

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

Daniel Mercado-Bettín¹, Ricardo Paíz^{2,3}, Valerie McCarthy², Eleanor Jennings³, Elvira de Eyto⁴, Angeles M. Gallegos⁵, Mary Dillane⁴, Juan C. Garcia⁵, José J. Rodríguez⁵, and Rafael Marcé¹

¹Centre for Advanced Studies of Blanes, Spanish National Research Council, Carrer Accés Cala Sant Francesc, 14, 17300, Blanes, Spain

²School of History and Geography, Dublin City University, D09 YT18, Dublin 9, Ireland

³Centre for Freshwater and Environmental Studies, Dundalk Institute of Technology, A91 K584 Dundalk, Co. Louth, Ireland

⁴Fisheries & Ecosystem Advisory Services, Marine Institute, F28 PF65 Newport, Co. Mayo, Ireland

⁵Ens d'Abastament d'Aigua Ter-Llobregat, Ctra. Aiguës, 6, 08440, Cardedeu, Spain

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

1 **Abstract.** Predicting water quality variables in lakes is critical for effective ecosystem management under climatic and human
2 pressures. Dissolved organic matter (DOM) serves as an energy source for aquatic ecosystems and plays a key role in their bio-
3 geochemical cycles. However, predicting DOM is challenging due to complex interactions between multiple potential drivers
4 in the aquatic environment and its surrounding terrestrial landscape. This study establishes an open and scalable workflow to
5 identify potential drivers and predict fluorescent DOM (fDOM) in the surface layer of lakes by exploring the use of supervised
6 machine learning models, including random forest, boosting methods, k-nearest neighbors, support vector regression and linear
7 model. The workflow was validated in two contrasting systems: one natural lake in Ireland with a relatively undisturbed catch-
8 ment, and one reservoir in Spain with a more human-influenced catchment. A total of 24 potential drivers were obtained from
9 global reanalysis data, and lake and river process-based modelling. SHapley Additive exPlanations (SHAP) were conducted
10 for the most influential drivers identified, with soil moisture, soil temperature, and Julian day being common to both study
11 sites. The best prediction was obtained when using the CatBoost model (during hold-out testing period, Irish site: KGE > 0.68,
12 $r^2 > 0.50$; Spanish site: KGE > 0.66, $r^2 > 0.54$). Interestingly, when only using drivers from globally accessible climate and
13 soil reanalysis data plus Julian day, the prediction capacity was maintained at both sites, showcasing potential for scalability.
14 Our findings highlight the complex interplay of environmental drivers and processes that govern DOM dynamics in lakes, and
15 contribute to the modelling of carbon cycling in aquatic ecosystems.

16 1 Introduction

17 Lakes are an essential component of global biogeochemical cycles, sustain biodiversity, and provide critical ecosystem services,
18 e.g., water supply, fishing and irrigation. However, their water quality is increasingly at risk due to climatic change and human
19 pressures (Bhateria and Jain, 2016). A key water quality variable is dissolved organic matter (DOM), which influences light
20 penetration, energy, oxygen dynamics and nutrient availability in any lake (Solomon et al., 2015). The dynamics of DOM
21 in lakes are driven by both external processes in the terrestrial environment and internal processes. Land cover, climate, and

22 topography regulate (either increase or decrease) carbon production in the catchment and carbon inputs into the lake (Li et al.,
23 2015). In the water body, the quantity and quality of DOM are controlled by physical and biogeochemical mechanisms such as
24 photodegradation, microbial processing and mixing dynamics, but can also be impacted by water abstraction or dam regulation
25 (Xenopoulos et al., 2021).

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

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

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

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

57 test how accurately can supervised ML models predict lake fDOM driven by reanalysis-based data, and hydrologic and lake
58 modeling outputs.

59 **2 Materials and methods**

60 **2.1 Study sites**

61 Lough Feeagh and Sau Reservoir are located in western Ireland (53° 56' N, 9° 35' W) and northeastern Spain (41° 58' N, 2°
62 23' E), respectively (Fig. 1). The study sites have contrasting attributes. Feeagh (depth of 46.8 m and area of 3.95 km²) is a
63 monomictic and oligotrophic lake surrounded by a relatively undisturbed landscape, while Sau (depth of 70 m and area of 5.8
64 km²) is an eutrophic system subjected to human activities and water abstraction. Feeagh has two primary inflows, the Black and
65 the Glenamong rivers, while Sau has one, the Ter river. The catchment of Feeagh is relatively small (84 km²), with mid-range
66 hills, and dominated by peatland. The catchment of Sau, in contrast, is larger (1525 km²), with a varying topography and land
67 uses (Fig.1; Fig. A1).

68 The dynamics of DOM in both study sites have been previously explored in Ryder et al. (2014) which identified that
69 natural dynamics related to soil temperature, river discharge and drought were important drivers in Feeagh, and in Marcé et al.
70 (2021), which showed human activities (e.g., wastewater effluents and agricultural runoff) were important for Sau, in addition
71 to its environmental dynamics. Catchment hydrology is key for carbon transport into both study sites, and contributes to a
72 distinct seasonality related to climate. Feeagh has a wet temperate climate, with cooler air temperatures and higher rainfall
73 levels that occur on more than 75% of days in the year. The variability of carbon inputs reflects the sensitivity of a peatland-
74 dominated landscape, which exacerbates climate-induced carbon release from the catchment into the lake. In contrast, Sau
75 has a Mediterranean climate, characterised by hot, dry summers and mild, wet winters, dictating water availability, thermal
76 stratification, and organic matter fluxes in the reservoir.

77 **2.2 Prediction workflow**

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

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

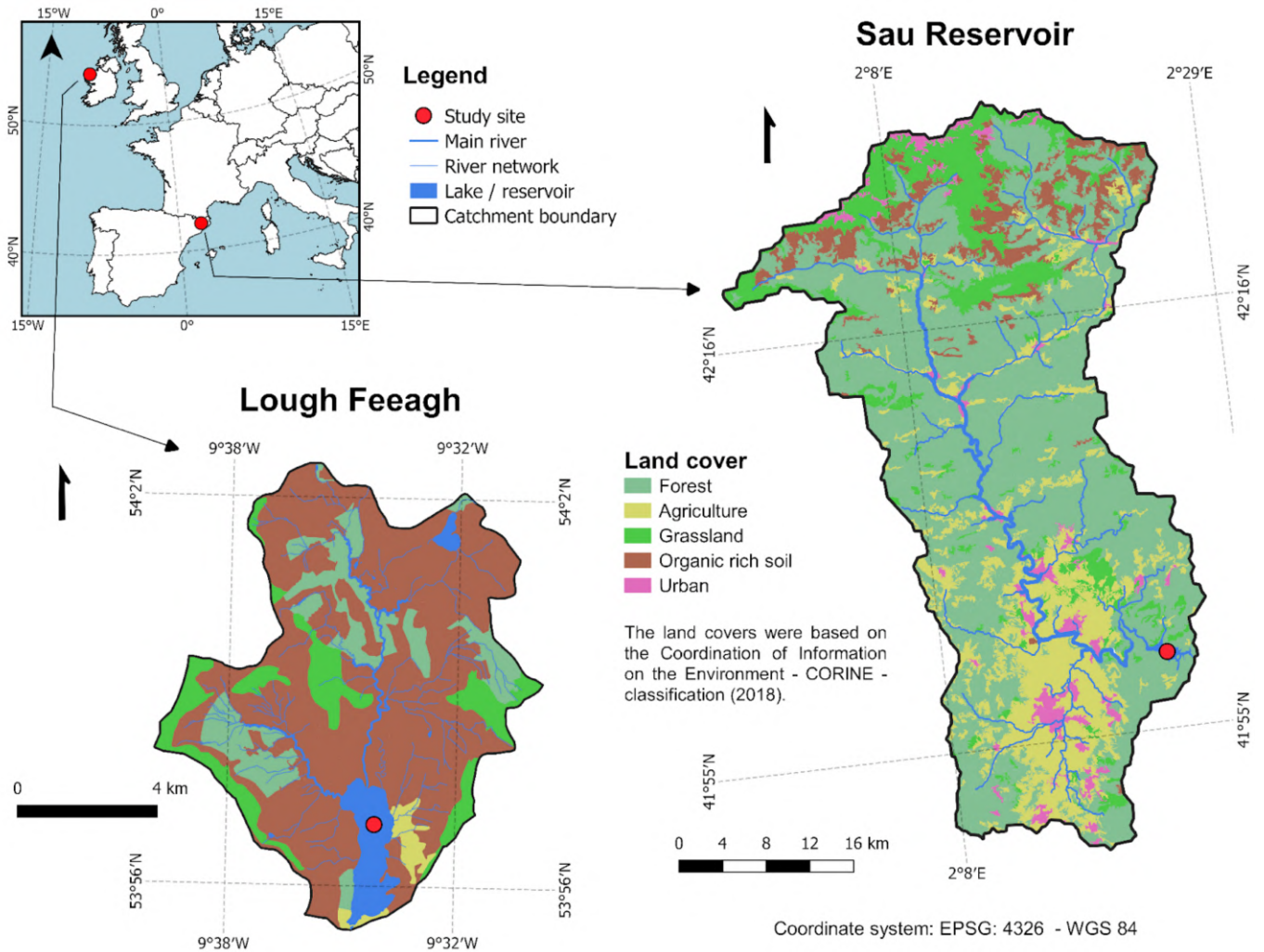


Figure 1. Two contrasting freshwater ecosystems: Lough Feeagh (Ireland) and Sau Reservoir (Spain). The figure shows their locations and catchment land cover. Lough Feeagh is a humic oligotrophic lake with a peatland-dominated catchment and temperate oceanic climate, whereas Sau is a eutrophic reservoir 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).

88 The same workflow was applied to both sites using the same data sources, allowing for comparison. Following the FAIR
 89 principles, all data and workflow scripts are available and fully reproducible in the following repository: https://github.com/danielmerbet/driver_attribution_fdom, a DOI will be provided in the final version of the manuscript. Large language models
 90 were used in this study to optimise the codes, improve the final plots and, for basic proofreading of the text.
 91

Workflow

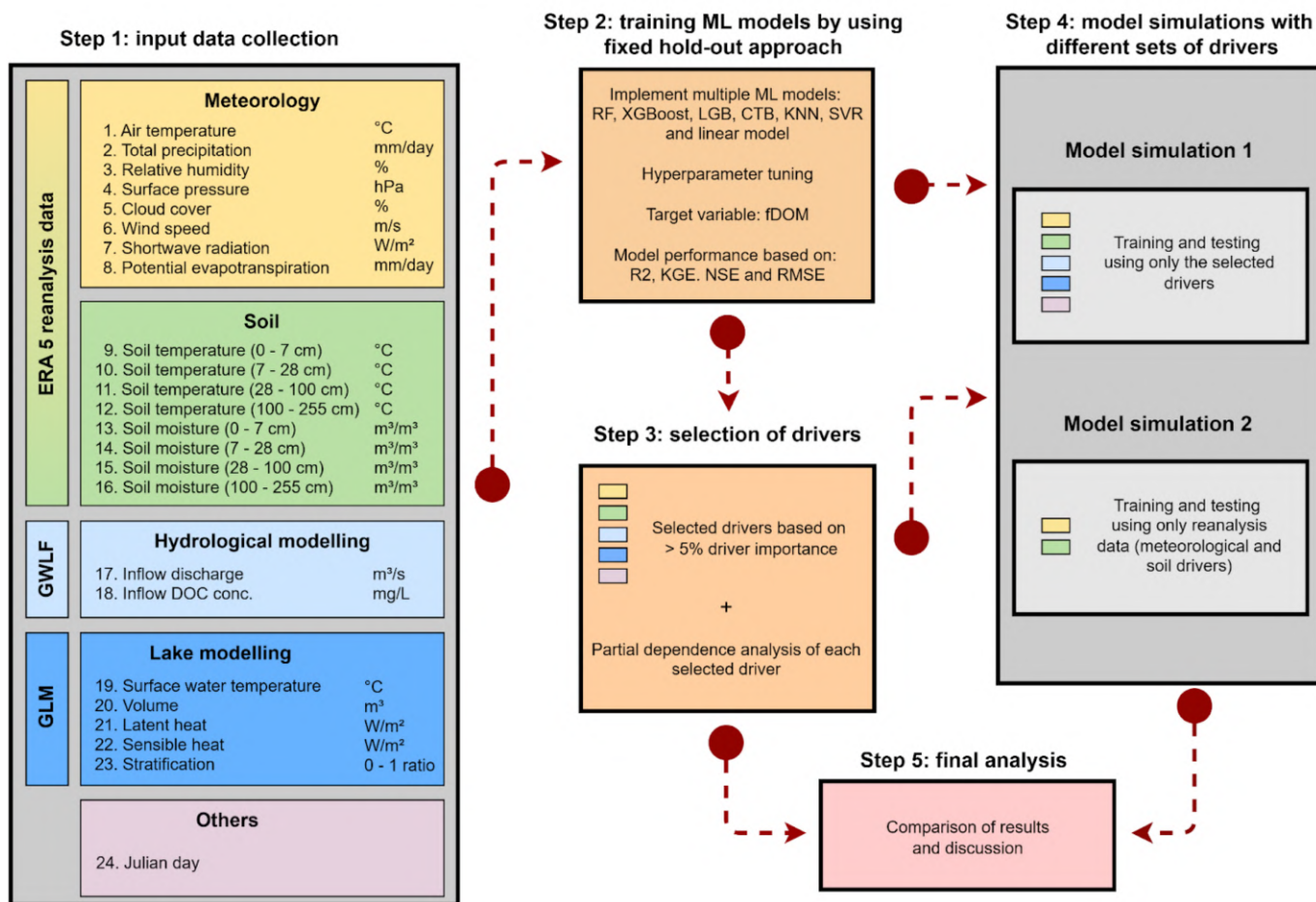


Figure 2. The workflow includes five steps: (1) compilation of input variables including meteorology (yellow) and soil (green) reanalysis data from ERA5 (the fifth-generation atmospheric reanalysis product produced by ECMWF), The Generalised Watershed Loading Functions Model (GWLF) outputs (light blue), 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

92 **2.3 Data**

93 **2.3.1 Target variable (fDOM)**

94 Daily surface fDOM values were obtained from high-frequency data for both sites (2-minute-resolution data averaged to obtain
95 daily values before analysis). For Feeagh, data were measured at 0.9 m depth, and for Sau, an average value was calculated
96 between the measurements for depths 0-5 m. fDOM data were expressed as quinine sulfate units (QSU) for the analysis. In
97 Feeagh, the data spanned from 1st of May 2012 to 19th of November 2019 (n = 2328), and for Sau from 4th of February 2017
98 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
99 constrained by the availability of fDOM data. Following exploratory data analysis (Fig. A3), no transformation was applied to
100 the fDOM data prior to training the ML models in the main manuscript, results for Sau are provided with a log₁₀ transformation
101 in the Supplementary Information for comparison (this is revisited in Results), given the moderate skewness of fDOM at this
102 study site.

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

109 **2.3.2 Driver data**

110 All input data variables, including their respective units and source, are displayed in Step 1 of Figure 2. These were grouped
111 into five categories: (1) meteorology, (2) soil, (3) process-based hydrological modelling, (4) process-based lake modelling
112 and (5) Julian day. All input data variables, including their respective units and source, are displayed in Step 1 of Figure 2.
113 Daily values of meteorology and soil variables were extracted from the European Centre for Medium-Range Weather Forecasts
114 (ECMWF) Reanalysis v5 (ERA5) dataset (Hersbach et al., Accessed on July 2025). This gridded dataset provides (pseudo)
115 observations at a global scale with a spatial resolution of 0.25°. Eight meteorological variables (traditionally employed in water
116 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)
117 were extracted for the grid-cell which contained each water body.

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

125 approximate calendar timing corresponds to the cosine of the Julian day: values close to 1 correspond to boreal winter, 0 to
126 spring/autumn, and -1 to boreal summer.

127 2.4 Supervised Machine Learning

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

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

140 Supervised machine-learning models were applied to predict fDOM dynamics, including Random Forest (RF) (Breiman,
141 2001), eXtreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016), Light Gradient Boosting (LGB) (Ke et al., 2017),
142 CatBoost (CTB) (Prokhorenkova et al., 2019), k-Nearest Neighbors (KNN) (Fix, 1985), Support Vector Regression (SVR)
143 (Cortes and Vapnik, 1995) and linear model. The three gradient-boosting frameworks were included to assess the robustness
144 of results across closely related implementations. Given their similarities, the main results shown are focused on the highest-
145 performing model among the three. Neural networks were not included because they typically require more complex architec-
146 ture design and calibration, which can increase methodological complexity and may reduce reproducibility of the workflow.
147 In contrast, the selected models allow more reproducible implementation due to their comparatively simpler hyperparameter
148 tuning.

149 2.4.1 Hyperparameter tuning

150 Hyperparameter tuning was implemented in all ML models to improve accuracy and generalization, and reduce the risk of over-
151 fitting. It was conducted in R using the caret framework where possible: for RF $mtry \in \{2, 4, 6\}$; for XGBoost $nrounds \in$
152 $\{1000, 2000\}$, $max_depth \in \{3, 6\}$, $eta \in \{0.01, 0.05, 0.1\}$, $gamma = 0$, $colsample_bytree = 1$, $min_child_weight = 1$,
153 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 im-
154 plementations were made: for LGB $learning_rate \in \{0.01, 0.05, 0.1\}$, $num_leaves \in \{15, 31, 63\}$, $max_depth \in \{5, 10, 1\}$,
155 $feature_fraction \in \{0.8, 1\}$, $bagging_fraction \in \{0.8, 1\}$, $nrounds = 2000$ (fixed); and for CTB $iterations \in \{300, 500, 1000\}$,
156 $depth \in \{4, 6, 8\}$, $learning_rate \in \{0.01, 0.05, 0.1\}$, and $l2_leaf_reg \in \{1, 3, 5\}$. Models were trained on the training subset

157 and evaluated on the hold-out test set using RMSE. The parameter configuration minimising test RMSE was selected. The
158 hyperparameter tuning can be reproduced by using the code “2_hyperparameter_tuning.R” in the repository.

159 For each case study, the dataset was chronologically split into training (85%) and hold-out test (15%) subsets. The final
160 15% of observations were reserved as an independent test set to evaluate predictive performance, mimicking a forecasting
161 exercise. Hyperparameter tuning was conducted exclusively on the training subset. Within the training data, model selection
162 was performed using 5-fold cross-validation. The data were randomly partitioned into five equally sized folds; four folds were
163 used for model fitting and one for validation, iteratively. The average cross-validated performance was used to select the optimal
164 hyperparameter configuration.

165 **2.4.2 Performance metrics**

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

171 **2.5 Driver importance and SHAP analysis**

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

181 **3 Results**

182 **3.1 Driver attribution**

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

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

192 **3.2 Driver influence on fDOM predictions**

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

195 Seasonal patterns in fDOM concentrations were observed in both Feeagh and Sau, with higher values in winter and lower
196 in summer, as reflected in the influence of Julian day. However, the key predictors and their effects differed substantially,
197 shaped by contrasting catchment and climate characteristics. In Feeagh, where precipitation is relatively high and sustained
198 year-round, deep soil temperature (100–255 cm) was the dominant and potentially limiting predictor, with fDOM increasing
199 with temperature up to a threshold, beyond which a drop in the water table may counteract the effect (see partial dependance
200 plots; Fig. A9 and A10). In addition, Feeagh showed minimal influence of mid-depth soil temperature (28-100 cm) and solar
201 radiation on fDOM. This temperature-fDOM relationship was less relevant in Sau, where deep soil temperature (100-255 cm)
202 had less explanatory power. Instead, fDOM dynamics in Sau, were driven primarily by soil moisture at both 28-100 cm and
203 100-255 cm depths, potentially depicting a limiting condition by water stress.

204 Water availability also shaped the role of other predictors differently across the two sites. For instance, surface water tem-
205 perature in Feeagh showed a clear threshold behaviour, with fDOM increasing relatively linearly beyond 6.5°C and stabilizing
206 around 7.5°C, while in Sau, water volume acted as a surrogate for fDOM production (see partial dependance plots; Fig. A9
207 and A10). Lower volumes corresponded to reduced fDOM values, which increase and stabilize beyond 146 hm³. Interest-
208 ingly, inflow DOC concentration influences Feeagh more than Sau, likely due to differences in hydroclimatological processes
209 governing these relationships. These differences highlight how catchment water availability fundamentally alters the relative
210 importance and behavior of fDOM drivers.

211 **3.3 Predicting DOM using Supervised Machine Learning**

212 Table 1 presents a comparative evaluation of the seven ML and statistical models employed in this study. In Lough Feeagh,
213 CatBoost (CTB) demonstrated the best overall performance with the highest R² (0.50) and KGE (0.68), and a relatively low
214 RMSE (8.17 QSU). Similarly, in Sau Reservoir, CTB again has the highest R² (0.54) and KGE (0.66), and one of the lowest
215 RMSE (7.11 QSU). While some models like RF and LM showed moderate performance at Feeagh, others such as SVR and
216 XGB performed poorly in Sau, with SVR even having a negative KGE (-0.82), suggesting significant model bias. Overall,
217 CatBoost (CTB) consistently outperformed all other models across both sites, supporting its selection for fDOM prediction
218 using the selected environmental drivers.

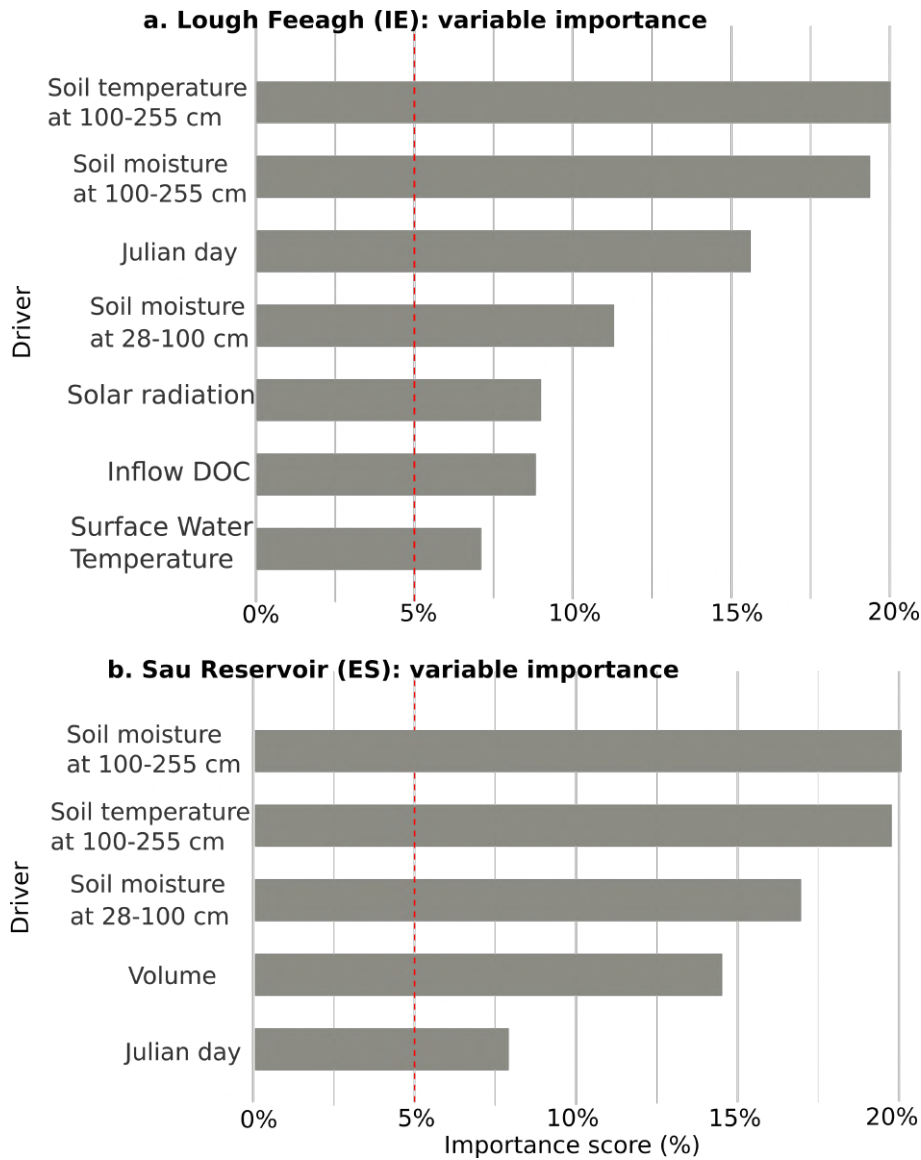


Figure 3. Selection of influential drivers for predicting fDOM. Twenty-four drivers from multiple sources were used to train the machine-learning models, including climate variables, soil variables, hydrologic and water-quality model outputs, lake model outputs, and cosine-transformed Julian day (see 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.

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

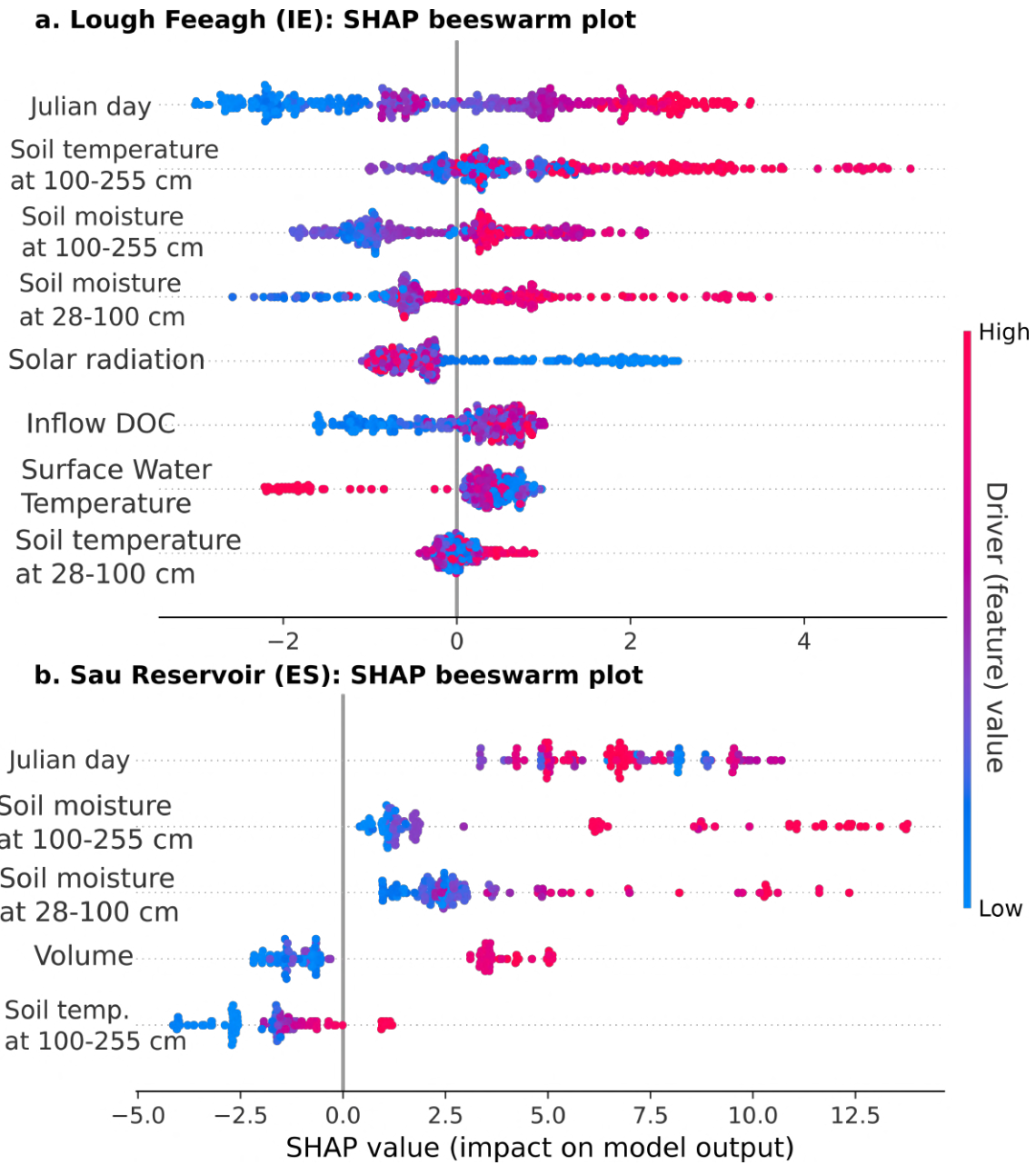


Figure 4. SHAP beeswarm plots showing the contribution of key drivers to fDOM predictions for (a) Lough Feagh and (b) Sau Reservoir based on the CatBoost model (CTB). Points represent individual observations, coloured by feature value, and are ordered by mean absolute SHAP value.

222 isability and overfitting. CatBoost (CTB), again, presented a more stable performance, showing slightly lower training metrics
 223 (especially in Feagh) 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.

Model/Metric	Lough Feeagh							Sau Reservoir						
	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

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

229 3.3.1 Model prediction using all drivers compared to a reduced set

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

235 For both lakes, the reduced driver models showed only a modest decline in predictive performance during the testing period.
 236 For example, in Feeagh, the model using all influential drivers achieved $R^2 = 0.51$ and $KGE = 0.69$, while the reduced model
 237 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
 238 model maintained a comparable $R^2 = 0.50$ and $KGE = 0.65$. Although the training performance was higher in Sau compared
 239 with the testing performance, indicating potential overfitting, the CatBoost model provided informative and generalisable
 240 predictions in both study sites.

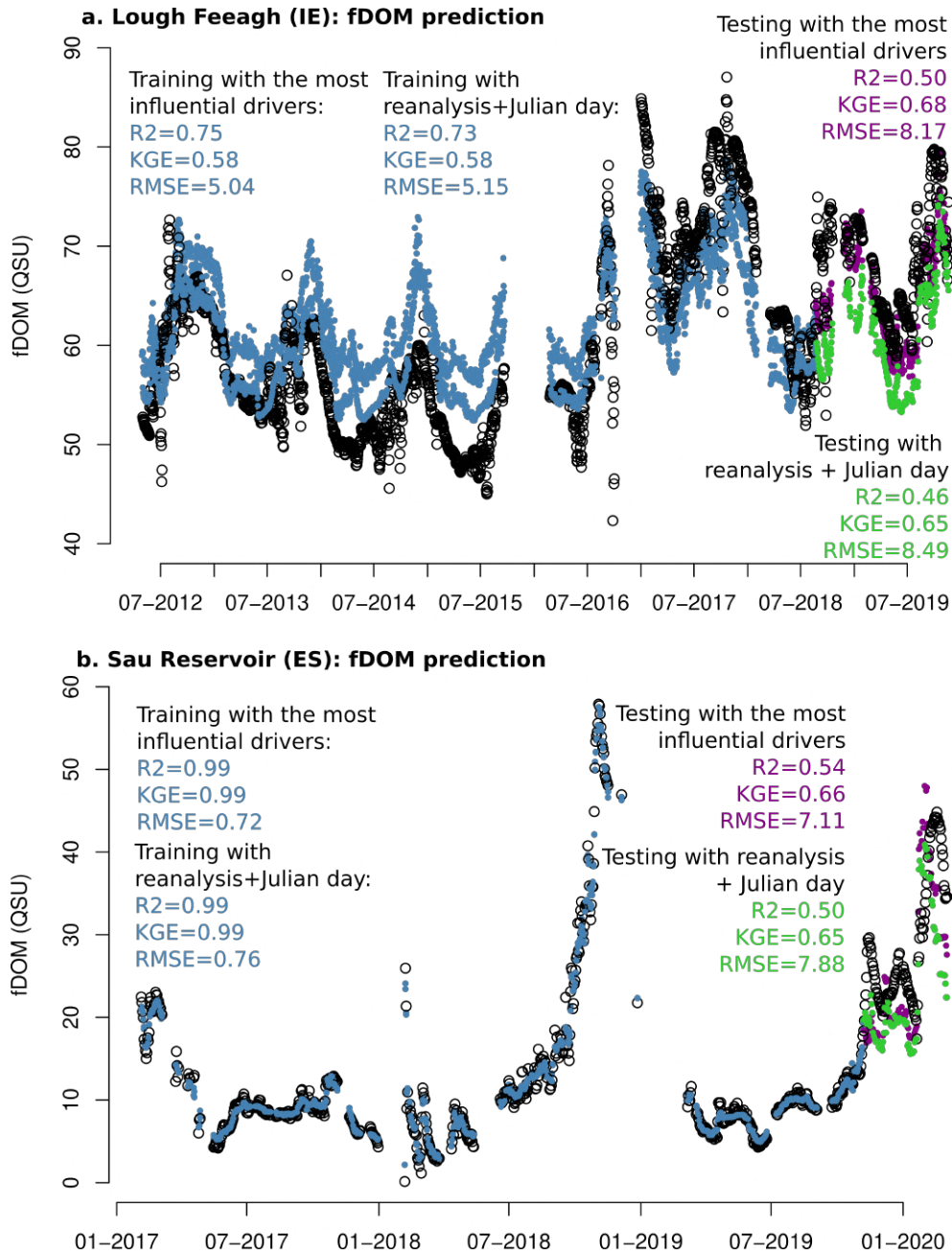


Figure 5. fDOM predictions obtained from the CatBoost model using different driver sets. Panels show training (blue; 85% of the time series) and testing periods (violet or green; 15%). (a) Feagh: simulations using the most influential drivers (violet) and a reduced subset of reanalysis-based drivers + 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.

241 4 Discussion

242 4.1 Driver attribution in contrasting conditions

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

254 DOM dynamics are driven by physical and biogeochemical processes in the soil that are sensitive to changes in temperature
255 and moisture, e.g., microbial processes that break down organic matter (Kalbitz et al., 2000). The fact that the deepest layers
256 of the soil were more important in the model for both sites than those shallower could be linked to potential carbon attenuation
257 processes, such as soil organic matter decomposition and retention in the soil, including sorption processes (Dubeux et al.,
258 2024; Rumpel and Kögel-Knabner, 2011). In any case, the production of carbon in the catchment that eventually ends up in
259 the lake requires concurrent downstream transport, governed by rainfall events.

260 Climate and topography (see supplementary Fig. A1) dictate the flushing of accumulated DOM during rainfall events, but
261 also can influence sustained baseflow DOM contributions. In Feeagh, carbon exports from the catchment have been observed
262 regularly throughout the entire annual cycle, with a seasonal variability (Doyle et al., 2019). In contrast, DOM in Sau accumu-
263 lates primarily during the summer and is mainly flushed out via surface runoff during the wetter winter months (Marcé et al.,
264 2021). These patterns are supported by the relationship between inflow DOC concentration and fDOM at both sites (Fig. 4,
265 Figures A9 and A10), which shows a slight increase in predicted fDOM under lower carbon input conditions. Thus, inflow
266 DOC concentration could reflect discharge pulses and dilution effects driven by precipitation (Jennings et al., 2020), following
267 the characteristic seasonality of each site.

268 Seasonality plays a crucial role in fDOM predictions, as evidenced by the relationship of the Julian day driver with DOM
269 dynamics at both sites (Fig. 4). At the Irish site, DOM seasonality is primarily shaped by natural environmental processes,
270 whereas in the Spanish site, human influence plays a much greater role. This distinction helps explain why surface lake water
271 temperature and solar radiation, two variables typically linked to strong seasonal patterns, were important only for Feeagh,
272 while reservoir volume was significant only for Sau (Figures 3 and 4; Figures A9 and A10). Volume and soil variables produce
273 a similar effect on fDOM as Julian day at Sau, given that higher volumes closer to the winter season can lead to higher fDOM
274 values. Incorporating Julian day into the workflow offers a simple yet effective way to represent seasonality, potentially replac-

275 ing seasonal variables (e.g., air temperature) (see the correlation matrix of all drivers in supplementary Fig. A12). This proves
276 that the use of machine learning approaches opens up opportunities to assess diverse drivers under contrasting conditions.

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

288 **4.2 Scalability**

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

299 **4.3 Limitations and future research**

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

308 water quality monitoring programmes, an aspect that is convenient for testing workflows such as the one described (Downing
309 et al., 2012).

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

316 **5 Conclusions**

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

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

327 *Code and data availability.* All data and codes used in this study are available in this repository: https://github.com/danielmerbet/driver_
328 [attribution_fdom](https://github.com/danielmerbet/driver_attribution_fdom). A DOI will be provided in the final version of the manuscript

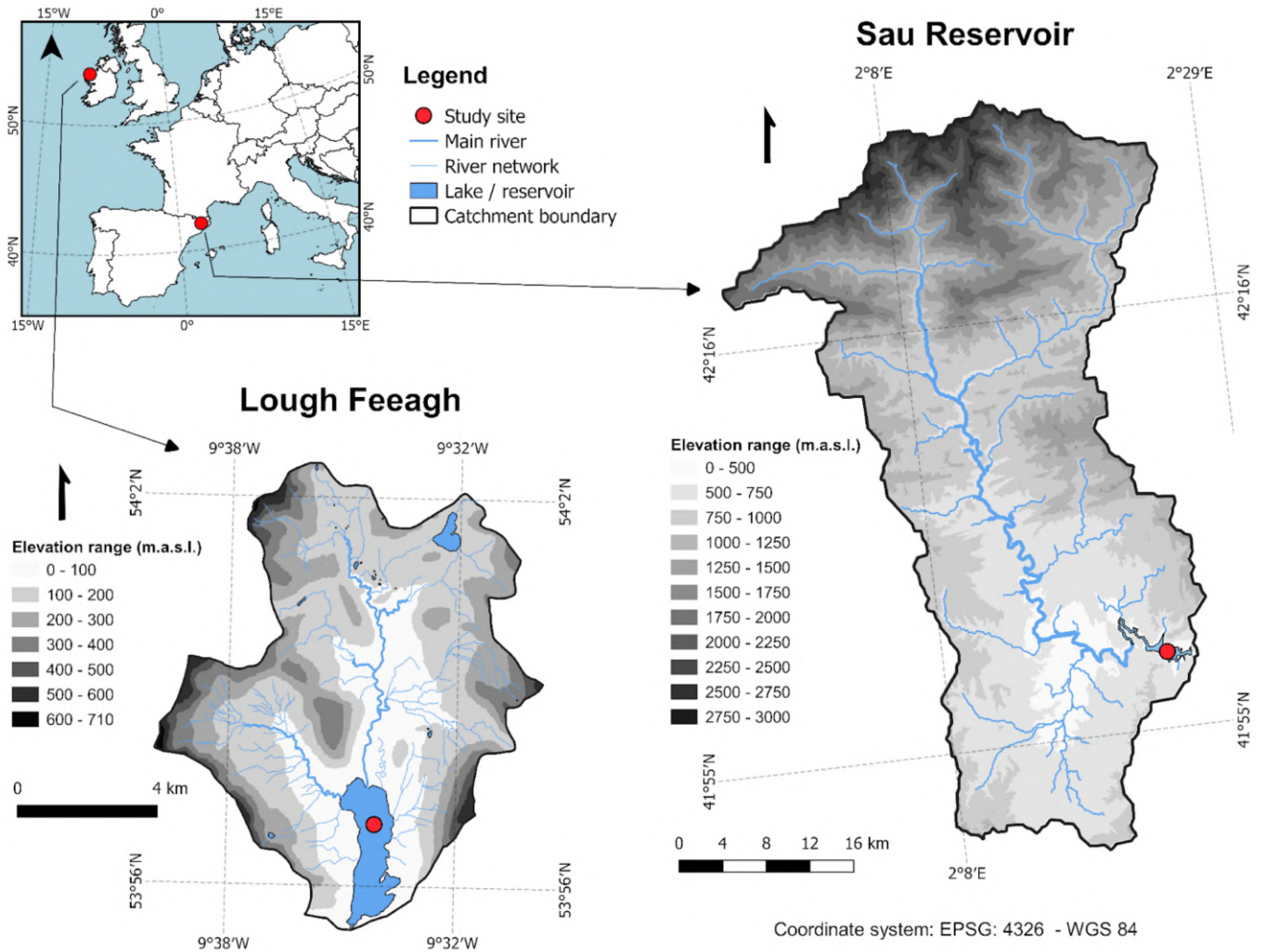


Figure A1. Elevation range for the two contrasting freshwater ecosystems: Lough Feeagh and Sau Reservoir.

329 **Appendix A: Supplementary information**

330 **Topography of Feeagh and Sau catchments**

331 Figure A1 contains the elevation range for the two contrasting freshwater ecosystems: Lough Feeagh and Sau Reservoir.

332 **CORINE land-cover classes and aggregation for both study catchments**

333 Here, we present Figure A2 listing all original CORINE Land Cover 2018 classes identified within each catchment and their
334 aggregation into the grouped categories used in Figure 1. These tables show how individual CORINE classes were consolidated
335 to represent dominant catchment characteristics while maintaining consistency with the original CORINE classification. This
336 information supports the visual interpretation presented in Figure 1.

a.

Lough Feeagh			
Code	Colour	CODE_18	CORINE Land Cover Classification
312		312	Coniferous forest
321		321	Natural grassland
322		322	Moors and heathland
324		324	Transitional woodland/shrubs
333		333	Sparsely vegetated areas
412		412	Peatbogs
511		511	Water courses
512		512	Water bodies

Land cover group	% of total area
Organic-rich soil LC	45%
Agriculture	0%
Grassland	28%
Urban	0%
Forest	25%
Watercourse	2%
	100%

b.

Sau Reservoir			
Code	Colour	CODE_18	CORINE Land Cover Classification
111		111	Continuous urban fabric
112		112	Discontinuous urban fabric
121		121	Industrial or commercial units, public services and military installations
131		131	Mineral extraction sites
142		142	Green urban areas
211		211	Non-irrigated arable land
212		212	Permanently irrigated arable land *
231		231	Pastures
242		242	Complex cultivation patterns
243		243	Land principally occupied by agriculture
311		311	Broad-leaved forest
312		312	Coniferous forest
313		313	Mixed forest
321		321	Natural grassland
322		322	Moors and heathland
323		323	Sclerophyllous vegetation*
324		324	Transitional woodland/shrubs
332		332	Bare rock
333		333	Sparsely vegetated areas
511		511	Water courses
512		512	Water bodies

Land cover group	% of total area
Organic-rich soil LC	6%
Agriculture	18%
Grassland	12%
Urban	4%
Forest	59%
Watercourse	0%
	100%

Figure A2. a. CORINE Land Cover 2018 classes identified in Feeagh and the grouped categories shown for visual clarity in the main manuscript in Figure 1; b. CORINE Land Cover 2018 classes identified in Sau and the grouped categories shown for visual clarity in the main manuscript in Figure 1.

337 **Data inspection and exploration**

338 Data inspection and exploration: the target variable (fDOM) at one of the study sites (Feeagh) exhibits low to moderate skew-
339 ness (approximately 0.5) with relatively few outliers, whereas the other study site (Sau) displays higher skewness (greater than
340 1 but below 2) and several extreme values. Zero-inflation does not appear to be an issue at either site.

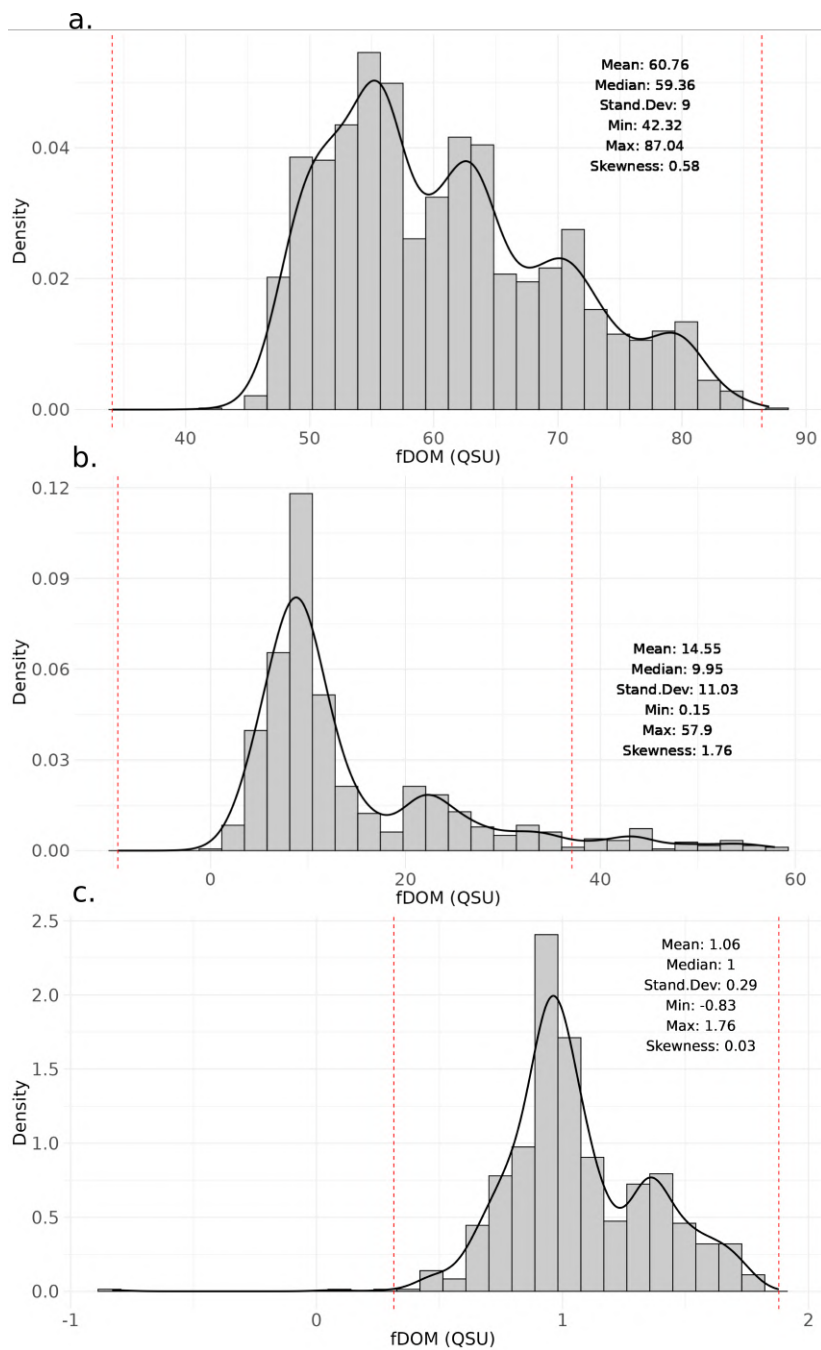


Figure A3. Exploratory data analysis for fDOM in both study sites, a. Shows the density plot of fDOM in Feeagh and b. Shows the density plot of fDOM in Sau, c. Shows the density plot of fDOM in Sau using log10 transformation. The red vertical line separates the outliers.

341 **Water temperature correction of fDOM for both study sites**

In Feeagh, fDOM data was corrected for the temperature quenching effect in previous scientific studies. The raw measurements were corrected using the compensation approach described by (Watras et al., 2011; Ryder et al., 2012). Corrected fluorescence was calculated relative to a reference temperature of 20°C using: Corrected fDOM = $\frac{\text{raw fDOM}}{1+(T-20)k}$

342 Where T is the sonde water temperature (°C) and k is the temperature correction coefficient (k = 0.0168 for Feeagh, derived
343 as the average of monthly values reported by Ryder et al. (2012)).

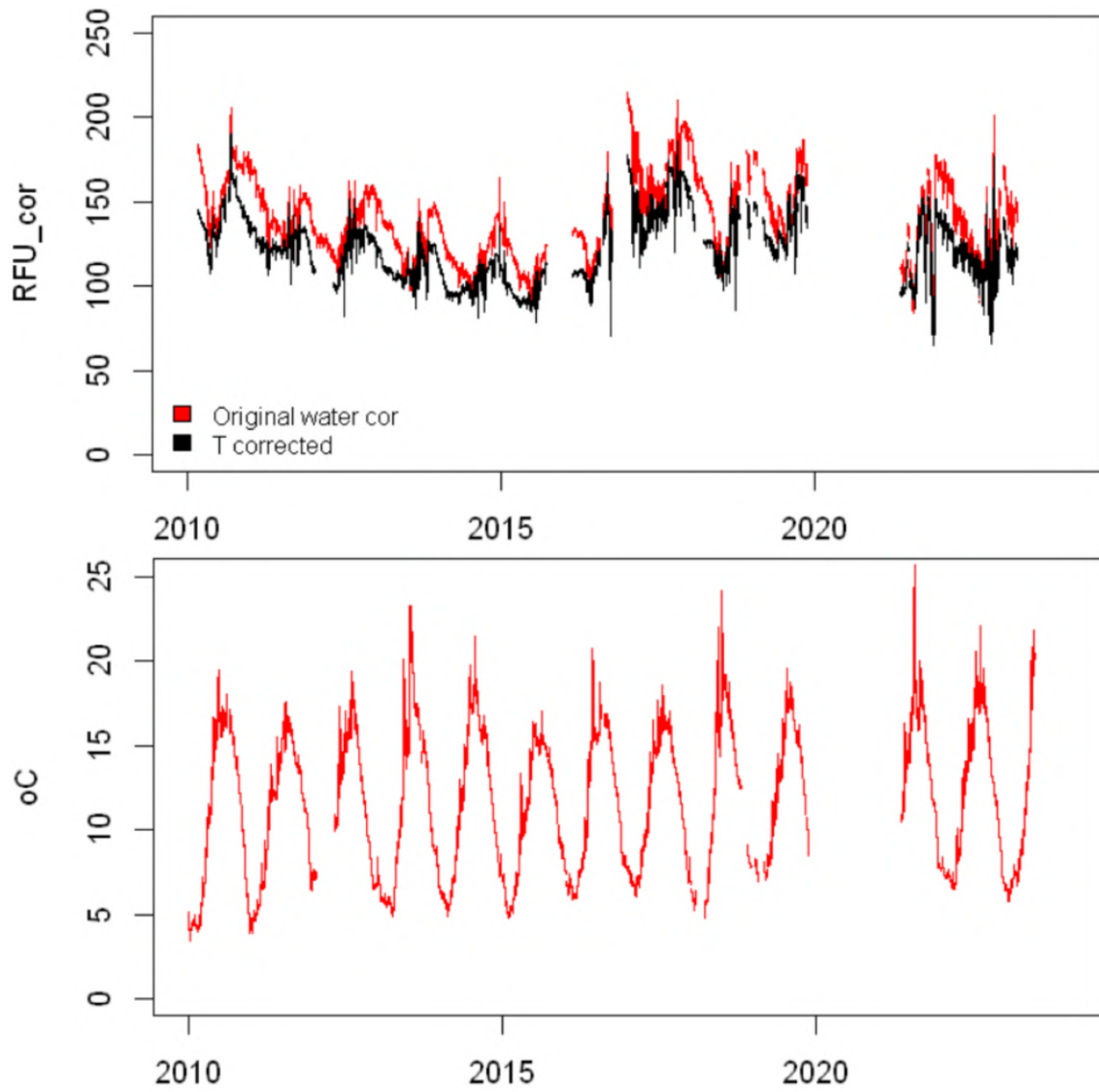


Figure A4. Upper panel: Raw (blue) and temperature-corrected (red) fDOM measurements for Feeagh; lowe panel: surface water temperature sonde measurements at Feeagh.

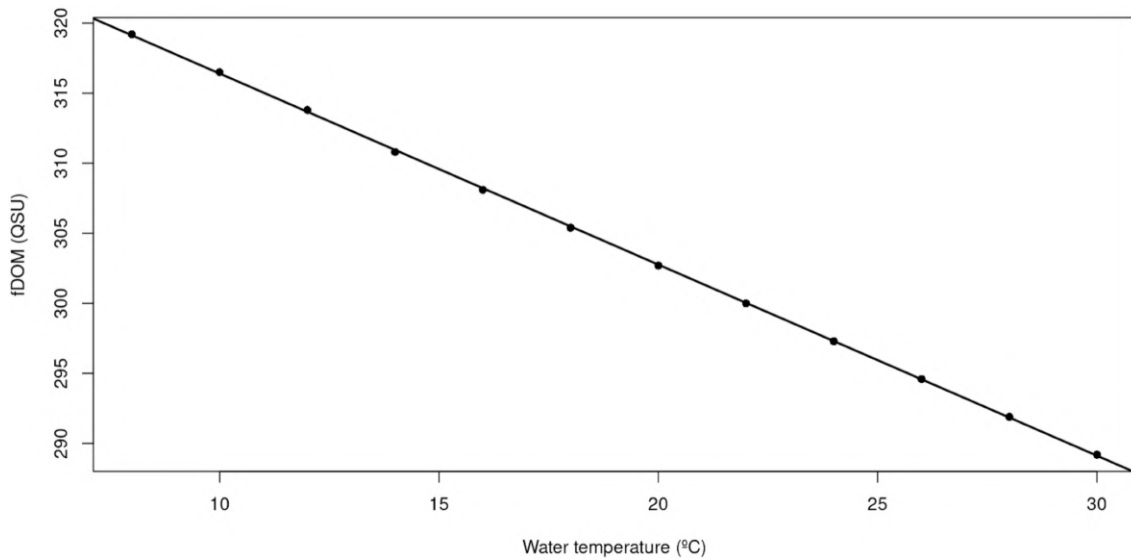


Figure A5. fDOM relation with temperature with a sample of 300 QSU provided by the manufacturer of the fDOM sensor in Sau.

344 In Sau, the fDOM data were corrected for the temperature quenching effect, following a test provided by the fDOM sensor
345 manufacturer, where they use a 300 QSU sample of water and change the temperature to get the effect of temperature in the
346 measurement, results can be found in Figure A5.

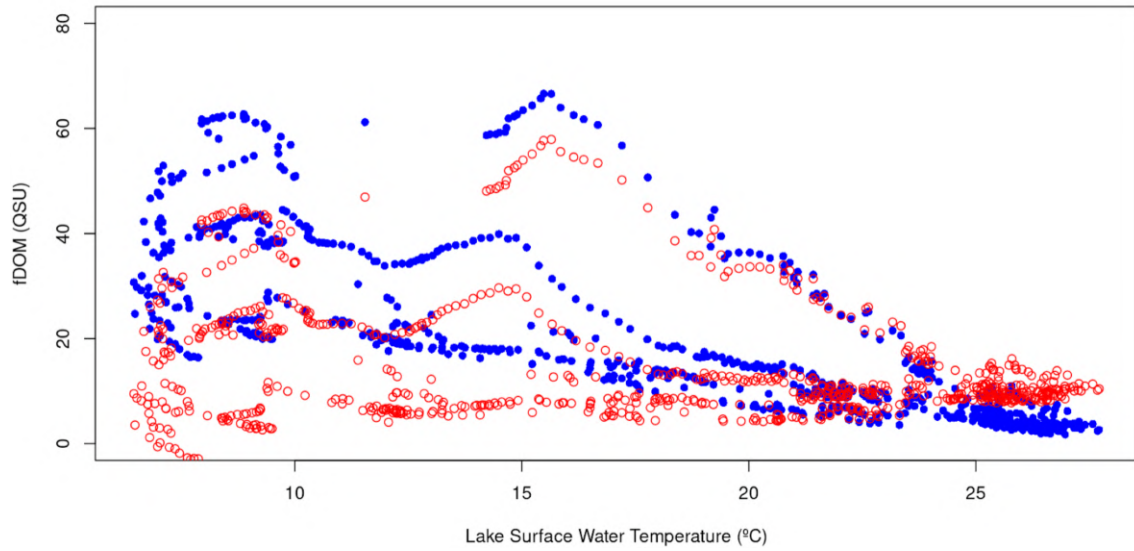


Figure A6. Uncorrected (blue) and corrected (red) fDOM data for Sau.

347 Then, fDOM data in Sau was corrected following this linear regression and surface water temperature on the lake. Figure
348 A6 presents the uncorrected values in blue and corrected values in red (8 negative values were removed from the total sample
349 of 777)

350 **Hydrologic Modelling**

351 Daily time series of inflow discharge and inflow DOC concentration into each site were generated using the Generalised Wa-
352 tershed Loading Functions Model (GWLF) coupled with a DOC module (GWLF-DOC). This model simulates catchment
353 hydrology (water balance and water distribution among the different hydrological pathways) and DOC dynamics (DOC pro-
354 duction and DOC washout) in a daily time step. The model input requirements include daily time series of two meteorological
355 variables: total precipitation and air temperature; as well as land cover, land use, and soil characterisation.

356 GWLF-DOC was applied to Feeagh based on previous model applications in the Irish catchment (Paíz et al., 2025a), for
357 which measured discharge data were used to calibrate and validate the hydrology (2013-2018 and 2019-2023, respectively),
358 and DOC concentration data were used to calibrate the DOC module (2016-2023). In Sau, observed inflow discharge data
359 were used to calibrate and validate the hydrology (2008-2011 and 2011-2024, respectively), and measured DOC concentration
360 data were used to calibrate and validate the DOC module (2008-2014 and 2016-2018, respectively) using the same calibration
361 strategy than for the Irish site. Calibration results were satisfactory for both hydrology (Feeagh: $R^2 = 0.64$ and $NSE = 0.64$;
362 Sau: $R^2 = 0.66$ and $NSE = 0.66$;) and DOC (Feeagh: $R^2 = 0.45$ and $NSE = 0.47$; Sau: $R^2 = 0.44$ and $NSE = 0.40$). Similarly,
363 validation results were satisfactory for both hydrology (Feeagh: $R^2 = 0.60$ and $NSE = 0.60$; Sau: $R^2 = 0.42$ and $NSE = 0.42$)
364 and DOC in the case of Sau ($R^2 = 0.50$ and $NSE = 0.46$).

365 **Lake Modelling**

366 Daily time series of 5 key lake variables (see Fig. 2) were obtained from the General Lake Model (GLM) run for each site. GLM
367 is an open-source, one-dimensional hydrodynamic model designed to simulate the vertical stratification and water balance of
368 lakes and reservoirs. It calculates vertical profiles of temperature, and density by accounting for factors such as inflows and
369 outflows, mixing processes, and surface heating and cooling (Hipsey et al., 2019). GLM was calibrated and validated by
370 evaluating the fit of modelled water temperature against measured water temperature profile data in Feeagh (2010-2015 and
371 2016-2017, respectively) and Sau (1997-2007 and 2008-2018, respectively). The calibration strategy was based on previous
372 lake modelling deployments at each site (Mercado-Bettín et al., 2021; Paíz et al., 2025b). Model performance was satisfactory
373 for both sites.

374 **Comparison of validation methods using Random Forest**

375 To pick the most suitable validation method, we implemented hold-out period method used in the main manuscript, k-fold
376 cross-validation method using $k=5$, and rolling window cross-validation using training size of two year, testing size of one
377 year, and a shift window every 90 days. Results are shown in Figure A7.

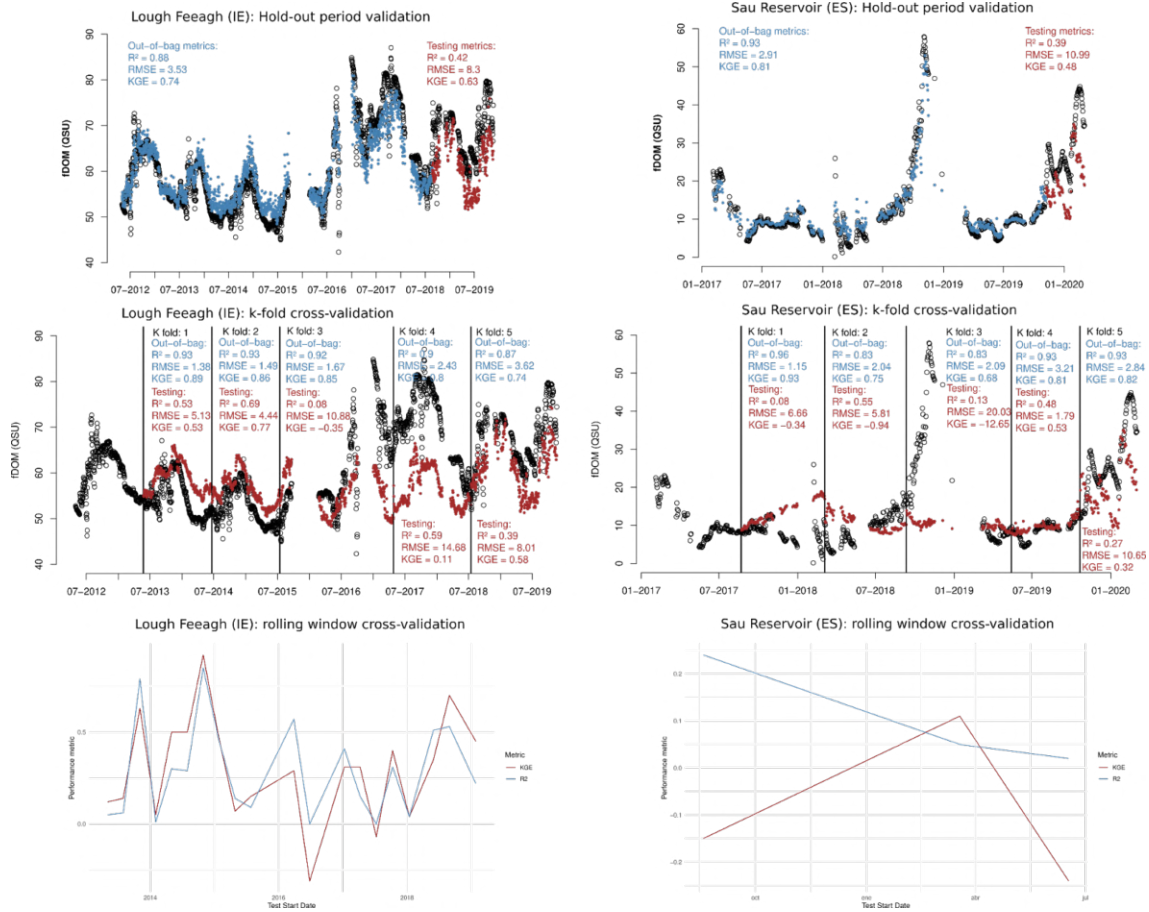


Figure A7. Comparison of validation methods using random forest for both study sites: hold-out period method used in the main manuscript, k-fold cross-validation method using k=5, and rolling window cross-validation using training size of two year, testing size of one year, and a shift window every 90 days. For this method, in the case of Sau Reservoir it is not possible to get a clear analysis due to the limited data.

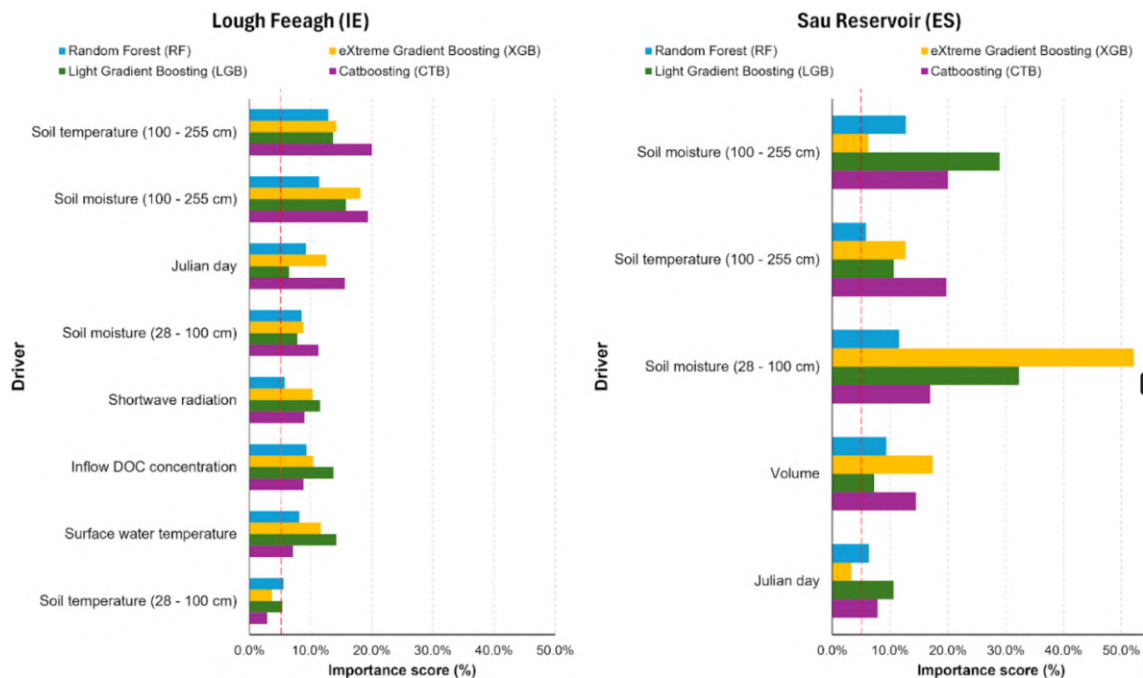


Figure A8. Selection of influential drivers for predicting fDOM. Twenty-four drivers from multiple sources were used to train the machine-learning models, including climate variables, soil variables, hydrologic and water-quality model outputs, lake model outputs, and cosine-transformed Julian day. Feature importance is shown for four models Random Forest, Light Gradient Boosting, eXtreme Gradient Boosting and Catboosting, and drivers exceeding 5% importance are highlighted.

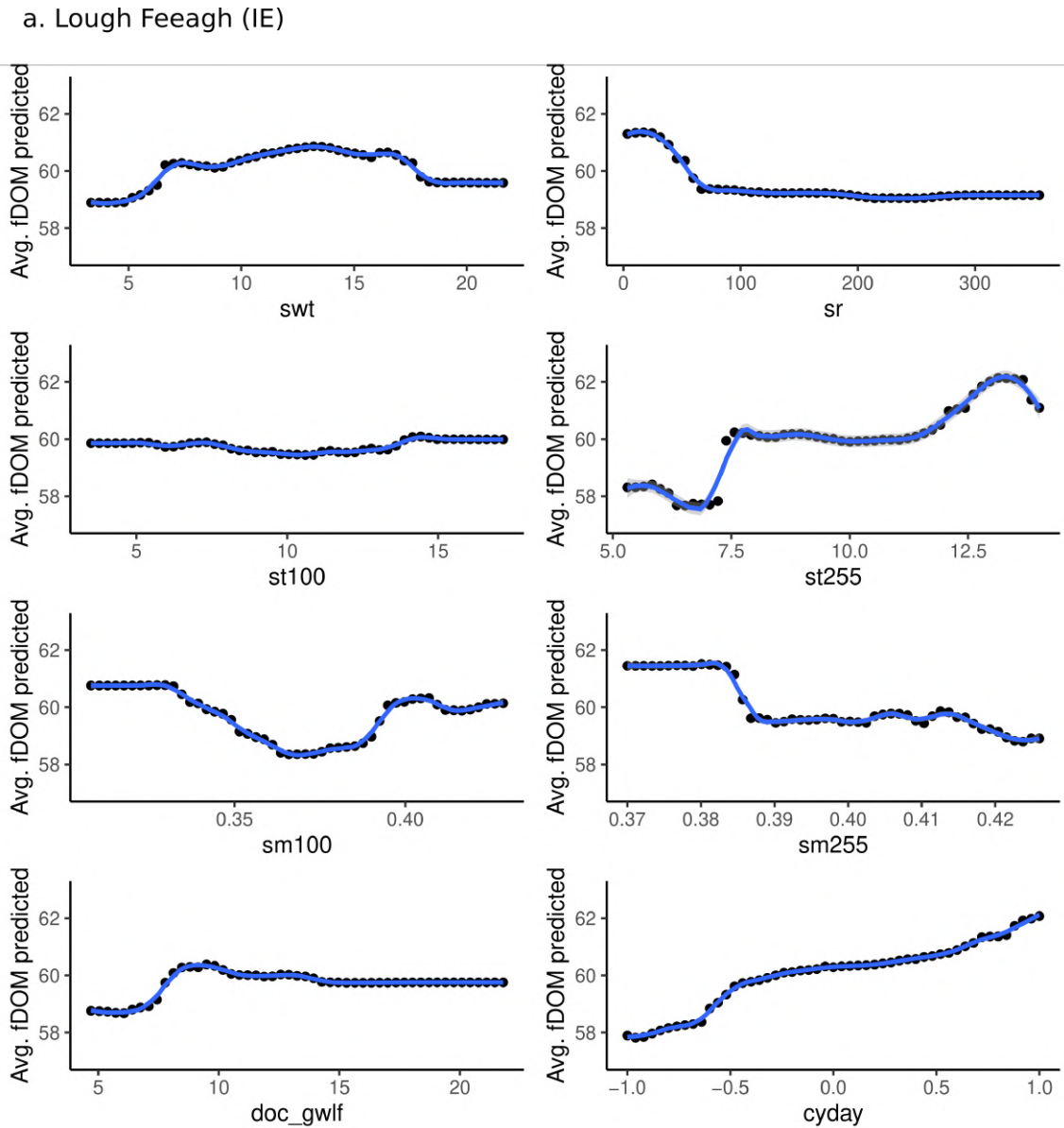


Figure A9. Partial dependence plots for the most influential drivers (feature importance > 5%, estimated using node purity) based on Random Forest model outputs for Lough Feeagh. Cosine-transformed Julian day (cyday) values are shown for seasonality; approximate calendar timing corresponds to $\cos(\text{Julian day}) = 1$ (winter), 0 (spring/autumn), and -1 (summer).

b. Sau Reservoir (ES)

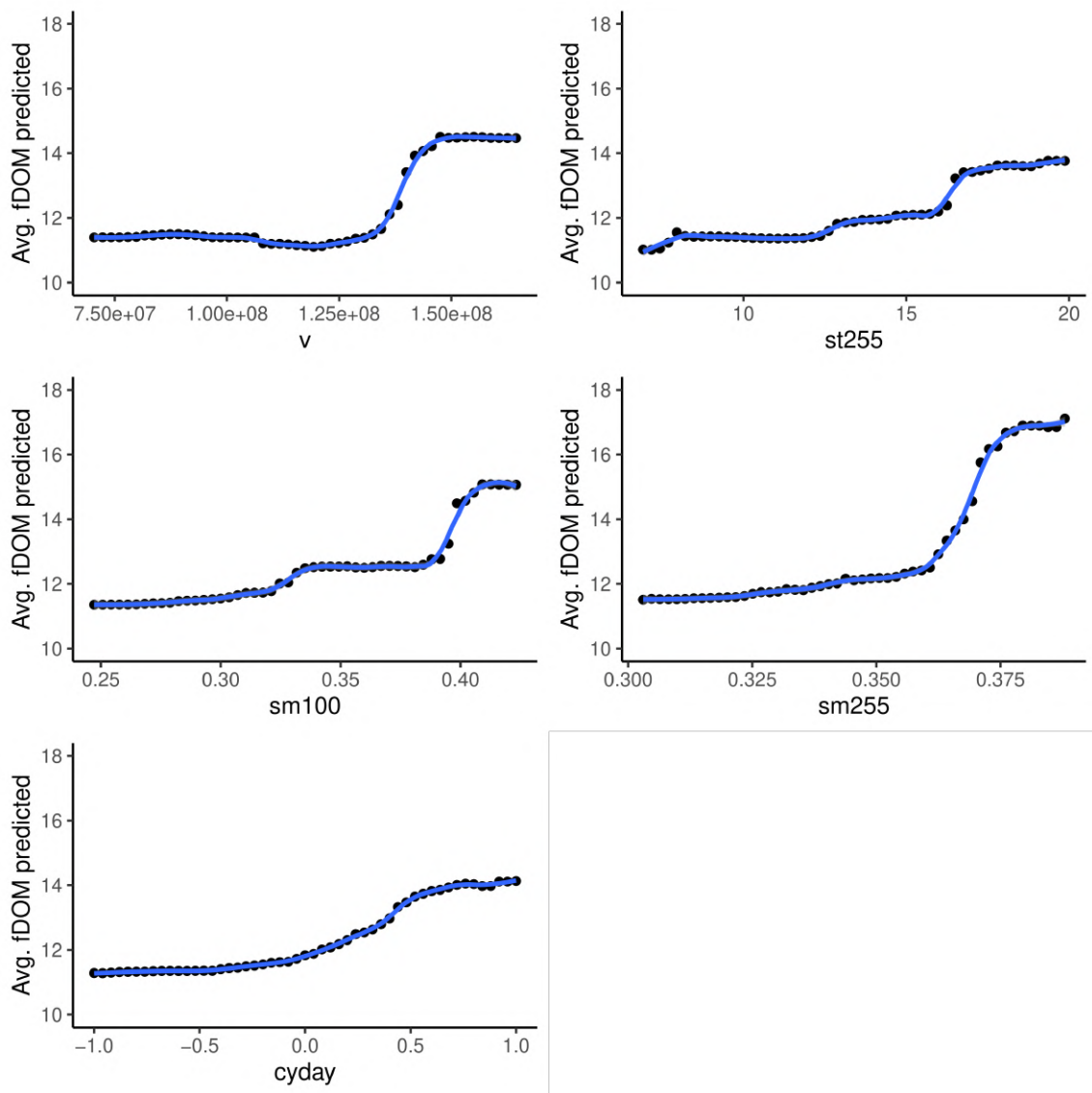


Figure A10. Partial dependence plots for the most influential drivers (feature importance > 5%, estimated using node purity) based on Random Forest model outputs for Sau Reservoir. Cosine-transformed Julian day (cyday) values are shown for seasonality; approximate calendar timing corresponds to $\cos(\text{Julian day}) = 1$ (winter), 0 (spring/autumn), and -1 (summer).

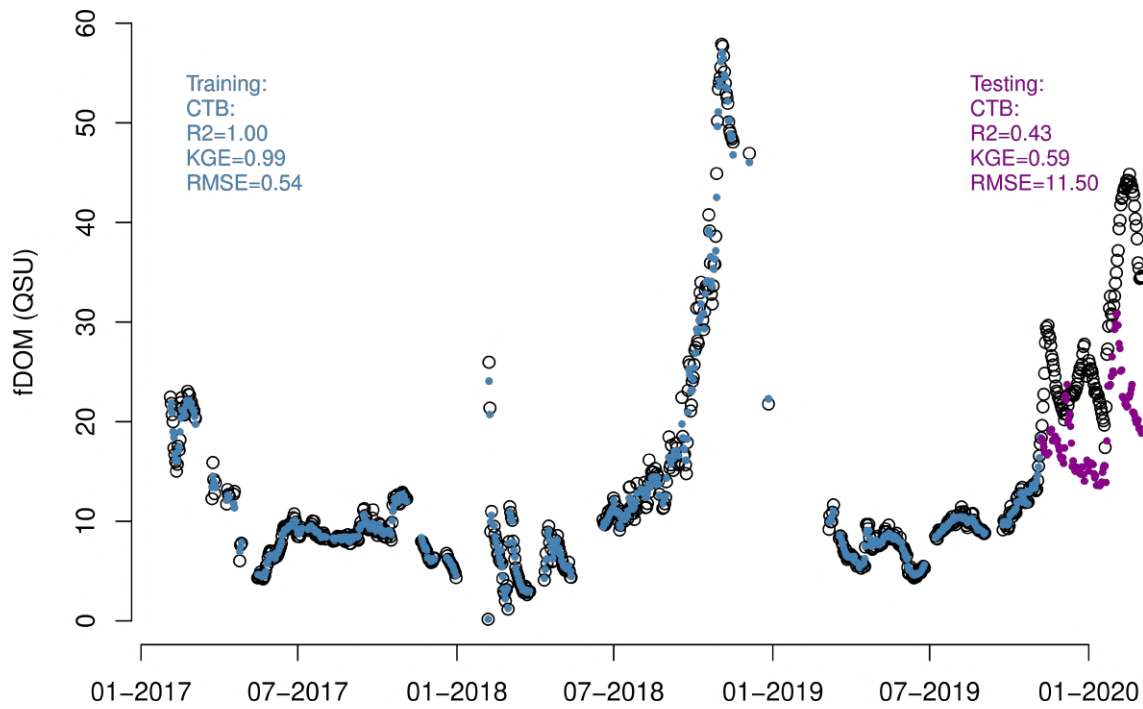


Figure A11. fDOM predictions obtained from the boosting (CatBoost) model using the most influential drivers with initial log10 transformation of the fDOM data.

381 **Correlation matrix of all drivers and fDOM for Feeagh and Sau.**

Correlation matrix of all drivers is presented in Figure A12.

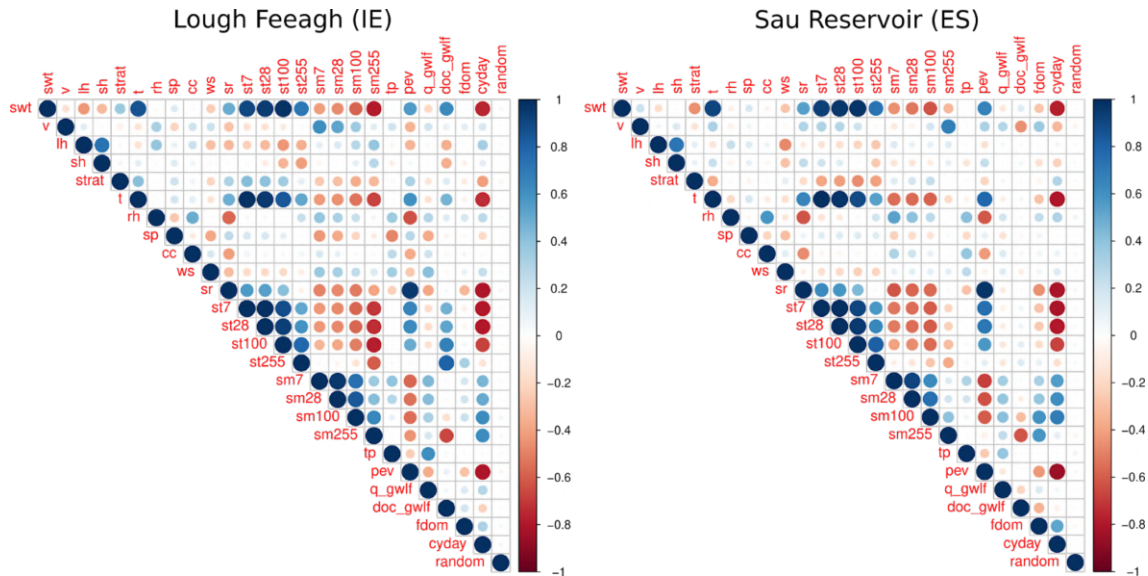


Figure A12. Correlation matrix of all drivers and fDOM for Feeagh and Sau.

382

383 **Model comparison to predict fDOM using the most influential drivers in both study sites.**

384 Resulting metrics during testing and training periods for all models are shown in Table A1 and A2, respectively.

Table A1. Model comparison during testing 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 Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting (LGB), Catboosting (CTB), k-Nearest Neighbors (KNN), Support Vector Regression (SVR) and linear model during training period.

	Lough Feeagh							Sau Reservoir						
Model/Metric	RF	XGB	LGB	CTB	LM	KNN	SVR	RF	XGB	LGB	CTB	LM	KNN	SVR
R^2	0.37	0.40	0.40	0.50	0.40	0.36	0.34	0.53	0.09	0.25	0.54	0.45	0.33	0.63
RMSE	9.07	9.28	8.97	8.17	8.52	9.85	10.54	8.51	12.22	11.35	7.11	6.58	9.23	17.05
KGE	0.54	0.51	0.50	0.68	0.55	0.12	0.30	0.55	0.25	0.22	0.66	0.31	0.43	-0.82

Table A2. Model comparison during training 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 Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting (LGB), Catboosting (CTB), k-Nearest Neighbors (KNN), Support Vector Regression (SVR) and linear model during training period.

	Lough Feeagh							Sau Reservoir						
Model/Metric	RF	XGB	LGB	CTB	LM	KNN	SVR	RF	XGB	LGB	CTB	LM	KNN	SVR
R^2	0.99	1.00	0.99	0.75	0.22	0.99	0.89	0.99	1.00	1.00	0.99	0.67	0.98	0.98
RMSE (QSU)	0.97	0.00	0.86	5.04	7.87	0.92	2.95	0.81	0.01	0.17	0.72	5.43	1.17	1.38
KGE	0.95	1.00	0.98	0.58	0.25	0.98	0.90	0.97	1.00	1.00	0.99	0.75	0.99	0.98

385 **Model comparison to predict fDOM using the most influential drivers in Sau with previous application of log10**
386 **transformation.**

387 Tables A3 and A4 presents the testing and training performance metrics results for Sau site with previous application of log10
388 transformation.

Table A3. Model comparison during testing to predict fDOM using the most influential drivers in Sau with previous application of log10 transformation. Statistic metrics (R^2 , RMSE and KGE) were calculated to compare the performance of the models Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting (LGB), Catboosting (CTB), k-Nearest Neighbors (KNN), Support Vector Regression (SVR) and linear model during training period.

Sau Reservoir							
Model/Metric	RF	XGB	LGB	CTB	LM	KNN	SVR
R^2	0.51	0.18	0.22	0.43	0.40	0.26	0.58
RMSE (QSU)	14.24	16.56	13.07	11.50	10.91	9.75	35.23
KGE	0.63	0.38	0.34	0.59	0.36	0.32	-1.35

Table A4. Model comparison during training to predict fDOM using the most influential drivers in Sau with previous application of log10 transformation. Statistic metrics (R^2 , RMSE and KGE) were calculated to compare the performance of the models Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting (LGB), Catboosting (CTB), k-Nearest Neighbors (KNN), Support Vector Regression (SVR) and linear model during training period.

Sau Reservoir							
Model/Metric	RF	XGB	LGB	CTB	LM	KNN	SVR
R^2	0.98	1.00	0.99	1.00	0.58	0.94	0.87
RMSE (QSU)	1.05	0.07	0.74	0.54	4.65	1.22	1.39
KGE	0.94	1.00	0.98	0.99	0.66	0.94	0.89

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

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

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

397 **References**

- 398 Asadollah, S. B. H. S., Safaeinia, A., Jarahizadeh, S., Alcalá, F. J., Sharafati, A., and Jodar-Abellan, A.: Dissolved organic carbon estimation
399 in lakes: Improving machine learning with data augmentation on fusion of multi-sensor remote sensing observations, *Water Research*,
400 277, 123 350, 2025.
- 401 Bhateria, R. and Jain, D.: Water quality assessment of lake water: A review, *Sustainable Water Resources Management*, 2, 161–173,
402 <https://doi.org/10.1007/s40899-015-0014-7>, 2016.
- 403 Biau, G. and Scornet, E.: A random forest guided tour, *Test*, 25, 197–227, 2016.
- 404 Breiman, L.: Random Forests, *Machine Learning*, 45, 5–32, <https://doi.org/10.1023/A:1010933404324>, 2001.
- 405 Chen, C., Chen, Q., Yao, S., He, M., Zhang, J., Li, G., and Lin, Y.: Combining physical-based model and machine
406 learning to forecast chlorophyll-a concentration in freshwater lakes, *Science of The Total Environment*, 907, 168 097,
407 <https://doi.org/10.1016/j.scitotenv.2023.168097>, 2024.
- 408 Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, in: *Proceedings of the 22nd ACM SIGKDD International Conference*
409 *on Knowledge Discovery and Data Mining*, pp. 785–794, <https://doi.org/10.1145/2939672.2939785>, 2016.
- 410 Cortes, C. and Vapnik, V.: Support-vector networks, *Machine Learning*, 20, 273–297, <https://doi.org/10.1007/BF00994018>, 1995.
- 411 Creed, I. F., Bergström, A.-K., Trick, C. G., Grimm, N. B., Hessen, D. O., Karlsson, J., Kidd, K. A., Kritzbeg, E., McKnight, D. M.,
412 Freeman, E. C., Senar, O. E., Andersson, A., Ask, J., Berggren, M., Cherif, M., Giesler, R., Hotchkiss, E. R., Kortelainen, P., Palta, M. M.,
413 and Weyhenmeyer, G. A.: Global change-driven effects on dissolved organic matter composition: Implications for food webs of northern
414 lakes, *Global Change Biology*, 24, 3692–3714, <https://doi.org/10.1111/gcb.14129>, 2018.
- 415 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
416 materials, and temperature on in situ measurements of DOM fluorescence in rivers and streams, *Limnology and Oceanography: Methods*,
417 10, 767–775, <https://doi.org/10.4319/lom.2012.10.767>, 2012.
- 418 Doyle, B. C., de Eyto, E., Dillane, M., Poole, R., McCarthy, V., Ryder, E., and Jennings, E.: Synchrony in catchment stream colour levels is
419 driven by both local and regional climate, *Biogeosciences*, 16, 1053–1071, <https://doi.org/10.5194/bg-16-1053-2019>, 2019.
- 420 Duan, H., Cao, Z., Luo, J., and Shen, M.: AI-driven opportunities and challenges in lake remote sensing, *Information Geography*, p. 100014,
421 <https://doi.org/10.1016/j.infgeo.2025.100014>, 2025.
- 422 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.,
423 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>,
424 2024.
- 425 Fix, E.: *Discriminatory Analysis: Nonparametric Discrimination, Consistency Properties*, Tech. rep., USAF School of Aviation Medicine,
426 1985.
- 427 Gobler, C. J.: Climate Change and Harmful Algal Blooms: Insights and perspective, *Harmful Algae*, 91, 101 731,
428 <https://doi.org/10.1016/j.hal.2019.101731>, 2020.
- 429 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,
430 V.: Predicting lake surface water phosphorus dynamics using process-guided machine learning, *Ecological Modelling*, 430, 109 136,
431 <https://doi.org/10.1016/j.ecolmodel.2020.109136>, 2020.
- 432 Harkort, L. and Duan, Z.: Estimation of dissolved organic carbon from inland waters at a large scale using satellite data and machine learning
433 methods, *Water Research*, 229, 119 478, <https://doi.org/10.1016/j.watres.2022.119478>, 2023.

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

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

441 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.,
442 and Winslow, L. A.: A General Lake Model (GLM 3.0) for linking with high-frequency sensor data from the Global Lake Ecological
443 Observatory Network (GLEON), *Geoscientific Model Development*, 12, 473–523, <https://doi.org/10.5194/gmd-12-473-2019>, 2019.

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

446 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
447 Highs to Lows: Changes in Dissolved Organic Carbon in a Peatland Catchment and Lake Following Extreme Flow Events, *Water*, 12,
448 2843, <https://doi.org/10.3390/w12102843>, 2020.

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

451 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
452 tree, in: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 3149–3157, 2017.

453 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
454 Change and the Biodiversity of Freshwater Ecosystems: Impacts on Linkages between Above-Sediment and Sediment Biota, *BioScience*,
455 50, 1099–1107, [https://doi.org/10.1641/0006-3568\(2000\)050\[1099:GCATBO\]2.0.CO;2](https://doi.org/10.1641/0006-3568(2000)050[1099:GCATBO]2.0.CO;2), 2000.

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

459 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
460 level: A case study of the Poyang Lake in China, *Hydrology Research*, 47, 69–83, <https://doi.org/10.2166/nh.2016.264>, 2016.

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

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

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

468 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.,
469 Doubek, J. P., et al.: Dynamic modeling of organic carbon fates in lake ecosystems, *Ecological Modelling*, 386, 71–82, 2018.

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

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

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

477 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
478 Lu, Y.: Machine-learning based approach to examine ecological processes influencing the diversity of riverine dissolved organic matter
479 composition, *Frontiers in Water*, 6, <https://doi.org/10.3389/frwa.2024.1379284>, 2024.

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

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

486 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,
487 S., and Jennings, E.: Near-term lake water temperature forecasts can be used to anticipate the ecological dynamics of freshwater species,
488 *Ecosphere*, 16, e70 335, <https://doi.org/10.1002/ecs2.70335>, 2025b.

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

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

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

494 Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., and Chica-Rivas, M.: Machine learning predictive models for mineral
495 prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines, *Ore Geology Reviews*,
496 71, 804–818, <https://doi.org/10.1016/j.oregeorev.2015.01.001>, 2015.

497 Rumpel, C. and Kögel-Knabner, I.: Deep soil organic matter—A key but poorly understood component of terrestrial C cycle, *Plant and Soil*,
498 338, 143–158, <https://doi.org/10.1007/s11104-010-0391-5>, 2011.

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

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

504 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,
505 J. E.: Ecosystem consequences of changing inputs of terrestrial dissolved organic matter to lakes: Current knowledge and future challenges,
506 *Ecosystems*, 18, 376–389, <https://doi.org/10.1007/s10021-015-9848-y>, 2015.

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

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

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

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

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

518 Xenopoulos, M. A., Barnes, R. T., Boodoo, K. S., Butman, D., Catalán, N., D’Amario, S. C., Fasching, C., Kothawala, D. N., Pisani,
519 O., Solomon, C. T., Spencer, R. G. M., Williams, C. J., and Wilson, H. F.: How humans alter dissolved organic matter composition
520 in freshwater: Relevance for the Earth’s biogeochemistry, *Biogeochemistry*, 154, 323–348, <https://doi.org/10.1007/s10533-021-00753-3>,
521 2021.

522 Zhang, D., Shi, K., Wang, W., Wang, X., Zhang, Y., Qin, B., Zhu, M., Dong, B., and Zhang, Y.: An optical mechanism-based
523 deep learning approach for deriving water trophic state of China’s lakes from Landsat images, *Water Research*, 252, 121181,
524 <https://doi.org/10.1016/j.watres.2024.121181>, 2024.

525 Zhang, Y., Yao, X., Wu, Q., Huang, Y., Zhou, Z., Yang, J., and Liu, X.: Turbidity prediction of lake-type raw water using random
526 forest model based on meteorological data: A case study of Tai lake, China, *Journal of Environmental Management*, 290, 112657,
527 <https://doi.org/10.1016/j.jenvman.2021.112657>, 2021.