A machine learning approach to driver attribution of dissolved organic matter dynamics in two contrasting freshwater systems

Daniel Mercado-Bettín[1], Ricardo Paíz[2,3], Valerie McCarthy[2], Eleanor Jennings[3], Elvira de Eyto[4], Angeles M. Gallegos[5], Mary Dillane[4], Juan C. Garcia[5], José J. Rodríguez[5], and Rafael Marcé[1]

[1]Centre for Advanced Studies of Blanes, Spanish National Research Council, Carrer Accés Cala Sant Francesc, 14, 17300, Blanes, Spain
[2]School of History and Geography, Dublin City University, D09 YT18, Dublin 9, Ireland
[3]Centre for Freshwater and Environmental Studies, Dundalk Institute of Technology, A91 K584 Dundalk, Co. Louth, Ireland
[4]Fisheries & Ecosystem Advisory Services, Marine Institute, F28 PF65 Newport, Co. Mayo, Ireland
[5]Ens d'Abastament d'Aigua Ter-Llobregat, Ctra. Aigües, 6, 08440, Cardedeu, Spain

Correspondence: Daniel Mercado-Bettín (daniel.mercado@ceab.csic.es)

Abstract. Predicting water quality variables in lakes is critical for effective ecosystem management under climatic and human pressures. Dissolved organic matter (DOM) serves as an energy source for aquatic ecosystems and plays a key role in their biogeochemical cycles. However, predicting DOM is challenging due to complex interactions between multiple potential drivers in the aquatic environment and its surrounding terrestrial landscape. This study establishes an open and scalable workflow to identify potential drivers and predict fluorescent DOM (fDOM) in the surface layer of lakes by exploring the use of supervised machine learning models, including random forest, boosting methods, k-nearest neighbors, support vector regression and linear model. The workflow was validated in two contrasting systems: one natural lake in Ireland with a relatively undisturbed catchment, and one reservoir in Spain with a more human-influenced catchment. A total of 24 potential drivers were obtained from global reanalysis data, and lake and river process-based modelling. SHapley Additive exPlanations (SHAP) were conducted for the most influential drivers identified, with soil moisture, soil temperature, and Julian day being common to both study sites. The best prediction was obtained when using the CatBoost model (during hold-out testing period, Irish site: KGE > 0.68, $r^2 > 0.50$; Spanish site: KGE > 0.66, $r^2 > 0.54$). Interestingly, when only using drivers from globally accessible climate and soil reanalysis data plus Julian day, the prediction capacity was maintained at both sites, showcasing potential for scalability. Our findings highlight the complex interplay of environmental drivers and processes that govern DOM dynamics in lakes, and contribute to the modelling of carbon cycling in aquatic ecosystems.

1 Introduction

Lakes are an essential component of global biogeochemical cycles, sustain biodiversity, and provide critical ecosystem services, e.g., water supply, fishing and irrigation. However, their water quality is increasingly at risk due to climatic change and human pressures (Bhateria and Jain, 2016). A key water quality variable is dissolved organic matter (DOM), which influences light penetration, energy, oxygen dynamics and nutrient availability in any lake (Solomon et al., 2015). The dynamics of DOM in lakes are driven by both external processes in the terrestrial environment and internal processes. Land cover, climate, and topography regulate (either increase or decrease) carbon production in the catchment and carbon inputs into the lake (Li et al., 2015). In the water body, the quantity and quality of DOM are controlled by physical and biogeochemical mechanisms such as photodegradation, microbial processing and mixing dynamics, but can also be impacted by water abstraction or dam regulation (Xenopoulos et al., 2021).

Increases in the concentrations of DOM can affect ecosystem stability and human water use (e.g., raw drinking water quality) by reducing oxygen levels, altering microbial communities and nutrient cycling (Lake et al., 2000). DOM is also a precursor to disinfection byproducts (DBPs) during drinking water treatment, substances which have negative human health implications (Li et al., 2014). Understanding the dynamics of DOM in lakes is essential for water quality management, especially as climate-driven processes are expected to increasingly influence DOM in freshwater systems (Creed et al., 2018). Moreover, the occurrence of extreme events such as eutrophication, algal blooms and hypoxic events, for which levels of DOM play a key role, is also expected to increase (Gobler, 2020). Hence, predicting DOM in lake water can improve water quality mitigation protocols and support adaptive water use management strategies.

Predicting DOM dynamics remains a challenge as it results from complex interactions in the environment, including multiple biogeochemical processes (Weyhenmeyer and Karlsson, 2009). Modelling tools offer an approach to simulate DOM in lake water. Process-based models have traditionally been used to better understand lake water quality, including DOM dynamics (McCullough et al., 2018). However, they generally require a large number of model parameters and governing equations, i.e., extensive parameterisation, to represent these dynamics. On the other hand, machine learning (ML) models do not rely on parameter calibration but instead incorporate large amounts of driver variables and data. This functionality can leverage the increasing amount of data being collected through satellite imagery, high frequency monitoring, and global climate and environmental modelling initiatives (Müller et al., 2024; Toming et al., 2020; Asadollah et al., 2025).

ML models have emerged as potential tools for modelling complex environmental variables, including those related to hydrology (Nearing et al., 2021) and water quality (Hanson et al., 2020). They have been recently employed in environmental applications, showing good predictive capabilities due to their ability to handle high-dimensional data, and capture nonlinear relationships (Li et al., 2016), for a diversity of parameters in lakes such as chlorophyll-a (Chen et al., 2024), turbidity (Zhang et al., 2021), and nutrient concentrations (e.g., phosphorus) (Hanson et al., 2020), suggesting potential for predicting DOM (Herzsprung et al., 2020).

This study introduces a workflow for predicting fluorescent DOM (fDOM) (a proxy for DOM) in lakes using a suite of supervised ML models driven by potential drivers either extracted from reanalysis data (climate and soil variables) or outputs from lake and catchment process-based models. The workflow was tested in two different study sites, one in Ireland and one in Spain, that represent contrasting settings for both the potential drivers and DOM dynamics. Model performance was first evaluated at each site using the most influential drivers to predict fDOM. Subsequently, a second simulation was performed using a subset of these drivers, limited to those sourced from reanalysis data, to evaluate the predictive capacity of the model in the context of higher scalability. A primary intention of the study is to use ML models to produce a robust, accessible, and reproducible workflow, rather than to conduct an extensive model benchmark comparison. The aims of this study are: (1) to find the most influential drivers (features) in predicting fDOM, and analyse their importance in two contrasting sites; (2) to test how accurately can supervised ML models predict lake fDOM driven by reanalysis-based data, and hydrologic and lake modeling outputs.

2 Materials and methods

2.1 Study sites

Lough Feeagh and Sau Reservoir are located in western Ireland (53° 56’ N, 9° 35’ W) and northeastern Spain (41° 58’ N, 2° 23’ E), respectively (Fig. 1). The study sites have contrasting attributes. Feeagh (depth of 46.8 m and area of 3.95 km^2) is a monomictic and oligotrophic lake surrounded by a relatively undisturbed landscape, while Sau (depth of 70 m and area of 5.8 km^2) is an eutrophic system subjected to human activities and water abstraction. Feeagh has two primary inflows, the Black and the Glenamong rivers, while Sau has one, the Ter river. The catchment of Feeagh is relatively small (84 km^2), with mid-range hills, and dominated by peatland. The catchment of Sau, in contrast, is larger (1525 km^2), with a varying topography and land uses (Fig.1; Fig. A1).

The dynamics of DOM in both study sites have been previously explored in Ryder et al. (2014) which identified that natural dynamics related to soil temperature, river discharge and drought were important drivers in Feeagh, and in Marcé et al. (2021), which showed human activities (e.g., wastewater effluents and agricultural runoff) were important for Sau, in addition to its environmental dynamics. Catchment hydrology is key for carbon transport into both study sites, and contributes to a distinct seasonality related to climate. Feeagh has a wet temperate climate, with cooler air temperatures and higher rainfall levels that occur on more than 75% of days in the year. The variability of carbon inputs reflects the sensitivity of a peatland-dominated landscape, which exacerbates climate-induced carbon release from the catchment into the lake. In contrast, Sau has a Mediterranean climate, characterised by hot, dry summers and mild, wet winters, dictating water availability, thermal stratification, and organic matter fluxes in the reservoir.

2.2 Prediction workflow

A five-step workflow was implemented to predict fDOM at each site (Fig. 2). First, all potential drivers for fDOM were collected as input data for each site. Second, an exploratory data analysis of fDOM was implemented and the ML models were trained using 85% of the available time series data (1978 out of 2328 fDOM measurements for Feeagh, and 653 out of 769 for Sau), while the remaining 15% (350 out of 2328 for Feeagh, and 116 out of 769 for Sau) future time series (hold-out period) was reserved for independent testing to evaluate performance and potential overfitting by comparing with the training period.

Third, a set of drivers was selected for each site based on the variable importance extracted from the ML models, retaining only those drivers that exceeded an importance threshold of 5%. SHAP and partial dependence analyses were applied to these specific drivers to evaluate how fDOM predictions at each site varied as a function of individually changing the selected input variables. Fourth, we ran simulations using (i) the drivers with a higher variable importance (> 5%), and (ii) using only drivers extracted from globally accessible reanalysis data, and assessed model.

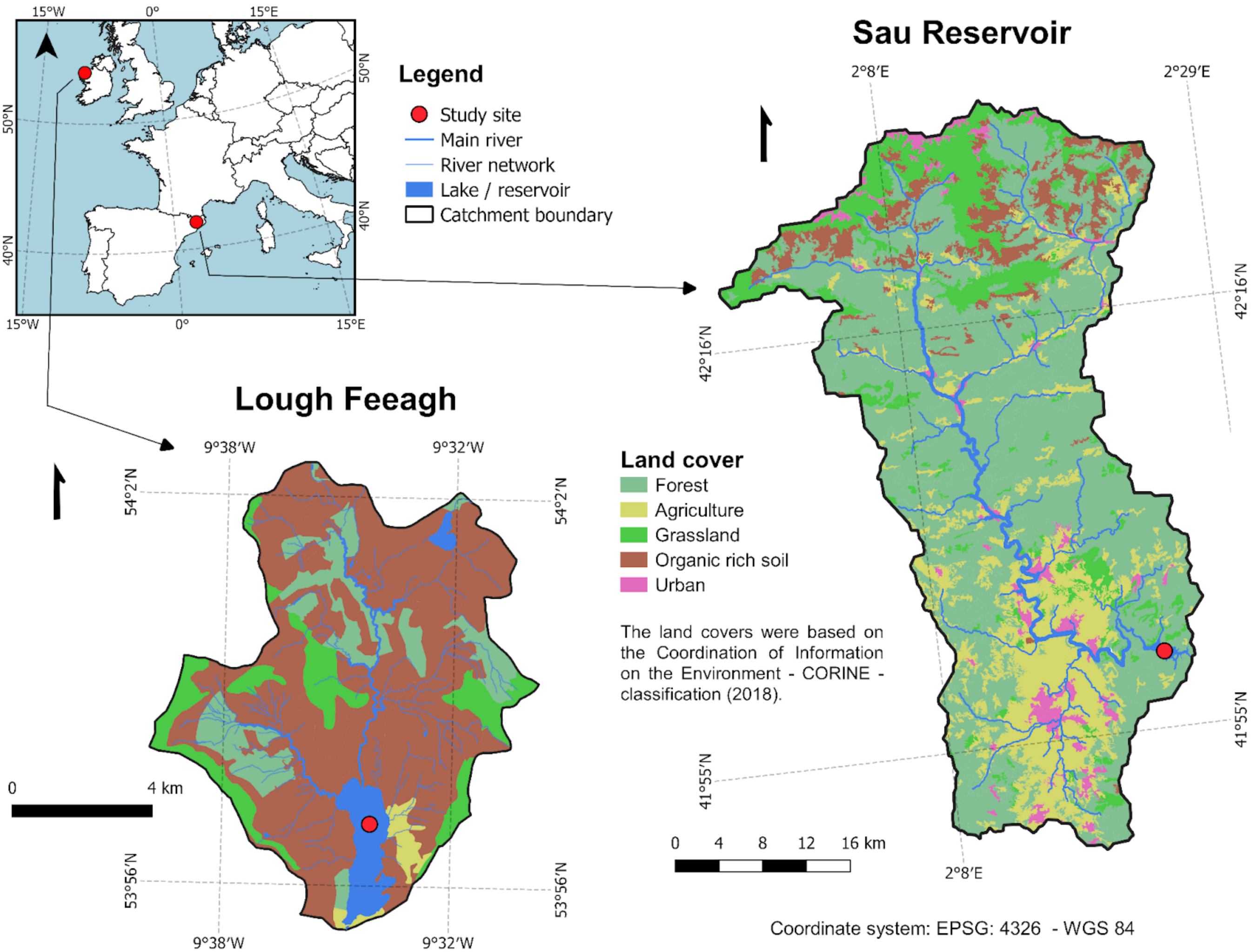

Figure 1. Two contrasting freshwater ecosystems: Lough Feeagh (Ireland) and Sau Reservoir (Spain). The figure shows their locations and catchment land cover. Lough Feeagh is a humic oligotrophic lake with a peatland-dominated catchment and temperate oceanic climate, whereas Sau is a eutrophic reservoir with a human-influenced catchment and Mediterranean climate. The CORINE Land Cover 2018 https://doi.org/10.2909/960998c1-1870-4e82-8051-6485205ebbac classes were grouped into the categories shown for visual clarity, following the aggregation approach described in the Supplementary Information (Figure A2).

The same workflow was applied to both sites using the same data sources, allowing for comparison. Following the FAIR principles, all data and workflow scripts are available and fully reproducible here: https://doi.org/10.5281/zenodo.19354955. Large language models were used in this study to optimise the codes, improve the final plots and, for basic proofreading of the text.

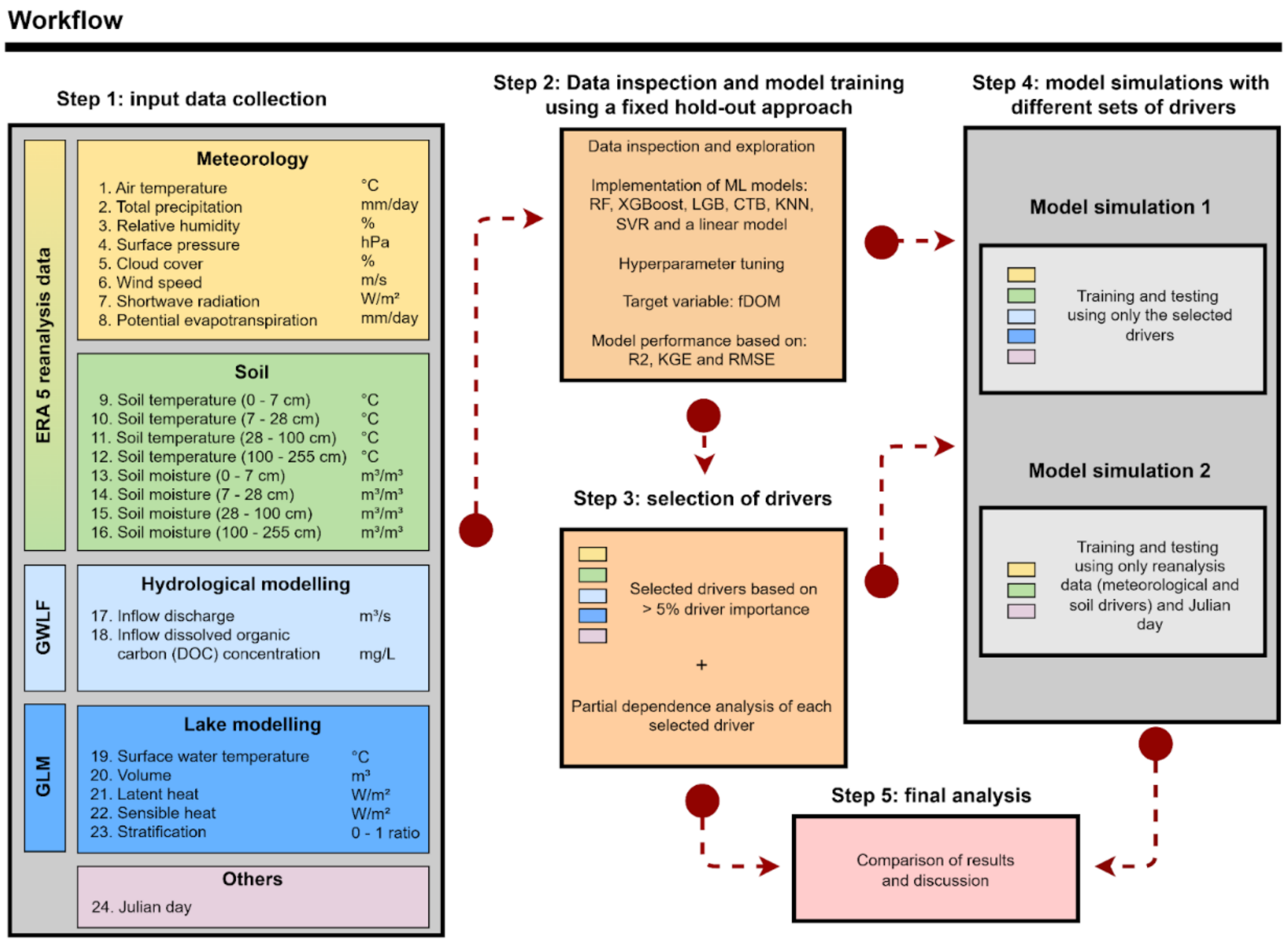

Figure 2. The workflow includes five steps: (1) compilation of input variables including meteorology (yellow) and soil (green) reanalysis data from ERA5 (the fifth-generation atmospheric reanalysis product produced by ECMWF), The Generalised Watershed Loading Functions Model (GWLF) outputs (light blue), The General Lake Model (GLM) outputs (blue), and Julian day (light magenta); (2) data inspection and model training using a split between training (85%) and testing (15%) periods; (3) selection of influential drivers based on feature-importance metrics; (4) model simulations using selected drivers and a reduced subset based on globally available reanalysis variables and Julian day; and (5) comparison of model outputs. DOC stands for Dissolved Organic Carbon. The workflow is available at: https://github.com/danielmerbet/driver_attribution_fdom

2.3 Data

2.3.1 Target variable (fDOM)

Daily surface fDOM values were obtained from high-frequency data for both sites (2-minute-resolution data averaged to obtain daily values before analysis). For Feeagh, data were measured at 0.9 m depth, and for Sau, an average value was calculated between the measurements for depths 0-5 m. fDOM data were expressed as quinine sulfate units (QSU) for the analysis. In Feeagh, the data spanned from 1st of May 2012 to 19th of November 2019 (n = 2328), and for Sau from 4th of February 2017 to 2nd of March 2020 (n = 769), with some gaps. All the other data (i.e., driver data) used in the workflow of this study were constrained by the availability of fDOM data. Following exploratory data analysis (Fig. A3), no transformation was applied to the fDOM data prior to training the ML models in the main manuscript, results for Sau are provided with a log10 transformation in the Supplementary Information for comparison (this is revisited in Results), given the moderate skewness of fDOM at this study site.

In Feeagh, fDOM data were collected using a Seapoint UV fluorometer sensor (Seapoint Sensors Inc., Exeter, NH, USA) and water temperature data were measured using a Hach Environmental Hydrolab Data Sonde X5 (UK OTT Hydrometry Ltd). In Sau, fDOM and water temperature data were collected using a fDOM Digital Smart Sensor and Multiparameter Sonde (YSI EXO sonde, Yellow Springs, OH, USA), respectively. Raw fDOM data were water temperature-corrected in both sites based on relationships established for each sensor (Ryder et al., 2012). Details about the fDOM corrections can be found in Supplementary Information (Fig. A4, A5 and A6).

2.3.2 Driver data

All input data variables, including their respective units and source, are displayed in Step 1 of Figure 2. These were grouped into five categories: (1) meteorology, (2) soil, (3) process-based hydrological modelling, (4) process-based lake modelling and (5) Julian day. All input data variables, including their respective units and source, are displayed in Step 1 of Figure 2. Daily values of meteorology and soil variables were extracted from the European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis v5 (ERA5) dataset (Hersbach et al., Accessed on July 2025). This gridded dataset provides (pseudo) observations at a global scale with a spatial resolution of 0.25°. Eight meteorological variables (traditionally employed in water modelling studies) and soil temperature and soil moisture data at four depths (0-7 cm, 7-28 cm, 28-100 cm, and 100-255 cm) were extracted for the grid-cell which contained each water body.

Daily values of inflow discharge and inflow DOC concentration into each site were generated using the Generalised Watershed Loading Functions Model (GWLF) coupled with a DOC module (GWLF-DOC). The GWLF-DOC version and calibration strategy applied are described in Paíz et al. (2025a). Calibration results can be found in the Supplementary Information. Daily values of five key lake variables (see Step 1, Fig. 2) were obtained from the General Lake Model (GLM) (Hipsey et al., 2019) run for both sites. Calibration strategies applied are described in Mercado-Bettín et al. (2021); Paíz et al. (2025b). Calibration results can be found in the Appendix. In addition, the cosine (to avoid an abrupt numerical change at every start of a year) of Julian day was included in the driver data inputs, given that seasonality is expected to influence DOM predictions; the approximate calendar timing corresponds to the cosine of the Julian day: values close to 1 correspond to boreal winter, 0 to spring/autumn, and -1 to boreal summer.

2.4 Supervised Machine Learning

Supervised ML models have advantages and limitations for time series prediction. In addition to capturing non-linear relationships typical in aquatic systems and water quality predictions (Hollister et al., 2016; Regier et al., 2023), they provide flexibility to assess multiple drivers, temporal indicators, and variables external to the system (Qi, 2012; Rodriguez-Galiano et al., 2015). ML models do not require a fixed set of drivers to predict fDOM effectively, unlike process-based models, which typically rely on predefined inputs. Additionally, there is no need for parameter calibration but hyperparameter tuning, simplifying the modelling process. While some may argue that the lack of parameterization suggests a "black box" approach, supervised ML can provide insights into the potential drivers for predicting a target variable (Biau and Scornet, 2016; Molnar, 2020).

However, due to the intrinsic autocorrelation in time series, e.g., when predicting DOM, these models tend to overfit when using out-of-bag samples during training. To overcome this issue, robust validation and training are required. Here, we used a hold-out period for validation during testing at both study sites. Prior to selecting this method, we compared it with two alternative validation methods using random forest: (1) 5-fold cross-validation and (2) rolling window cross-validation with a two-year training period, a one-year testing period, and a window shift every 90 days (see supplementary Figure A7).

Supervised machine-learning models were applied to predict fDOM dynamics, including Random Forest (RF) (Breiman, 2001), eXtreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016), Light Gradient Boosting (LGB) (Ke et al., 2017), CatBoost (CTB) (Prokhorenkova et al., 2019), k-Nearest Neighbors (KNN) (Fix, 1985), Support Vector Regression (SVR) (Cortes and Vapnik, 1995) and linear model. The three gradient-boosting frameworks were included to assess the robustness of results across closely related implementations. Given their similarities, the main results shown are focused on the highest-performing model among the three. Neural networks were not included because they typically require more complex architecture design and calibration, which can increase methodological complexity and may reduce reproducibility of the workflow. In contrast, the selected models allow more reproducible implementation due to their comparatively simpler hyperparameter tuning.

2.4.1 Hyperparameter tuning

Hyperparameter tuning was implemented in all ML models to improve accuracy and generalization, and reduce the risk of overfitting. It was conducted in R using the caret framework where possible: for RF $mtry \in \{2,4,6\}$; for XGBoost $nrounds \in \{1000, 2000\}$, $max\_depth \in \{3,6\}$, $eta \in \{0.01, 0.05, 0.1\}$, $gamma = 0$, $colsample\_bytree = 1$, $min\_child\_weight = 1$, and $subsample = 1$; for kNN $k \in \{3,5,7\}$; for SVR $C \in \{0.1, 1, 10\}$, and $sigma \in \{0.01, 0.1\}$. Additional model-specific implementations were made: for LGB $learning\_rate \in \{0.01, 0.05, 0.1\}$, $num\_leaves \in \{15, 31, 63\}$, $max\_depth \in \{5, 10, 1\}$, $feature\_fraction \in \{0.8, 1\}$, $bagging\_fraction \in \{0.8, 1\}$, $nrounds = 2000$ (fixed); and for CTB $iterations \in \{300, 500, 1000\}$, $depth \in \{4, 6, 8\}$, $learning\_rate \in \{0.01, 0.05, 0.1\}$, and $l2\_leaf\_reg \in \{1, 3, 5\}$. Models were trained on the training subset and evaluated on the hold-out test set using RMSE. The parameter configuration minimising test RMSE was selected. The hyperparameter tuning can be reproduced by using the code "2_hyperparameter_tuning.R" in the repository.

For each case study, the dataset was chronologically split into training (85%) and hold-out test (15%) subsets. The final 15% of observations were reserved as an independent test set to evaluate predictive performance, mimicking a forecasting exercise. Hyperparameter tuning was conducted exclusively on the training subset. Within the training data, model selection was performed using 5-fold cross-validation. The data were randomly partitioned into five equally sized folds; four folds were used for model fitting and one for validation, iteratively. The average cross-validated performance was used to select the optimal hyperparameter configuration.

2.4.2 Performance metrics

To evaluate and facilitate inter-model comparison, three complementary performance metrics were computed for both training and test datasets: Coefficient of determination (R^2), Root Mean Square Error (RMSE), and Kling–Gupta Efficiency (KGE). These metrics were selected because they capture complementary aspects of model performance such as correlation strength (R^2), error magnitude (RMSE), and combined accuracy in correlation, bias, and variability (KGE). In addition, they are widely used in water-quality modelling of aquatic ecosystems, allowing comparison of the results with other studies.

2.5 Driver importance and SHAP analysis

To assess the influence of the most important drivers, we included SHapley Additive exPlanations (SHAP) plots, using the best boosting method. The SHAP plots measures how much a single driver (feature) value contributes to moving the prediction away from the average value, the Y-axis has the input drivers ranked by overall importance (from top to bottom), X-axis has the SHAP value representing the impact on model output for a single prediction, each point is a single data instance and the color reflects the driver value (blue = low, red = high). In addition, in supplementary information, partial dependence plots were generated to support and illustrate the individual influence of each driver on fDOM predictions by varying the driver's values across its entire range while keeping all other drivers constant in their average value. These partial dependence plots were generated using the outputs from the Random Forest model. To implement SHAP analysis and partial dependence plots, the shap package in Python and the pdp package in R were used, respectively.

3 Results

3.1 Driver attribution

The most influential predictors of fDOM at each site were identified from all 24 potential drivers based on the 5% threshold of the variable importance extracted from the CTB model. This resulted in seven influential drivers being identified for Feeagh and five for Sau (Figure 3), four of which were common to both sites. This selection was consistent across multiple machine-learning models (Fig. A8).

The variables for the deepest soil layer were relevant for both study sites. Soil temperature and soil moisture at 100-255 cm, were the most influential drivers for Feeagh. Similarly, for Sau, soil moisture and temperature at similar depths were the most influential drivers. Another key driver that was shared between Feeagh and Sau was Julian day. Lake volume was only important for Sau, while solar radiation, the amount of carbon entering the water body (indicated by the DOC inflow concentration) and both soil moisture and temperature at 28-100 cm were only influential for Feeagh.

3.2 Driver influence on fDOM predictions

Figure 4 introduces SHAP beeswarm plots for the most influential drivers selected in Figure 3, enabling the assessment of the individual effect of each driver on fDOM predictions.

Seasonal patterns in fDOM concentrations were observed in both Feeagh and Sau, with higher values in winter and lower in summer, as reflected in the influence of Julian day. However, the key predictors and their effects differed substantially, shaped by contrasting catchment and climate characteristics. In Feeagh, where precipitation is relatively high and sustained year-round, deep soil temperature (100–255 cm) was the dominant and potentially limiting predictor, with fDOM increasing with temperature up to a threshold, beyond which a drop in the water table may counteract the effect (see partial dependance plots; Fig. A9 and A10). In addition, Feeagh showed minimal influence of mid-depth soil temperature (28-100 cm) and solar radiation on fDOM. This temperature-fDOM relationship was less relevant in Sau, where deep soil temperature (100-255 cm) had less explanatory power. Instead, fDOM dynamics in Sau, were driven primarily by soil moisture at both 28-100 cm and 100-255 cm depths, potentially depicting a limiting condition by water stress.

Water availability also shaped the role of other predictors differently across the two sites. For instance, surface water temperature in Feeagh showed a clear threshold behaviour, with fDOM increasing relatively linearly beyond 6.5ºC and stabilizing around 7.5ºC, while in Sau, water volume acted as a surrogate for fDOM production (see partial dependance plots; Fig. A9 and A10). Lower volumes corresponded to reduced fDOM values, which increase and stabilize beyond 146 hm^3. Interestingly, inflow DOC concentration influences Feeagh more than Sau, likely due to differences in hydroclimatological processes governing these relationships. These differences highlight how catchment water availability fundamentally alters the relative importance and behavior of fDOM drivers.

3.3 Predicting DOM using Supervised Machine Learning

Table 1 presents a comparative evaluation of the seven ML and statistical models employed in this study. In Lough Feeagh, CatBoost (CTB) demonstrated the best overall performance with the highest R^2 (0.50) and KGE (0.68), and a relatively low RMSE (8.17 QSU). Similarly, in Sau Reservoir, CTB again has the highest R^2 (0.54) and KGE (0.66), and one of the lowest RMSE (7.11 QSU). While some models like RF and LM showed moderate performance at Feeagh, others such as SVR and XGB performed poorly in Sau, with SVR even having a negative KGE (-0.82), suggesting significant model bias. Overall, CatBoost (CTB) consistently outperformed all other models across both sites, supporting its selection for fDOM prediction using the selected environmental drivers.

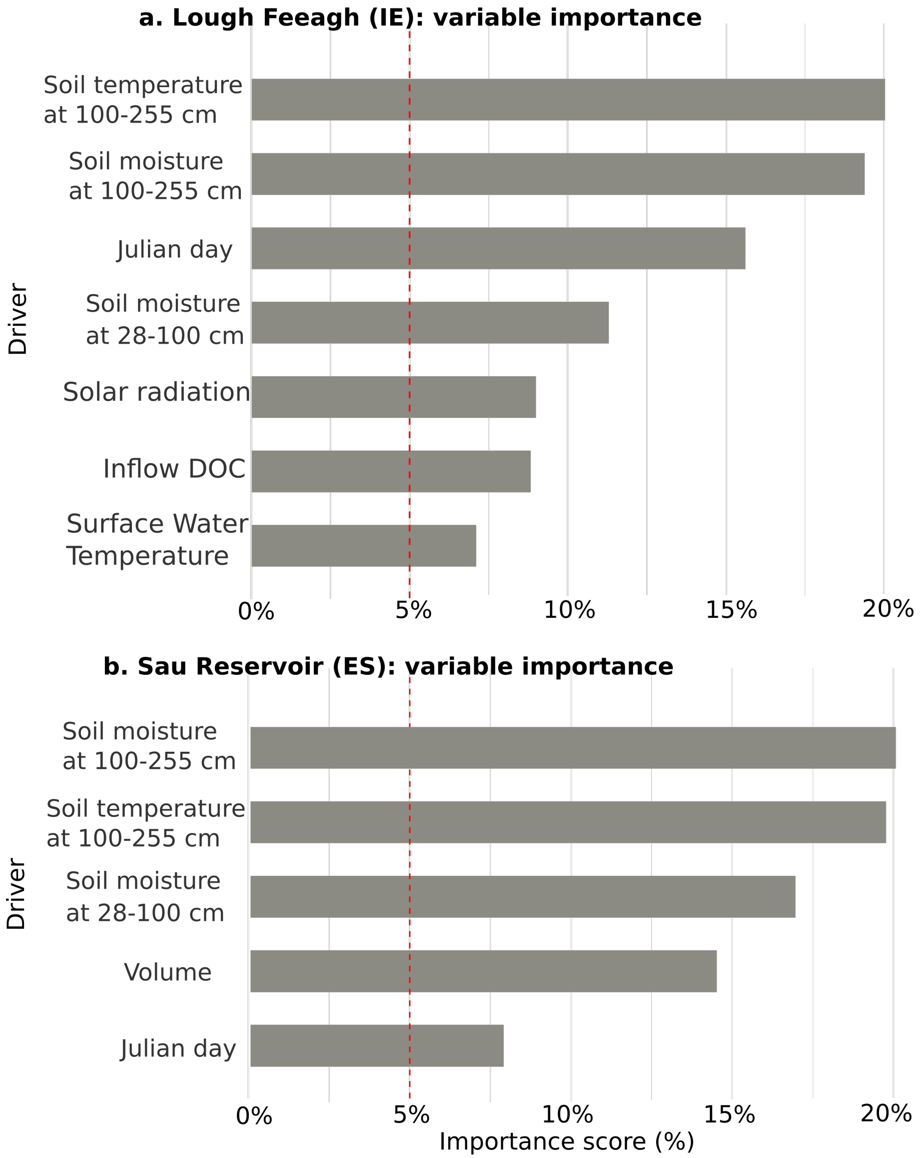

Figure 3. Selection of influential drivers for predicting fDOM. Twenty-four drivers from multiple sources were used to train the machine-learning models, including climate variables, soil variables, hydrologic and water-quality model outputs, lake model outputs, and cosine-transformed Julian day (see Figure 2 for more details). Feature importance is shown for the best-performing model (CatBoost) using gain contribution, and drivers exceeding 5% importance are highlighted.

The results during the training phase (supplementary Table A2) confirm that most ML models, especially XGB, RF, and KNN, had a high performance during training. However, the performance dropped in some models (e.g., XGB at Sau Reservoir) during the testing phases, underscoring the importance of evaluating models on independent test data to assess general-

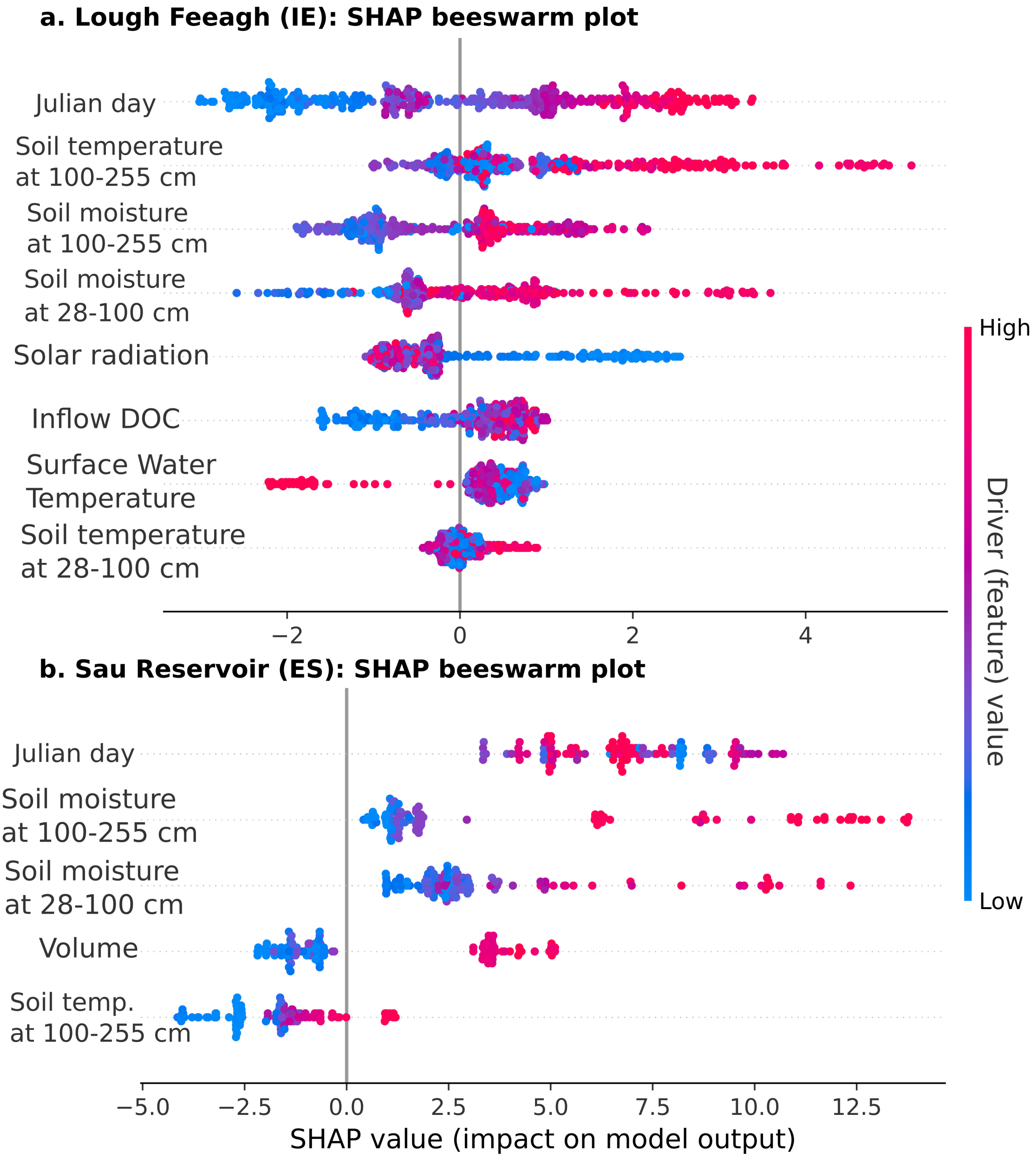

Figure 4. SHAP beeswarm plots showing the contribution of key drivers to fDOM predictions for (a) Lough Feeagh and (b) Sau Reservoir based on the CatBoost model (CTB). Points represent individual observations, coloured by feature value, and are ordered by mean absolute SHAP value.

isability and overfitting. CatBoost (CTB), again, presented a more stable performance, showing slightly lower training metrics (especially in Feeagh) compared with the other ML models and better generalisation when comparing with the metrics during

Table 1. Model comparison to predict fDOM using all selected drivers in both study sites. Statistic metrics (R^2, RMSE and KGE) were calculated to compare the performance of the models during the testing (hold-out) period. The table support the selection of the best model for each study site between Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting (LGB), CatBoost (CTB), k-Nearest Neighbors (KNN), Support Vector Regression (SVR) and linear model. Additionally, we tested other boosting methods, performance metrics for testing phase can be found in supplementary Table A2. Supplementary Table A1 contains the results during the training period for comparison.

	Lough Feeagh							Sau Reservoir						
Model/Metric	**RF**	**XGB**	**LGB**	**CTB**	**LM**	**KNN**	**SVR**	**RF**	**XGB**	**LGB**	**CTB**	**LM**	**KNN**	**SVR**
R^2	0.37	0.41	0.47	0.51	0.40	0.36	0.34	0.53	0.11	0.45	0.54	0.45	0.33	0.63
RMSE (QSU)	9.04	9.27	8.22	8.29	8.52	9.85	10.54	8.24	12.4	9.98	7.11	6.58	9.23	17.05
KGE	0.55	0.52	0.63	0.69	0.55	0.12	0.30	0.55	0.21	0.33	0.66	0.31	0.43	-0.82

testing. Feeagh exhibited more consistent model performance, with less overfitting between training and testing phases, across the different validation methods (Fig A7), compared to Sau. This difference is likely attributable to the limited amount of available data at Sau. In addition, we assess whether this overfitting could be related to the moderate skewness of the fDOM data in Sau, however similar results and performance metrics were found when applying log transformation (to remove skewness) before applying the ML models (see Fig. A11 and Tables A3 and A4).

3.3.1 Model prediction using all drivers compared to a reduced set

Figure 5 presents the fDOM prediction performance of the CatBoost model (best ML model overall) for Feeagh and Sau, using different input configurations. The models were trained on 85% of the time series (blue points) and tested on the remaining 15% using a hold-out approach (violet and green points). Two scenarios were compared: one using the most influential drivers (8 for Feeagh, 5 for Sau), and a second using only a reduced subset of reanalysis-based and easily accessible drivers, specifically soil temperature, soil moisture, and Julian day.

For both lakes, the reduced driver models showed only a modest decline in predictive performance during the testing period. For example, in Feeagh, the model using all influential drivers achieved $R^2 = 0.51$ and KGE = 0.69, while the reduced model still attained $R^2 = 0.48$ and KGE = 0.67. Similarly, in Sau, the full model scored $R^2 = 0.54$ and KGE = 0.66, whereas the reduced model maintained a comparable $R^2 = 0.50$ and KGE = 0.65. Although the training performance was higher in Sau compared with the testing performance, indicating potential overfitting, the CatBoost model provided informative and generalisable predictions in both study sites.

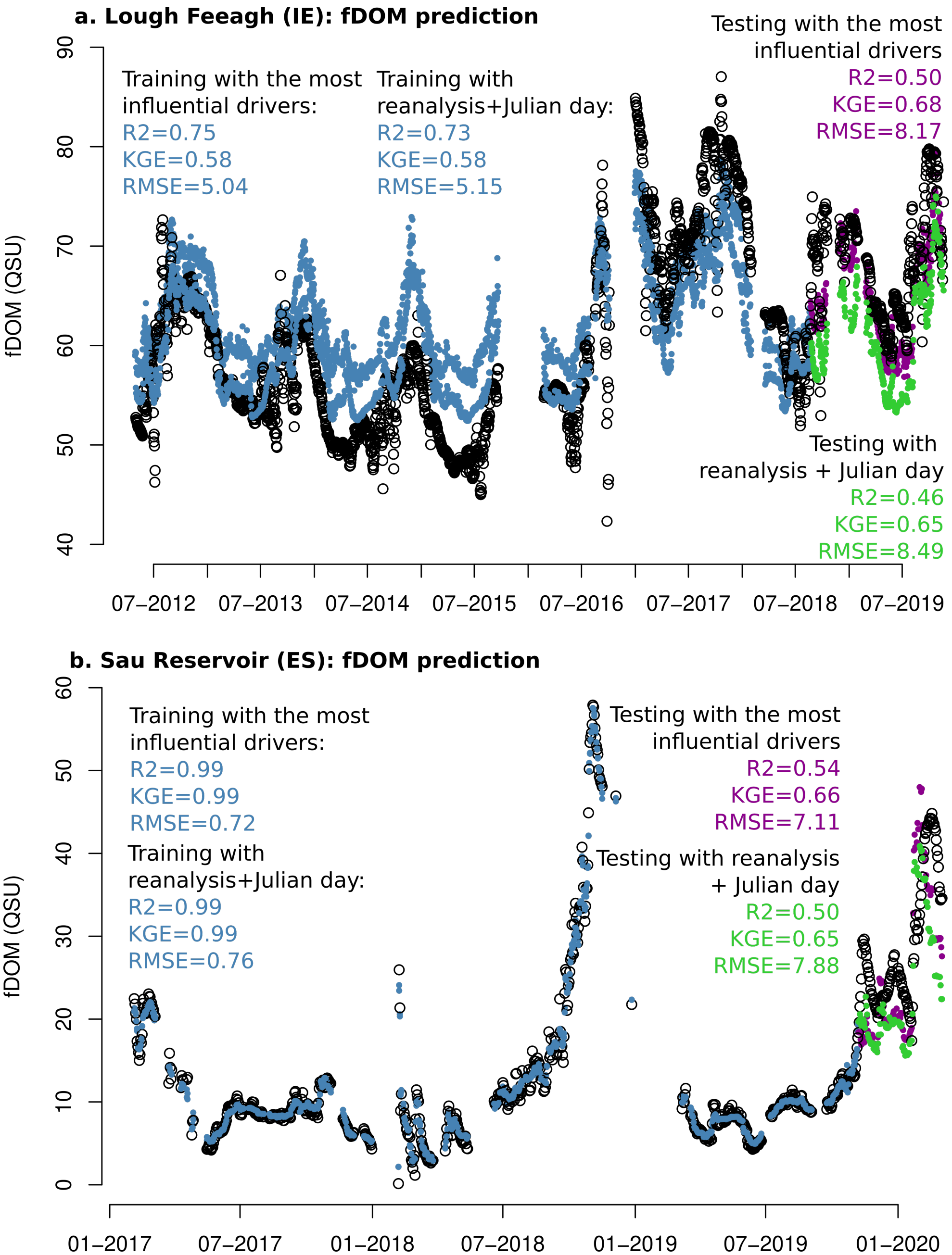

Figure 5. fDOM predictions obtained from the CatBoost model using different driver sets. Panels show training (blue; 85% of the time series) and testing periods (violet or green; 15%). (a) Feeagh: simulations using the most influential drivers (violet) and a reduced subset of reanalysis-based drivers + Julian day (green). (b) Sau: simulations using the most influential drivers (violet) and a reduced subset of reanalysis-based drivers + Julian day (green). Model performance metrics (R^2, KGE, RMSE) are reported for both periods.

4 Discussion

4.1 Driver attribution in contrasting conditions

The most influential drivers for fDOM identified for each site suggest potential carbon-producing processes that drive DOM dynamics in each lake. Interestingly, the most influential drivers were related to temperature and moisture at the deepest soil layers, indicating a common relevance for carbon inputs from terrestrial processes in both catchments. The Irish site is dominated by peat, a highly organic soil known to be sensitive to temperature-induced carbon release, especially as temperature levels rise. This is supported by the relation of soil temperatures with fDOM at this site (Ryder et al., 2014). In contrast, water availability constraints are much more pronounced in drier soils such as those of the Spanish site, as is validated by the partial dependence relationship of soil moisture with fDOM (Šimek et al., 2011). A plausible explanation is that, in Sau, increased soil moisture could boost biological activity, enhancing fDOM production, while in Feeagh, soil moisture showed a bell-shaped relationship, suggesting a more nuanced interplay between oxygen availability and microbial processes (Fig 4; Fig. A9 and A10). In addition, higher soil moisture can reflect stronger soil–lake hydrological connectivity, particularly during wet periods and flushing events, facilitating the transport of terrestrially derived DOM into the lake.

DOM dynamics are driven by physical and biogeochemical processes in the soil that are sensitive to changes in temperature and moisture, e.g., microbial processes that break down organic matter (Kalbitz et al., 2000). The fact that the deepest layers of the soil were more important in the model for both sites than those shallower could be linked to potential carbon attenuation processes, such as soil organic matter decomposition and retention in the soil, including sorption processes (Dubeux et al., 2024; Rumpel and Kögel-Knabner, 2011). In any case, the production of carbon in the catchment that eventually ends up in the lake requires concurrent downstream transport, governed by rainfall events.

Climate and topography (see supplementary Fig. A1) dictate the flushing of accumulated DOM during rainfall events, but also can influence sustained baseflow DOM contributions. In Feeagh, carbon exports from the catchment have been observed regularly throughout the entire annual cycle, with a seasonal variability (Doyle et al., 2019). In contrast, DOM in Sau accumulates primarily during the summer and is mainly flushed out via surface runoff during the wetter winter months (Marcé et al., 2021). These patterns are supported by the relationship between inflow DOC concentration and fDOM at both sites (Fig. 4, Figures A9 and A10), which shows a slight increase in predicted fDOM under lower carbon input conditions. Thus, inflow DOC concentration could reflect discharge pulses and dilution effects driven by precipitation (Jennings et al., 2020), following the characteristic seasonality of each site.

Seasonality plays a crucial role in fDOM predictions, as evidenced by the relationship of the Julian day driver with DOM dynamics at both sites (Fig. 4). At the Irish site, DOM seasonality is primarily shaped by natural environmental processes, whereas in the Spanish site, human influence plays a much greater role. This distinction helps explain why surface lake water temperature and solar radiation, two variables typically linked to strong seasonal patterns, were important only for Feeagh, while reservoir volume was significant only for Sau (Figures 3 and 4; Figures A9 and A10). Volume and soil variables produce a similar effect on fDOM as Julian day at Sau, given that higher volumes closer to the winter season can lead to higher fDOM values. Incorporating Julian day into the workflow offers a simple yet effective way to represent seasonality, potentially replacing seasonal variables (e.g., air temperature) (see the correlation matrix of all drivers in supplementary Fig. A12). This proves that the use of machine learning approaches opens up opportunities to assess diverse drivers under contrasting conditions.

Improving the accuracy of DOM predictions in lakes can enhance efforts to reduce further water quality deterioration and support lake management. This study demonstrated a feasible approach for simulating daily fDOM in two contrasting lakes, especially when using the Catboost model given its good generalisability. The performance metrics (Fig. 5) obtained at each site for the different model simulations lie in a similar or better range than comparable studies that modelled carbon dynamics in lakes (e.g., Harkort and Duan, 2023; Liu et al., 2021; Zhang et al., 2021, 2024). It is important to be aware, however, that previous studies are based on different frameworks. These include variations in the machine learning algorithms used, the target variable for quantifying carbon dynamics, with most studies having focused on DOC, whereas fDOM is the target variable here, as well as differences in input driver data and site-specific conditions. While both fDOM and DOC are widely used indicators of dissolved organic matter in lakes, they differ in measurement principles and in the fractions of organic matter they represent. Both variables, however, remain ecologically relevant for understanding DOM sources, transport, and transformation processes, which aligns with the context of this work.

4.2 Scalability

Our results suggest high potential for scalability, as predictive performance remained consistent across different driver sets even for two contrasting study sites, and performed good using only reanalysis data that is globally available and Julian day (Fig. 5). Importantly, this consistency remained even when specific highly-influential drivers were removed from the driver set. For instance, in Sau, where human intervention makes future reservoir outflows difficult to predict, avoiding reliance on water volume as a driver proved advantageous, as its removal from the set of drivers maintained model performance, despite being identified as an influential variable. It is likely that soil moisture at the deepest layer, a variable that showed a behavior similar to that of volume (Fig. 4), may have contributed to maintain the predictive performance when volume is removed from the driver set. In addition, for Feeagh, the predictive capacity was also maintained when using only meteorological and soil drivers. This demonstrates that a large driver dataset, such as the 24-variable set used in this study, would not be necessary to produce an accurate prediction when modelling fDOM using supervised machine learning.

4.3 Limitations and future research

Our approach offers the opportunity to validate and deploy a workflow capable of delivering daily DOM predictions in both undisturbed and anthropized sites, even when only limited data on input drivers are available, while at the same time providing insights into the dominant drivers. However, a site-specific model validation, including identification of appropriate training and testing periods, hyperparameter tuning for each specific study site and assessment of overfitting is essential. In terms of driver attribution, it is of note that the relationships identified using machine learning may not always be related at a process level (Sullivan, 2022). In our case, however, many of these same drivers had already been identified for river DOC levels in the Feeagh catchment (e.g., Doyle et al., 2019; Ryder et al., 2014). While the workflow can be easily replicated, fDOM data or data for another proxy for DOM are required. It is of note that such proxies of DOM are increasingly being incorporated into water quality monitoring programmes, an aspect that is convenient for testing workflows such as the one described (Downing et al., 2012).

The workflow presented here is not recommended for climate change studies, as the drivers of DOM variability can significantly change under entirely new and unrecorded climatic conditions. Consequently, supervised machine learning may fail to capture the signal from the time series. Moreover, the method's reliance on historical patterns limits its ability to extrapolate beyond the range of observed environmental conditions (Mi et al., 2024). Future research could expand the application of this framework to a broader range of lakes, integrate additional drivers such as remote sensing-derived terrestrial and aquatic quantity and quality parameters (Duan et al., 2025).

5 Conclusions

By identifying the key environmental drivers of lake dissolved organic matter (DOM) dynamics, this study presents an open, robust and scalable workflow for daily DOM prediction using different ML algorithms. Validated in two hydroclimatic contrasting sites in Ireland (Lough Feeagh) and Spain (Sau Reservoir), the approach revealed that deep soil temperature is the dominant driver in the peat-rich, temperate Irish catchment, whereas deep soil moisture plays a more critical role in the drier, Mediterranean setting of the Spanish site. These primary drivers are further shaped by hydrological processes, seasonal variability, and human activities.

The workflow showed good predictive performance even when based solely on globally available reanalysis data, supporting its potential applicability to other freshwater systems worldwide. In addition to expanding the set of approaches available for lake DOM prediction, the workflow offers transparent driver attribution, contributing valuable insights into the natural and anthropogenic processes governing carbon cycling in aquatic ecosystems.

Code and data availability. All data and codes used in this study are available here: https://doi.org/10.5281/zenodo.19354955.

Author contributions. DMB wrote the original draft and conducted the main analysis. RP contributed to the main analysis and, the writing and revision of manuscript. DMB, RP, VM, EJ, and RM conceptualized and designed the study. VM, EJ, EE, and RM contributed to the writing and revision of the manuscript. EE, AG, MD, JG, and JJ collected and provided in-situ data and offered expert feedback.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This research was funded through "Horizon Europe funding program under Grant Agreement number 101081728" https://doi.org/10.3030/101081728, funded by the European Commision, as a part of the "Innovative tools to control organic matter and disinfection byproducts in drinking water" (intoDBP) project https://intodbp.eu/. DMB and RM were funded by the Catalan Water Agency (Agència Catalana de l'Aigua - ACA) as a part of the NEREIDA project number RDI001/24/000044 (subvenció R+D+I de la convocatòria ACC/1362/2024).

References

Asadollah, S. B. H. S., Safaeinia, A., Jarahizadeh, S., Alcalá, F. J., Sharafati, A., and Jodar-Abellan, A.: Dissolved organic carbon estimation in lakes: Improving machine learning with data augmentation on fusion of multi-sensor remote sensing observations, Water Research, 277, 123 350, 2025.

Bhateria, R. and Jain, D.: Water quality assessment of lake water: A review, Sustainable Water Resources Management, 2, 161–173, https://doi.org/10.1007/s40899-015-0014-7, 2016.

Biau, G. and Scornet, E.: A random forest guided tour, Test, 25, 197–227, 2016.

Breiman, L.: Random Forests, Machine Learning, 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Chen, C., Chen, Q., Yao, S., He, M., Zhang, J., Li, G., and Lin, Y.: Combining physical-based model and machine learning to forecast chlorophyll-a concentration in freshwater lakes, Science of The Total Environment, 907, 168 097, https://doi.org/10.1016/j.scitotenv.2023.168097, 2024.

Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, https://doi.org/10.1145/2939672.2939785, 2016.

Cortes, C. and Vapnik, V.: Support-vector networks, Machine Learning, 20, 273–297, https://doi.org/10.1007/BF00994018, 1995.

Creed, I. F., Bergström, A.-K., Trick, C. G., Grimm, N. B., Hessen, D. O., Karlsson, J., Kidd, K. A., Kritzberg, E., McKnight, D. M., Freeman, E. C., Senar, O. E., Andersson, A., Ask, J., Berggren, M., Cherif, M., Giesler, R., Hotchkiss, E. R., Kortelainen, P., Palta, M. M., and Weyhenmeyer, G. A.: Global change-driven effects on dissolved organic matter composition: Implications for food webs of northern lakes, Global Change Biology, 24, 3692–3714, https://doi.org/10.1111/gcb.14129, 2018.

Downing, B. D., Pellerin, B. A., Bergamaschi, B. A., Saraceno, J. F., and Kraus, T. E. C.: Seeing the light: The effects of particles, dissolved materials, and temperature on in situ measurements of DOM fluorescence in rivers and streams, Limnology and Oceanography: Methods, 10, 767–775, https://doi.org/10.4319/lom.2012.10.767, 2012.

Doyle, B. C., de Eyto, E., Dillane, M., Poole, R., McCarthy, V., Ryder, E., and Jennings, E.: Synchrony in catchment stream colour levels is driven by both local and regional climate, Biogeosciences, 16, 1053–1071, https://doi.org/10.5194/bg-16-1053-2019, 2019.

Duan, H., Cao, Z., Luo, J., and Shen, M.: AI-driven opportunities and challenges in lake remote sensing, Information Geography, p. 100014, https://doi.org/10.1016/j.infgeo.2025.100014, 2025.

Dubeux, J. C. B., Lira Junior, M. d. A., Simili, F. F., Bretas, I. L., Trumpp, K. R., Bizzuti, B. E., Garcia, L., Oduor, K. T., Queiroz, L. M. D., Acuña, J. P., and Mendes, C. T. E.: Deep soil organic carbon: A review, CABI Reviews, 19, https://doi.org/10.1079/cabireviews.2024.0024, 2024.

Fix, E.: Discriminatory Analysis: Nonparametric Discrimination, Consistency Properties, Tech. rep., USAF School of Aviation Medicine, 1985.

Gobler, C. J.: Climate Change and Harmful Algal Blooms: Insights and perspective, Harmful Algae, 91, 101 731, https://doi.org/10.1016/j.hal.2019.101731, 2020.

Hanson, P. C., Stillman, A. B., Jia, X., Karpatne, A., Dugan, H. A., Carey, C. C., Stachelek, J., Ward, N. K., Zhang, Y., Read, J. S., and Kumar, V.: Predicting lake surface water phosphorus dynamics using process-guided machine learning, Ecological Modelling, 430, 109 136, https://doi.org/10.1016/j.ecolmodel.2020.109136, 2020.

Harkort, L. and Duan, Z.: Estimation of dissolved organic carbon from inland waters at a large scale using satellite data and machine learning methods, Water Research, 229, 119 478, https://doi.org/10.1016/j.watres.2022.119478, 2023.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N.: ERA5 hourly data on single levels from 1940 to present, Copernicus Climate Change Service (C3S) Climate Data Store (CDS), https://doi.org/10.24381/cds.adbb2d47, retrieved April 14, 2025, from https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab=overview, Accessed on July 2025.

Herzsprung, P., Wentzky, V., Kamjunke, N., von Tümpling, W., Wilske, C., Friese, K., Boehrer, B., Reemtsma, T., Rinke, K., and Lechtenfeld, O. J.: Improved Understanding of Dissolved Organic Matter Processing in Freshwater Using Complementary Experimental and Machine Learning Approaches, Environmental Science & Technology, 54, 13 556–13 565, https://doi.org/10.1021/acs.est.0c02383, 2020.

Hipsey, M. R., Bruce, L. C., Boon, C., Busch, B., Carey, C. C., Hamilton, D. P., Hanson, P. C., Read, J. S., de Sousa, E., Weber, M., and Winslow, L. A.: A General Lake Model (GLM 3.0) for linking with high-frequency sensor data from the Global Lake Ecological Observatory Network (GLEON), Geoscientific Model Development, 12, 473–523, https://doi.org/10.5194/gmd-12-473-2019, 2019.

Hollister, J. W., Milstead, W. B., and Kreakie, B. J.: Modeling lake trophic state: A random forest approach, Ecosphere, 7, e01 321, https://doi.org/10.1002/ecs2.1321, 2016.

Jennings, E., de Eyto, E., Moore, T., Dillane, M., Ryder, E., Allott, N., Nic Aonghusa, C., Rouen, M., Poole, R., and Pierson, D. C.: From Highs to Lows: Changes in Dissolved Organic Carbon in a Peatland Catchment and Lake Following Extreme Flow Events, Water, 12, 2843, https://doi.org/10.3390/w12102843, 2020.

Kalbitz, K., Solinger, S., Park, J.-H., Michalzik, B., and Matzner, E.: Controls on the dynamics of dissolved organic matter in soils: A review, Soil Science, 165, 277–300, 2000.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y.: LightGBM: A highly efficient gradient boosting decision tree, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 3149–3157, 2017.

Lake, P. S., Palmer, M. A., Biro, P., Cole, J., Covich, A. P., Dahm, C., Gibert, J., Goedkoop, W., Martens, K., and Verhoeven, J.: Global Change and the Biodiversity of Freshwater Ecosystems: Impacts on Linkages between Above-Sediment and Sediment Biota, BioScience, 50, 1099–1107, https://doi.org/10.1641/0006-3568(2000)050[1099:GCATBO]2.0.CO;2, 2000.

Li, A., Zhao, X., Mao, R., Liu, H., and Qu, J.: Characterization of dissolved organic matter from surface waters with low to high dissolved organic carbon and the related disinfection byproduct formation potential, Journal of Hazardous Materials, 271, 228–235, https://doi.org/10.1016/j.jhazmat.2014.02.009, 2014.

Li, B., Yang, G., Wan, R., Dai, X., and Zhang, Y.: Comparison of random forests and other statistical methods for the prediction of lake water level: A case study of the Poyang Lake in China, Hydrology Research, 47, 69–83, https://doi.org/10.2166/nh.2016.264, 2016.

Li, M., del Giorgio, P. A., Parkes, A. H., and Prairie, Y. T.: The relative influence of topography and land cover on inorganic and organic carbon exports from catchments in southern Quebec, Canada, Journal of Geophysical Research: Biogeosciences, 120, 2562–2578, https://doi.org/10.1002/2015JG003073, 2015.

Liu, D., Yu, S., Xiao, Q., Qi, T., and Duan, H.: Satellite estimation of dissolved organic carbon in eutrophic Lake Taihu, China, Remote Sensing of Environment, 264, 112 572, https://doi.org/10.1016/j.rse.2021.112572, 2021.

Marcé, R., Verdura, L., and Leung, N.: Dissolved organic matter spectroscopy reveals a hot spot of organic matter changes at the river–reservoir boundary, Aquatic Sciences, 83, 67, https://doi.org/10.1007/s00027-021-00823-6, 2021.

McCullough, I. M., Dugan, H. A., Farrell, K. J., Morales-Williams, A. M., Ouyang, Z., Roberts, D., Scordo, F., Bartlett, S. L., Burke, S. M., Doubek, J. P., et al.: Dynamic modeling of organic carbon fates in lake ecosystems, Ecological Modelling, 386, 71–82, 2018.

Mercado-Bettín, D., Clayer, F., Shikhani, M., Moore, T. N., Frías, M. D., Jackson-Blake, L., Sample, J., Iturbide, M., Herrera, S., French, A. S., Norling, M. D., Rinke, K., and Marcé, R.: Forecasting water temperature in lakes and reservoirs using seasonal climate prediction, Water Research, 201, 117 286, https://doi.org/10.1016/j.watres.2021.117286, 2021.

Mi, C., Tilahun, A. B., Flörke, M., Dürr, H. H., and Rinke, K.: Climate warming effects in stratified reservoirs: Thorough assessment for opportunities and limits of machine learning techniques versus process-based models in thermal structure projections, Journal of Cleaner Production, 454, 142 347, https://doi.org/10.1016/j.jclepro.2024.142347, 2024.

Molnar, C.: Interpretable machine learning, Lulu. com, 2020.

Müller, M., D'Andrilli, J., Silverman, V., Bier, R. L., Barnard, M. A., Lee, M. C. M., Richard, F., Tanentzap, A. J., Wang, J., de Melo, M., and Lu, Y.: Machine-learning based approach to examine ecological processes influencing the diversity of riverine dissolved organic matter composition, Frontiers in Water, 6, https://doi.org/10.3389/frwa.2024.1379284, 2024.

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., and Gupta, H. V.: What Role Does Hydrological Science Play in the Age of Machine Learning?, Water Resources Research, 57, e2020WR028 091, https://doi.org/10.1029/2020WR028091, 2021.

Paíz, R., Pierson, D. C., Lindqvist, K., Naden, P. S., de Eyto, E., Dillane, M., McCarthy, V., Linnane, S., and Jennings, E.: Accounting for model parameter uncertainty provides more robust projections of dissolved organic carbon dynamics to aid drinking water management, Water Research, 276, 123 238, https://doi.org/10.1016/j.watres.2025.123238, 2025a.

Paíz, R., Thomas, R. Q., Carey, C. C., de Eyto, E., Jones, I. D., Delany, A. D., Poole, R., Nixon, P., Dillane, M., McCarthy, V., Linnane, S., and Jennings, E.: Near-term lake water temperature forecasts can be used to anticipate the ecological dynamics of freshwater species, Ecosphere, 16, e70 335, https://doi.org/10.1002/ecs2.70335, 2025b.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A.: CatBoost: Unbiased boosting with categorical features, Tech. Rep. arXiv:1706.09516, arXiv, https://doi.org/10.48550/arXiv.1706.09516, 2019.

Qi, Y.: Random Forest for Bioinformatics, pp. 307–323, Springer, https://doi.org/10.1007/978-1-4419-9326-7_11, 2012.

Regier, P., Duggan, M., Myers-Pigg, A., and Ward, N.: Effects of random forest modeling decisions on biogeochemical time series predictions, Limnology and Oceanography: Methods, 21, 40–52, https://doi.org/10.1002/lom3.10523, 2023.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., and Chica-Rivas, M.: Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines, Ore Geology Reviews, 71, 804–818, https://doi.org/10.1016/j.oregeorev.2015.01.001, 2015.

Rumpel, C. and Kögel-Knabner, I.: Deep soil organic matter—A key but poorly understood component of terrestrial C cycle, Plant and Soil, 338, 143–158, https://doi.org/10.1007/s11104-010-0391-5, 2011.

Ryder, E., Jennings, E., de Eyto, E., Dillane, M., NicAonghusa, C., Pierson, D. C., Moore, K., Rouen, M., and Poole, R.: Temperature quenching of CDOM fluorescence sensors: Temporal and spatial variability in the temperature response and a recommended temperature correction equation, Limnology and Oceanography: Methods, 10, 1004–1010, https://doi.org/10.4319/lom.2012.10.1004, 2012.

Ryder, E., de Eyto, E., Dillane, M., Poole, R., and Jennings, E.: Identifying the role of environmental drivers in organic carbon export from a forested peat catchment, Science of The Total Environment, 490, 28–36, https://doi.org/10.1016/j.scitotenv.2014.04.091, 2014.

Solomon, C. T., Jones, S. E., Weidel, B., Buffam, I., Fork, M. L., Karlsson, J., Larsen, S., Lennon, J. T., Read, J. S., Sadro, S., and Saros, J. E.: Ecosystem consequences of changing inputs of terrestrial dissolved organic matter to lakes: Current knowledge and future challenges, Ecosystems, 18, 376–389, https://doi.org/10.1007/s10021-015-9848-y, 2015.

Sullivan, E.: Understanding from Machine Learning Models, The British Journal for the Philosophy of Science, 73, 109–133, https://doi.org/10.1093/bjps/axz035, 2022.

Toming, K., Kotta, J., Uuemaa, E., Sobek, S., Kutser, T., and Tranvik, L. J.: Predicting lake dissolved organic carbon at a global scale, Scientific Reports, 10, 8471, https://doi.org/10.1038/s41598-020-65010-3, 2020.

Šimek, K., Comerma, M., García, J.-C., Nedoma, J., Marcé, R., and Armengol, J.: The Effect of River Water Circulation on the Distribution and Functioning of Reservoir Microbial Communities as Determined by a Relative Distance Approach, Ecosystems, 14, 1–14, https://doi.org/10.1007/s10021-010-9388-4, 2011.

Watras, C., Hanson, P., Stacy, T., Morrison, K., Mather, J., Hu, Y.-H., and Milewski, P.: A temperature compensation method for CDOM fluorescence sensors in freshwater, Limnology and Oceanography: Methods, 9, 296–301, 2011.

Weyhenmeyer, G. A. and Karlsson, J.: Nonlinear response of dissolved organic carbon concentrations in boreal lakes to increasing temperatures, Limnology and Oceanography, 54, 2513–2519, https://doi.org/10.4319/lo.2009.54.6_part_2.2513, 2009.

Xenopoulos, M. A., Barnes, R. T., Boodoo, K. S., Butman, D., Catalán, N., D'Amario, S. C., Fasching, C., Kothawala, D. N., Pisani, O., Solomon, C. T., Spencer, R. G. M., Williams, C. J., and Wilson, H. F.: How humans alter dissolved organic matter composition in freshwater: Relevance for the Earth's biogeochemistry, Biogeochemistry, 154, 323–348, https://doi.org/10.1007/s10533-021-00753-3, 2021.

Zhang, D., Shi, K., Wang, W., Wang, X., Zhang, Y., Qin, B., Zhu, M., Dong, B., and Zhang, Y.: An optical mechanism-based deep learning approach for deriving water trophic state of China's lakes from Landsat images, Water Research, 252, 121 181, https://doi.org/10.1016/j.watres.2024.121181, 2024.

Zhang, Y., Yao, X., Wu, Q., Huang, Y., Zhou, Z., Yang, J., and Liu, X.: Turbidity prediction of lake-type raw water using random forest model based on meteorological data: A case study of Tai lake, China, Journal of Environmental Management, 290, 112 657, https://doi.org/10.1016/j.jenvman.2021.112657, 2021.