression and neural network models over a range of tolerance values.

Model	Tolerance	ROC AUC	Number of features
Logistic Regression	-0.01	0.901	4
	-0.001	0.948	12
	-0.0001	0.949	14
Neural network	-0.01	0.908	4
	-0.001	0.926	7
	-0.0001	0.933	21

To create the classification model, the [IGI Machine Learning Tools](#) [5] are used. These tools enable the exploration of a number of types of classification models with clear visualisations showing the effects of different model parameter selections. The tools also include automated parameter tuning and this was used to evaluate a range of parameter combinations and identify the most suitable configuration for the dataset. The input to this analysis is an 8 column (7 features plus 1 class label), 2,353 row matrix, consisting

only of the Learning catchments. There are slightly more rows (events) than during the sequential feature selection because, with only 7 features, more events have no missing features.

Parameter tuning was conducted using standardised data and a 70:30 train/test split. Several models achieved good predictive performance, with balanced accuracy exceeding 0.8. To minimise model complexity while maintaining high performance, a multilayer perceptron was selected. The final model comprised two hidden layers of 10 neurons with tanh activation function and an L2 regularisation (smoothing) parameter of 0.01. The resulting model performance, visualised using the [IGI Machine Learning Tools](#) [5], is seen in Figure 5a.

The selected neural network was then trained using all learning catchments and evaluated on the independent validation catchments. The model achieved a balanced accuracy of 76%, an overall accuracy of 86%, and a ROC AUC of 0.84, indicating good discrimination between livestock and arable fields. Performance was higher for livestock (precision = 0.90, recall = 0.92, F1 = 0.91) than for arable (precision = 0.64, recall = 0.59, F1 = 0.61), reflecting the class imbalance within the validation dataset. Compared with the learning data, overall accuracy decreased slightly (95% to 86%), suggesting the model generalised well to previously unseen catchments without substantial overfitting. The validation confusion matrix and ROC curve are shown in Figure 5b.

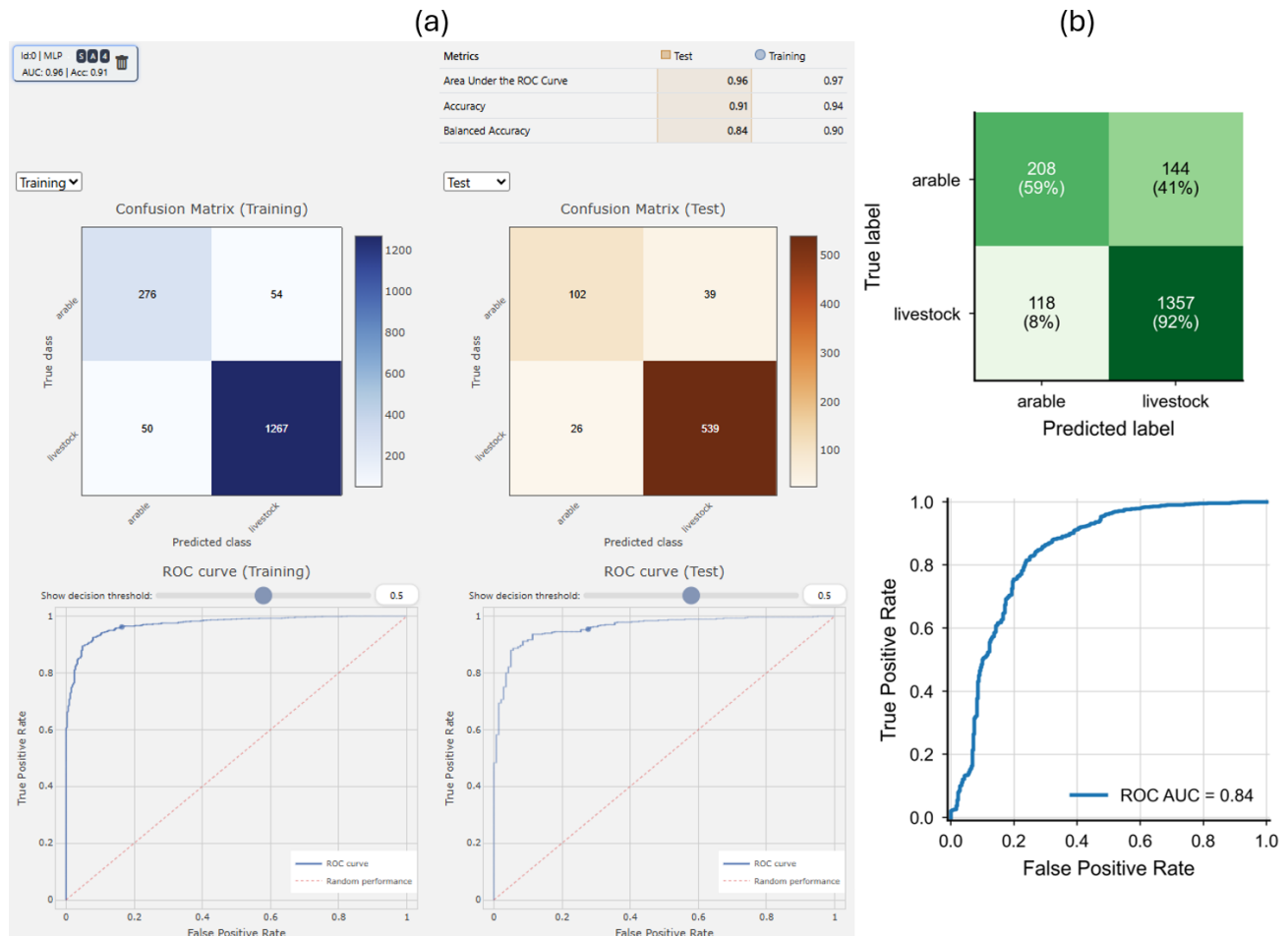


Figure 5. (a) Results from the IGI Machine Learning Tools [5], (b) Validation catchment results

The [IGI Machine Learning Tools](#) [5] also allow the exploration of the model, including the Feature Importance Ranking based on analysis of variance concepts. For the model chosen, the top highest rank features were:

- Conductivity: Baseline mean, 3 hours before event
- Nitrite and Nitrate: Flow normalised event load
- Turbidity: Peak (or trough)

Figure 6 shows the average rainfall event response for these selected parameters and another that was not found to be important (water temperature), separated according to whether the catchment was managed as arable or livestock.

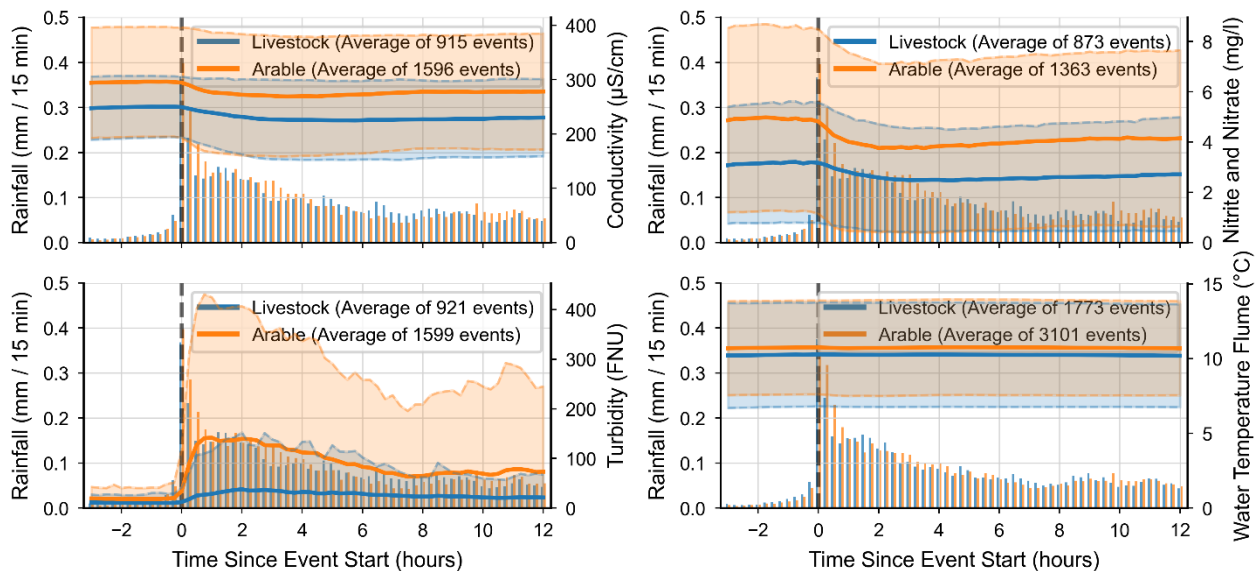


Figure 6. Average responses to rainfall events, mean and standard deviation

Turbidity shows the clearest distinction between the two land uses, with substantially larger event peaks observed under arable management. This is consistent with the expectation that ploughed and cultivated soils are generally more susceptible to erosion and sediment mobilisation during rainfall events, resulting in increased sediment transport within surface runoff.

Conductivity and nitrite and nitrate flux also show consistently higher responses under arable management. This reflects differences in nutrient availability, soil disturbance, and hydrological connectivity associated with arable systems, leading to enhanced transport of dissolved ions and nutrients during runoff events.

For comparison, water temperature is also shown in Figure 6. In contrast to the other parameters, water temperature exhibits very limited separation between arable and livestock management. This is expected, as water temperature is likely to be influenced more strongly by meteorological conditions and seasonal variability than by field management practices.

5 ADDITIONAL CASES

Although a comprehensive sensitivity analysis of all parameters and extracted features is beyond the scope of this study, three additional field use targets were investigated. The first two were classification tasks: distinguishing between livestock fields of high sugar grass pastures (deep rooted) and of permanent pastures (shallow rooted) and distinguishing between livestock fields grazed by sheep and by cattle. The third case, Section 6, used regression rather than classification, testing whether the water quality signals could estimate the number of days since fertiliser was last applied to a field.

For the additional classification cases, the same process as that described in Section 4 was applied, with the only change being the class label assigned to the target. Shallow rooted grass was present in

catchments 4, 5, 6, 12 and 13, and fields within (all green farm), while the remaining catchments, and fields within, were deep rooted (red and blue farms). Only the livestock fields were used for this classification, and two catchments of each type were held back for validation. The results are seen in Table 3, after the sequential feature selection, 11 features were retained for this case. The results indicate only moderate discrimination between pastures with deep rooted and shallow rooted grass. This suggests that any water quality signatures distinguishing grass rooting depth are weak compared with the clearer arable/livestock distinction found in Section 4.

Table 3. Summary of additional classification cases

		Catchments	Split	Network Architecture	Activation function	Alpha	Accuracy	Balanced accuracy	ROC AUC
Grass roots (Deep: Shallow)	Learning	4, 6, 7, 8, 13, 14	747 : 686 52% : 48%	[7, 7]	relu	0.1	0.93	0.93	0.98
	Validation	5, 9, 11, 12	436 : 399 52% : 48%				0.65	0.65	0.73
Livestock (Sheep: Cattle)	Learning	2, 3, 5, 6, 7, 9, 11, 13, 14, 15	440 : 54 89% : 11%	[10]	tanh	1	0.98	0.93	0.99
	Validation	1, 4, 8, 10, 12	187 : 23 89% : 11%				0.85	0.70	0.84

Data were more limited for the cattle/sheep classification because only days when livestock were present in the field were used. For example, if a field was empty and cattle had left the previous day, that data was excluded, whereas in the arable/livestock classification of Section 4 it would have been used and classed as livestock. This stricter criterion was necessary because sheep were sometimes placed in a field the day after cattle had left, in such cases, evidence of the cattle from the previous day would likely persist even though the field was now classed as sheep, adding uncertainty to the classification.

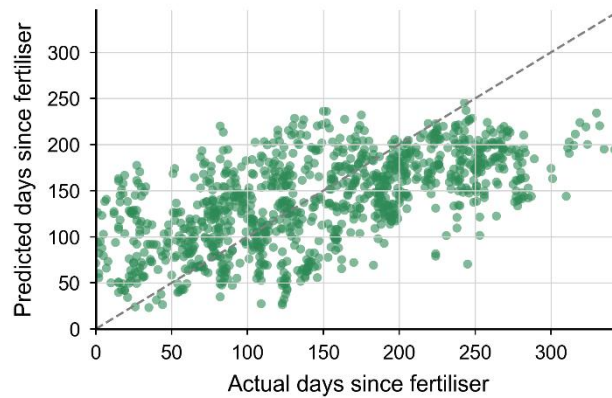
Sequential feature selection retained 17 features for this case, the results, seen in Table 3, indicate moderate discrimination between cattle and sheep. On the surface this is broadly comparable to the arable/livestock model, if somewhat weaker, but the result should be treated with considerable caution because of the limited cattle data. Cattle accounted for only 54 of the 494 learning events and just 23 of the 210 validation events, so the model had very few cattle examples to learn from, and the validation scores rest on a small enough sample that a handful of misclassifications would shift them substantially. The sharp fall in performance from the learning to the validation data, balanced accuracy dropping from 0.93 to 0.70 and ROC AUC from 0.99 to 0.84, is also consistent with the model having overfitted to the few cattle examples available. The cattle/sheep distinction therefore warrants further data collection before any firm conclusion can be drawn.

6 REGRESSION

For the regression case, the method described in Section 4 was applied with two modifications appropriate for a continuous target. First, the Sequential Feature Selector used ridge regression with default parameters in place of logistic regression. Second, the classifier was replaced by a multilayer perceptron regressor, comprising a single hidden layer of 6 neurons with tanh activation, an L2 regularisation parameter of 1, a maximum of 10,000 iterations, and a fixed random seed of 42. At -0.001 tolerance, the Sequential Feature Selector retained 28 of the 31 features; because dropping only three of the 31 features offered little benefit in parsimony or computational cost, all 31 were retained. The results are seen in Table 4 and Figure 7.

Table 4. Summary of additional regression case

		Catchments	Split	Network Architecture	Activation function	Alpha	R ²	RMSE	MAE
Fertiliser (regression)	Learning	2, 4, 5, 6, 8, 9, 10, 12, 13, 14	2021	[12]	logistic	10	0.52	49	38
	Validation	3, 7, 11, 15	1075				0.36	61	50

**Figure 7.** Predicting the days since fertiliser applied on the validation data.

These results are appreciably weaker than the classification cases in Section 4. A validation R^2 of 0.36 indicates that the model accounts for only around a third of the variance in the target, and the drop from the learning R^2 of 0.52 points to limited generalisation to unseen catchments rather than a robust underlying relationship. The mean absolute error of roughly 50 days is likely too large for the estimate to be useful in practice. This weak performance is perhaps unexpected, as fertiliser application might be expected to leave a nutrient signature in runoff. In practice, that signal is likely obscured by variation in fertiliser type and rate, uptake and retention in the soil, rainfall timing, and the dilution and transformation of nutrients before they reach the sensor. As with the cattle/sheep case, the fertiliser regression is best regarded as an exploratory result that motivates further data collection and analysis rather than a validated predictive model.

7 DISCUSSION

The results demonstrate that continuous water quality monitoring contains measurable signals related to agricultural land use. Although the classification performance was moderate rather than definitive, the ability to distinguish between livestock and arable management using rainfall event responses suggests that characteristic water quality signatures are present within the dataset. In particular, turbidity, conductivity, and nutrient related responses showed consistent differences between the two land use classes, supporting existing understanding of sediment and nutrient mobilisation under different agricultural practices.

The three additional cases were less successful. Neither the grass rooting depth classification nor the fertiliser regression achieved performance strong enough to be considered reliable, and the cattle/sheep result, while moderate, rested on very limited data. These cases should therefore be regarded as exploratory and provide useful direction for future investigation as more data becomes available.

A key limitation of this study is that only a limited subset of possible derived variables, event definitions, and extracted features were explored. A more comprehensive sensitivity analysis could improve model performance by systematically investigating alternative rainfall event thresholds, additional feature

extraction approaches, and parameter interactions. In addition, validation was performed on catchments held back from the same three farms, performance on hydrologically and geologically distinct catchments elsewhere remains to be tested.

Despite these limitations, the study demonstrates the potential for combining continuous water quality sensing with machine learning classification to estimate likely sources of river water pollution. In the future, extending this approach to include additional sources, such as wastewater inputs or urban runoff, could support near real time pollution source attribution within river catchments. This has the potential to provide faster and more scalable identification of water quality pollution sources compared to traditional approaches.

8 References

1. Environment Agency (2024) World Rivers Day: what are the biggest causes of river pollution and what's being done about them?
<https://environmentagency.blog.gov.uk/2024/09/22/world-rivers-day-what-are-the-biggest-causes-of-river-pollution-and-whats-being-done-about-them/>
2. Shiferaw N, Kim J, Seo D (2023) Identification of pollutant sources and evaluation of water quality improvement alternatives of a large river. *Environmental Science and Pollution Research*, 30(11), 31546–31560.
<https://doi.org/10.1007/s11356-022-24431-5>
3. Yue F-J, Li S-L, Liu C-Q, Zhao Z-Q, Ding H (2017) Tracing nitrate sources with dual isotopes and long term monitoring of nitrogen species in the Yellow River, China. *Scientific Reports*, 7, 8537.
<https://doi.org/10.1038/s41598-017-08756-7>
4. Environment Agency (2026) Yorkshire Water fined £733k after polluting country park stream.
<https://www.gov.uk/government/news/yorkshire-water-fined-733k-after-polluting-country-park-stream>
5. IGI Ltd (2026) Machine Learning.
<https://igilttd.com/energy/software/machine-learning>
6. Hassall KL, Coleman K, Dixit PN, Granger SJ, Zhang Y, Sharp RT, Wu L, Whitmore AP, Richter GM, Collins AL, Milne AE (2022) Exploring the effects of land management change on productivity, carbon and nutrient balance: application of an ensemble modelling approach to the upper River Taw observatory, UK. *Science of the Total Environment*, 824, 153824.
<https://doi.org/10.1016/j.scitotenv.2022.153824>
7. Irisarri JGN, Pembleton KG, Lai Y, Griffith B, Shield I, Le Grice P, Harris P (2026) Evaluating the transition of marginal grasslands to arable systems for human-protein production through an in silico long-term experiment. *Journal of Agriculture and Food Research*, 28, 102980.
<https://doi.org/10.1016/j.jafr.2026.102980>
8. Guo Y, Collins AL, Zhang Y, Carswell A, Segura C, Mead A, Olde L, Loick N, Cardenas L (2026) Conversion from improved pasture to arable exacerbates unintended environmental consequences: evidence from the North Wyke Farm Platform, UK. *Agriculture, Ecosystems & Environment*, 408, 110480.
<https://doi.org/10.1016/j.agee.2026.110480>
9. Rothamsted Research (2026) North Wyke Farm Platform.
<https://nw-farmplatform.rothamsted.ac.uk/>
10. Oulaid B, Harris P, Maas E, Fakeye IA, Baker C (2025) Geographically weighted quantile machine learning for probabilistic soil moisture prediction from spatially resolved remote sensing. *Remote Sensing*, 17(16), 2907.
<https://doi.org/10.3390/rs17162907>
11. Curceac S, Atkinson PM, Milne AE, Wu L, Harris P (2020) Adjusting for conditional bias in process model simulations of hydrological extremes: an experiment using the North Wyke Farm Platform. *Frontiers in Artificial Intelligence*, 3, 565859.
<https://doi.org/10.3389/frai.2020.565859>
12. scikit-learn developers (2026) scikit-learn: Machine Learning in Python.
<https://scikit-learn.org/>