History of Watersheds

Jakob Nyström
MSc Data Science - Machine Learning and Statistics
jakob.nystrom.5563@student.uu.se

Muhammad Faran Khalid
MSc Data Science - Machine Learning and Statistics
muhammad-faran.khalid.0898@student.uu.se

Venugopal Srinivas
MSc Data Science - Machine Learning and Statistics
venugopal.srinivas.1001@student.uu.se

***Abstract*—The objectives of this project was to look for trends in the historical watershed data of routinely monitored lakes in Sweden. The research questions we tried to answer were the shift in the Total Organic Concentrations (TOC) with changing climate, hydrology, temperature, land use and lake nutrients. We developed different models to try and explain how the variables in the dataset influence the TOC concentrations of the lakes. In terms of TOC levels, organic N, metals (Mg, Fe, Si), forest and wetland in the catchment area had the largest positive coefficients; P, SO4 and pH had the largest negative impact. Marginal trends were found in the spatial models, it would have been interesting if we had access to data for more than 110 lakes. No significant trends were found in the hydrograph trend analysis for the lakes. ARIMA forecasting for Waterflow resulted in low variance due to lack of enough historical data.**

***Index Terms*—Organic carbon, boreal lakes, mixed effects models, spatial models, general additive models, ARIMA.**

I. Introduction

Organic matter is the decomposing material from plant and animal life, typically found in large quantities in soils and wetlands. Organic matter is also found in high concentrations in many boreal lakes.[1] An important subset of this is total organic carbon (TOC), which comprises dissolved and non-dissolved carbon particles. Most lakes are net contributors of CO2 to the atmosphere; they therefore play an important role in the global CO2 cycle.

The level of TOC in a given lake can exhibit high variability over time, and differences between lakes can be large. In this study, we examine how water chemistry, weather and water flows, physical properties and the catchment area, influence TOC concentrations, based on panel data from 2001-2022. In addition, our study includes hydrological trend analysis across lakes, focused on total waterflow in the years 2010-2021, and time series predictions for 2023-2025. Our research questions can be formulated as:

- What are the factors associated with differences in TOC concentrations within and across lakes?
- Are there any spatial patterns in how and which variables influence the TOC concentrations in lakes?
- Are the lakes and catchments getting wetter or dryer over time?

.

[1]Lakes situated in sub-arctic climates.

To answer these questions, we deploy and evaluate a few different models: linear mixed effects, spatial models, generalized additive models and ARIMA. Section II provides a brief overview of previous research related to this topic, followed by a description of our methodologies in section III. Results of the data analysis and different modelling approaches are found in section IV, and the results are discussed in section V. The code for this project can be found in this GitHub repository. The authors of each section and subsection are specified in the list below:

- Jakob Nyström: I; II (excl. last paragraph); III A (LME); IV A.1-2 (chemistry, weather data); IV C.1 (LME); V (LME).
- Muhammad Faran Khalid: II (hydrology); III C-D; (GAM, ARIMA); IV A.4 (hydrology data); IV C.3-5 (GAM, hydrographs, ARIMA); V (GAM, ARIMA).
- Venugopal Srinivas: Abstract; III B (GWR, MGWR, spatial lag; IV A.3 (catchment data); IV C.2 (GWR, MGWR, spatial lag); V (spatial models).

II. Related work

Several studies ([1], [2], [3]) have looked at how aspects of organic matter in lakes are related to lake chemistry and the surrounding environment. At the same time, data on lake chemistry from the Swedish lake monitoring program [4] provides a unique opportunity to deeper analyze TOC and water chemistry over time. Based on these studies, and domain knowledge provided by the supervisors, we outline hypotheses about how TOC relates to other attributes below.

Nitrogen (N) and phosphorus (P) are nutrients and expected to correlate positively with TOC, either directly or as nutrients for microorganisms. Silicon (Si) can also be a nutrient for primary producers. Iron (Fe) binds to TOC, and should thus be positively correlated. Sulfate (SO4) stemming from acid rains have been correlated with TOC, but the relationship is complex.

In terms of catchment area land use-land cover, wetlands [3], [1] and forests contain lots of organic matter, that should contribute to higher TOC concentrations. Agriculture should contribute too, due to the use of fertilizer. Lakes in mountain areas generally contain less organic matter. Hence altitude [3], latitude and longitude could also influence TOC. Organic matter has also been linked to temperature [1]. Precipitation

can, on the one hand, lead to an inflow of organic matter from the catchment, but on the other hand have a dilutive effect [3].

Spring and summer should imply higher inflow of organic matter from the catchment, but also higher precipitation dilution and more outflow, so the net effect is not clear. Temperature has been found to positively correlate with TOC [3].

In terms of hydrology, [5] looks into hydrological behavior of various environments and discusses how minerotrophic fens have been affected by increasing water levels, which can lead to shallow lakes. The paper also underlines some statistical techniques of using hydrographs to understand hydrological trends of water bodies.

III. Methodology

A. Mixed effects models

A linear mixed effects model (LME) is a regression framework suitable for panel data, in our case repeated measurements over time at different locations; this means we need to deal with observations that are not i.i.d.[2] In an LME, we estimate population level coefficients, called fixed effects (FE), and group level coefficients, called random effects (RE) [6]–[8]. The FE represent averages effects across all lakes, while RE are specific to each lake; called random since they are drawn from multivariate Gaussian distributions derived from the data. The model for a lake i can be expressed as

$$Y_i = X_i\beta + Z_i\gamma_i + \epsilon_i = \mathcal{N}(Y_i; X_i\beta + Z_i\gamma_i, R_i) \quad (1)$$

where Y_i is the response variable vector; X_i is the FE design matrix; β is a vector of FE coefficients; Z_i is the RE design matrix; γ_i is the corresponding random coefficient vector with mean zero and covariance matrix D_i; and ϵ_i is a vector of i.i.d. and zero-mean normal errors, assumed to be independent both within and between lakes. β, D_i and R_i are estimated from the data using (restricted) maximum likelihood, while γ_i and ϵ_i are randomly sampled [9]. The design matrices X_i and Z_i can fully or partially overlap, but do not need to. Each lake will have its own random intercept, and depending the choice of Z_i, random slopes as well. The conditional distribution of Y_i is assumed to be Gaussian, why it is important to evaluate the residuals ϵ_i in the analysis.

LMEs are available for `Python` in the `statsmodels` package [9]. Compared to its `R` counterparts and libraries like `sklearn`, the `statsmodels` API is less mature. We therefore need to build a test and validation pipeline from the ground up, with custom functions to perform e.g. train-test split, predictions and sequential feature selection.

B. Spatial models

Spatial models are an extension of regression models where the non-stationarity in relationships being measured is allowed to vary spatially. To model spatial regression *spreg* [10] library in R language was used. To study the spatial lag models, Moran Plot and Moran's I tools supported by PySAL library in python [11] is used.

1) Geographically weighted regression:

$$y_i = \beta_{i0} + \sum_{k=1}^{p} \beta_{ik}x_{ik} + \epsilon_i \quad (2)$$

where the terms in the equation are akin to regular linear regression with additional location information embedded in i [12] [13]. A moving kernel window approach is used to calibrate the model spatially. Each observation is given a weight which is a monotonically decreasing function of the distance.

Multiscale geographically weighted regression (MGWR) is an extension of GWR where the neighborhood (bandwidth) of a explanatory variable is not same at all locations and vary independently. Here, neighborhood infers to both physical space and parameter space.

$$y = \sum_{j=1}^{k} f_j + \epsilon \quad (3)$$

where f_j is a smoothing function applied to the j-th explanatory variable that may be characterized by distinct bandwidth parameter [14]. Each variable is associated with a distinct bandwidth by recasting GWR as a Generalized Additive Model (GAM) [13].

2) Spatial lag model (WX):

$$y = X\beta + WX\gamma + \epsilon \quad (4)$$

The matrix multiplication of the weights W with a covariate is called a spatial lag [15]. Spatial lag can get introduced into the regression model through predictor variable(WX), response variable(Wy) or an error term. Spatially lagged terms are accompanied by parameters for the lags. In this case γ is a vector of parameters indicative of the lags from the predictor variable side, hence called Spatial Lag (WX) model. If the lag was from the response variable side then it would have been a Spatial Lag (Wy) model. Fig 35 shows the taxonomy of autoregressive models. γ indicates the strength of autocorrelation among the observations with respect to the variable in focus [15].

Moran's I score [16] indicates global autocorrelation and is given by

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} z_i z_j}{\sum_i z_i^2} \quad (5)$$

where n is the number of observations; z_i is the standardized value of the variable of interest at location i and w_{ij} is the cell corresponding to the i-th row and j-th column of W i.e. weights matrix. The graphical interpretation of this is the Moran Plot 40. It can be considered as a scatterplot in which the variable is displayed against its spatial lag.

Spatial lag was introduced into the model by deriving another attribute from the forest cover attribute obtained from the landcover data. And lakes that had more than 50% forest

[2]Independent and identically distributed, a property assumed in many statistical models such as standard linear regression.

cover were given a value '1' and '0' if less than 50%. This derived attribute w_forest is the spatial lag term. When evaluating weight matrix, the lakes with more than 50% forest cover were chosen as neighbors. Spatial Lag models are an extension of Ordinary Least Squares (OLS) regression models [15] but unlike in OLS where the coefficients are considered to be constant, it is estimated by weighted least squares. More weights are given to nearest neighbors in comparison to distant neighbors [17].

C. Generalized additive models

Generalized additive models (GAMs) are flexible models used in statistical modelling as they also allow for non-linear relationships between independent variables and the dependent variable. Unlike linear models, which assume a straight-line relationship, GAMs allows modelling of complex non-parametric relationships through smooth functions such as splines.

For our research we used the Linear Generalized Additive Model (GAM) which is a type of GAM model that extends the Generalized Linear Model (GLM) and allows non-linear functions of predictor variables while maintaining overall linearity in terms of the link function. Linear GAMs use smooth functions (like splines) to model these relationships, offering greater flexibility [18].
The Linear GAM model can be expressed as follows:

$$g(E[Y|X]) = \beta_0 + f_1(x_1) + f_2(x_2) + \ldots + f_p(x_p) \quad (6)$$

where $g(\cdot)$ is the link function, $E[Y|X]$ is the expected value of the response Y given predictors X, β_0 is the intercept, and $f_i(x_i)$ are smooth functions of the predictors.

The $f_i(x_i)$ is a spline function. It can be defined as:

$$f_i(x_i) = \sum_{k=1}^{K} \beta_{ik} B_{ik}(x_i) \quad (7)$$

where $B_{ik}(x_i)$ are the basis functions for the spline and β_{ik} are the coefficients to be estimated.
To improve the accuracy of the Linear GAM we used an ensemble technique (xGBoost) [19]. XGBoost can reduce overfitting in models by by optimizing both the loss function and a regularization term. XGBoost can also handle large number of features in an effective manner which was necessary in the context of our problem.
The xGBoost model [19] can be expressed as follows:

$$\text{Obj} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (8)$$

where l is a differentiable convex loss function that measures the difference between the predicted value $\hat{y}_i$ and the actual target y_i, and $\Omega(f_k)$ is the regularization term, which penalizes the complexity of the model. This regularization term helps in reducing overfitting.

The regularization term Ω is defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \quad (9)$$

where T is the number of leaves in the tree, w is the vector of scores on leaves, γ is the complexity control for each tree, and λ is the L2 regularization term on the leaf weights.

To implement Linear GAM we used the `pygam` library [20] in `Python` as it is the most commonly used library for implementing GAM models.

D. AutoRegressive Integrated Moving Average

ARIMA is a statistical model commonly used for forecasting time series data [21]. It has three main components with three specialized parameters. AutoRegressive (AR) it represents relationship between an observation and a number of lagged observations. It is represented by parameter p. Integrated (I) this is a method for making non-stationary data stationary which is vital when using ARIMA for time series analysis and forecasting. It is represented by parameter d. Finally, Moving Average (MA) this captures relationship between an observation and a residual error from a moving average. It is represented by parameter q.

Overall the ARIMA model can be represented as:

$$(p, d, q) : (1 - \sum_{i=1}^{p} \phi_i L^i)\nabla^d Y_t = (1 + \sum_{i=1}^{q} \theta_i L^i)\epsilon_t \quad (10)$$

where the variables and parameters are defined as:

- ϕ_i: Coefficients of the autoregressive (AR) part of the model.
- L^i: The i-th lag operator, where $L^i Y_t = Y_{t-i}$.
- $\nabla^d Y_t$: The d-th differenced series, used to achieve stationarity.
- θ_i: Coefficients of the moving average (MA) part of the model.
- ϵ_t: Error term at time t, representing random fluctuations.

The left side of the equation represents the autoregressive and differencing components, and the right side includes the moving average component and error term.

To implement ARIMA model we used the `statsmodel` library [22] in `Python` as it provides extensive functionality and extensions for ARIMA model.

IV. Results

A. Data and exploratory analysis

1) Lake chemistry data: Data on water chemistry comes from the Swedish lake monitoring program [4], covering 110 lakes during 2001-2022. Lake locations are visualized in Fig. 1. For most lakes, there are four sample dates each year. Some lakes have been added or removed during the period, and consequently have fewer observations.

Densities are plotted for all variables. Due to the LME assumptions, we ideally want them to be Gaussian, but in fact most are Gamma distributed with long right tails. This emphasizes the need for evaluating model residuals, and potential variable transformations.

Next, we perform correlation analysis on TOC and all other variables, shown in Fig. 2. Due to the structure of the data,

Fig. 1. Lake Locations

correlation coefficients are first calculated for each lake, then averaged across them. The highest positive correlations are found with organic N, Fe, Si and organic P. In terms of negative correlation, SO4 and pH have the largest coefficients. Most of these are in line with the initial hypotheses, particularly N, P, Fe and Si. Overall, the coefficients are quite small.

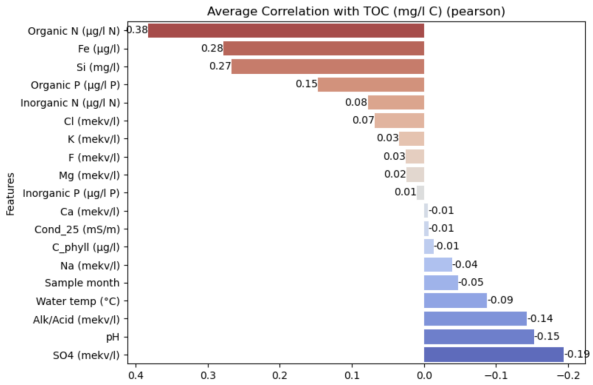

Fig. 2. Correlation between TOC and lake chemistry variables.

There is also substantial temporal autocorrelation of first (AR1) and second (AR2) order: five variables have > 0.4 correlation coefficients for AR1, and six for AR2 (details in Fig. 21, 22 in Appendix A).

2) Meteorology data: Data on temperature and precipitation are taken from the Swedish Meteorological and Hydrological Institute (SMHI) [23]. We first identify the active weather stations closest to each lake. Stations do not measure all types of meteorological data, so the candidate sets for temperature and precipitation are different. For each lake, we pick the two with the shortest distances. We then pull the available data and calculate rolling averages for different window sizes. Due to some SMHI stations becoming inactive or opening during this period, data is not always complete (as evident in Fig. 23, 24 in Appendix B).[3]

Almost all temperature measurements are within 40 km of the lake; for precipitation, most stations are within 30 km (Fig. 25, 26, Appendix B). While this is more coarse than ideal, it should still be indicative of local precipitation and temperature at most lake sites. As might be expected, the temperature variables follow multimodal distributions due to seasonality. The precipitation variables are Gamma distributed, with long right tails.

Compared to the chemical properties of lake water, correlations between TOC and the weather variables are more limited, as seen in Fig. 3. The highest positive correlations are with yearly precipitation levels, and temperature over the past weeks is most negatively correlated. Since these are calculated as rolling averages, and temperature is highly seasonal, it is not surprising that we find high temporal autocorrelation as well (not shown here).

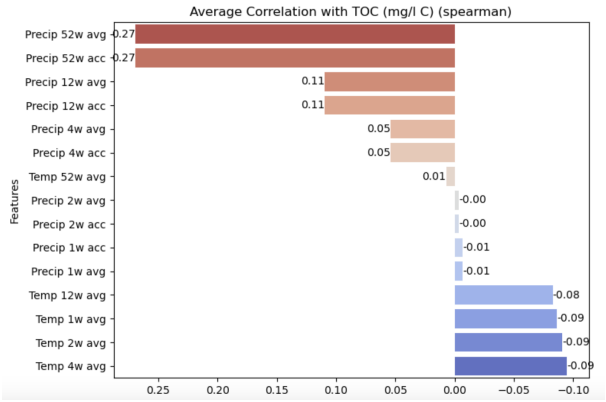

Fig. 3. Correlation between TOC and weather-related variables.

3) Catchment area and physical property data: The land cover data consisted of the total lake catchment area. Along with that information, it also gave us the altitude of the catchment area, the percentage cover of forests, marshland, farmland, open ground, mountains, tree felled land and urban land of all the monitored lakes. Since all the lakes have a unique MD-MVM-ID, they were mapped to the lake chemistry dataset and used in the analysis. The land cover data was statistic unlike the lake chemistry observations which was spanned over two decades.

[3]Completing these gaps would require significant additional (manual) effort, and we jointly chose to deprioritize this.

4) Hydrology data: Hydrology data for lakes which includes data on annual waterflow, total waterflow, local waterflow and water temperature was taken from SMHI [24]. The data spans over a time period of 11 years (2010 - 2021). We also calculated the 7-days and 30-days rolling averages for the 'Total Waterflow' for all lakes. These rolling averages were calculated for achieving better accuracy in our models. We also calculated the Water Residence Time (WRT) for lakes for which we had the mean depth i.e. mean(MEDDJ) available. WRT = mean lake depth $\times$ lake area$/$ mean annual waterflow

B. Experimental setup

Since all the data sets are relatively small, we are able to run all analysis and models locally in `Jupyter` notebooks.

C. Results of modelling

1) Results of linear mixed effects models: To make interpretation easier[4], and since data sets cover different time periods, we fit separate models based on different sets of variables. The response variable is TOC in all cases.

- Model 1.1: Lake chemistry variables.
- Model 1.2: Smaller model with select chemistry variables, based on correlation analysis, as a benchmark to M 1.1.
- Model 1.3: Lake chemistry variables, with RE slopes included.
- Model 2: Physical attributes, catchment area, and weather.
- Model 3: Seasonal and time variables.

Variables with more than 20% missing values are excluded from the LME models, since simple interpolation techniques become unreliable for large numbers and we do not want to drop too many observations. (see Table VII, Fig. 20 in Appendix C). Remaining missing values, after excluding the variables mentioned in the previous subsection, are interpolated using the mean for each lake and variable. For Model 3, dummy variables are generated for the yearly seasons (winter, spring, summer, fall). The Time count variable used in this model starts with a value of zero at the first measurement timestamp and is incremented for each subsequent measurement.

Both response and independent variables are min-max scaled. We make ten train-test runs with a 25% test fraction and evaluate average goodness of fit (R^2) [25] and accuracy (1 – sMAPE) [26]. It should be noted that R^2 is an ambiguous measure when comparing LMEs [6]. It will capture both the impact of the FE coefficients, and the random intercepts (and slopes), so it is not directly comparable to the R^2 of a standard OLS model.

A sequential feature selection (SFS) algorithm is used for all models except 1.2; it first applies a p-value criterion to ensure statistical significance, then Akaike and Bayesian information criteria (AIC, BIC) to limit the number of variables [27], [28]. Results are presented in Table I.

[4]There are complex causation chains between variables in the different data sets and TOC.

TABLE I
LINEAR MIXED EFFECTS MODELS BENCHMARK

	M1.1	M1.2	M1.3	M2	M3
R^2 (test)	0.863	0.855	0.881	0.792	0.801
R^2 (train)	0.866	0.863	0.889	0.810	0.808
1 – sMAPE (test)	0.843	0.845	0.864	0.828	0.823
1 – sMAPE (train)	0.846	0.849	0.868	0.830	0.827
Nb of variables	10	6	9 + 2	5	4

The models with lake chemistry variables (M1.1-1.3) have R^2 scores in the range of 0.85-0.88 and accuracies of about 84-86%. We can see that including more variables (M1.1 vs M1.2) helps explain somewhat more of the variation in the data on the test set (not substantially so, though), but test accuracy does not change. Adding RE slopes (M1.3) improves both model fit and accuracy; by how much depends on the RE variable combinations included (see Table X in Appendix C).[5] The suggested variable transformations from section IV-A (log and log-power) do not improve M1.1 (not shown here). Overall, a few important FE variables explain most of the variation in the data.

The model based on physical attributes, catchment area, and weather (M2), gives an R^2 of 0.79. This decrease compared to the models above, is likely because catchment area data is static, and the weather variables are more indirect than the chemical properties. Seasonal and time variables (M3) lead to similar results as M2 despite few variables included, potentially because the random intercepts already capture a lot of inter-lake data variation, but this result is still somewhat surprising. Prediction plots for all models are show in Appendix C.

Using a Shapiro-Wilks test [29], we reject the null hypothesis of residuals being drawn from a Gaussian. However, the residuals are still symmetrically distributed except for a thin right tail, across all models (Fig. 33 in Appendix C shows an example). Correlation between independent variables were generally in the range of 0.3-0.4 (see Fig. 34 in Appendix C), which is an acceptable level.

Table II shows the statistically significant variables from the final fit of each model. Since the data is scaled, the magnitudes can be compared relative to each other, but not interpreted in terms of their original units. In the lake chemistry models, organic N, Si and Fe exhibit coefficients in line with the hypotheses in section II. They are also quite consistent in magnitude across models. The sign of P is surprising based on the initial hypothesis, and correlation analysis. Note that when SO_4 is included as a RE in M1.3, the FE is no longer significant and thus excluded from the table.

In M2, the shares of Forest and Agriculture have the expected effect on TOC. The net effect of precipitation

[5]Doing an exhaustive grid search of all possible combinations would have been computationally infeasible. For M1.3 shown here, we pick the best-performing pair, SO4 and pH.

TABLE II
LINEAR MIXED EFFECTS MODELS OUTPUT SUMMARY

	Coef.	Std. Err.	p-val.	Conf. interval
Model 1.1				
Organic N	0.569	0.016	0.000	(0.538, 0.599)
Mg	0.431	0.034	0.000	(0.364, 0.499)
Fe	0.224	0.011	0.000	(0.203, 0.245)
Si	0.120	0.007	0.000	(0.106, 0.134)
F	0.065	0.015	0.000	(0.036, 0.094)
Chlorophyll	-0.039	0.011	0.001	(-0.062, -0.017)
pH	-0.121	0.009	0.000	(-0.138, -0.104)
Inorganic P	-0.174	0.022	0.000	(-0.217, -0.131)
Organic P	-0.248	0.030	0.000	(-0.306, -0.190)
SO4	-0.435	0.032	0.000	(-0.496, -0.373)
Model M1.2				
Organic N	0.751	0.018	0.000	(0.715, 0.787)
Fe	0.283	0.015	0.000	(0.254, 0.312)
Si	0.139	0.009	0.000	(0.122, 0.156)
Alk/Acid	0.089	0.018	0.000	(0.053, 0.125)
SO4	-0.094	0.024	0.000	(-0.142, -0.046)
pH	-0.096	0.010	0.000	(-0.115, -0.077)
Organic P	-0.421	0.030	0.000	(-0.480, -0.361)
Model 1.3				
Organic N	0.682	0.020	0.000	(0.643, 0.722)
Mg	0.436	0.049	0.000	(0.340, 0.532)
Fe	0.224	0.013	0.000	(0.199, 0.248)
Si	0.146	0.010	0.000	(0.127, 0.165)
F	0.111	0.017	0.000	(0.077, 0.144)
pH	0.059	0.015	0.000	(0.030, 0.088)
Chlorophyll	-0.062	0.020	0.002	(-0.101, -0.023)
Inorganic P	-0.189	0.029	0.000	(-0.247, -0.132)
Organic P	-0.214	0.038	0.000	(-0.288, -0.139)
SO4 RE				
pH RE				
Model 2				
Forest	0.200	0.014	0.000	(0.173, 0.228)
Agriculture	0.150	0.047	0.001	(0.058, 0.241)
Precip 52w avg	0.088	0.006	0.000	(0.075, 0.101)
Precip 4w avg	0.012	0.004	0.004	(0.004, 0.021)
Temp 2w avg	-0.025	0.003	0.000	(-0.031, -0.019)
Model 3				
Time count	0.028	0.002	0.000	(0.024, 0.031)
Winter	0.006	0.001	0.000	(0.003, 0.009)
Summer	-0.004	0.001	0.002	(-0.007, -0.001)
Spring	-0.007	0.001	0.000	(-0.010, -0.005)

seems to be an inflow of organic matter to the lakes, but the coefficient magnitudes are small. The effect of temperature is surprising, but on the other hand the effect is very small.[6] Based on M3, there seems to be a small positive trend in TOC, on average, but this varies between lakes. Somewhat surprising as well, Winter has a positive coefficient while Summer and Spring are negative.

2) Results of the spatial models: The variation in TOC values across 4 years in the autumn season can be seen in the Figure 36. Principal Component Analysis (PCA) was conducted. The variables were selected after evaluating their variance using PCA tool [30]. The predictor variables selected were influenced from the background knowledge of the lake bio-geo-chemical processes of the domain experts. Spatial pattern of TOC observations can be seen in Fig 37. It can also be seen from the Fig 37 that higher values of TOC were found mostly in the southern part of Sweden. Autocorrelation is a similarity measure with respect to distance, attribute values or both. The similarity measure can be positive or negative. If the attribute and the lagged attribute are positively or negatively autocorrlated then you will find more data points in the upper-right side and lower-left of the plot and if they are not autocorrelated then you will find more data points in the lower-right side and upper-left side of the Moran's plot 40 [31]. Moran's statistic of +1 indicates complete spatial autocorrelation and a score of -1 indicates absolutely dispersed data with no correlation. Moran's I score of 0.26 was obtained to indicate that there is a certain degree of autocorrelation, but not significant enough.

R package *spgwr* [10] was used for the GWR model to compare with OLS model. *spgwr* package of R language comes with in-built tests [32] to compare the results with OLS model. In all of the statistical tests, the Null is OLS. Results of the OLS model is in fig. 4, GWR model fig. 5, BFC 1999 Test [33] fig. 6, BFC 2002 Test fig. 7, LMZ F1 Test fig. 8, LMZ F2 Test fig. 9. Except for the LMZ F1 Test, in all other Tests the alternate hypothesis is True. In this case, the alternate being than the GWR model performed better than the OLS model in capturing and accounting for the variations in data.

The python library used for the exploratory analysis [34], building the MGWR & Spatial lag models was PySAL [35] which comes with tools to analyse and visualize spatial data. The TOC values for the 2020 Autumn season were used in building the models and tested for consistency & stability using the 2015 Autumn season as shown in the TableIII. Even though the R-squared and Adjusted R-squared statistic was same for the MGWR and Spatial-lag model, MGWR model saw significant reduction in sum squared residuals. Given the fact that the Moran's I score was 0.26, it explains the closeness in model statistics of MGWR and Spatial-Lag models.

```
Residual standard error: 1.288 on 86 degrees of freedom
Multiple R-squared:  0.9633,    Adjusted R-squared:  0.9556
F-statistic: 125.5 on 18 and 86 DF,  p-value: < 2.2e-16
```

Fig. 4. OLS Model Summary

```
Number of data points: 105
Effective number of parameters (residual: 2traceS - traceS'S): 66.80324
Effective degrees of freedom (residual: 2traceS - traceS'S): 38.19676
Sigma (residual: 2traceS - traceS'S): 0.8855667
Effective number of parameters (model: traceS): 57.43883
Effective degrees of freedom (model: traceS): 47.56117
Sigma (model: traceS): 0.7936122
Sigma (ML): 0.5341213
AICc (GWR p. 61, eq 2.33; p. 96, eq. 4.21): 435.6349
AIC (GWR p. 96, eq. 4.22): 223.7182
Residual sum of squares: 29.95499
Quasi-global R2: 0.9922942
```

Fig. 5. GWR Model Summary

[6]We also trained a standard linear regression model (not shown here) on average TOC for each lake and the catchment variables. In this model, Wetland had a positive effect on TOC (as expected) while the share of Water had a negative impact.

```
        Brunsdon, Fotheringham & Charlton (1999) ANOVA

data:  gwr.fit3
F = 4.9504, df1 = 91.256, df2 = 60.166, p-value = 3.455e-10
alternative hypothesis: greater
sample estimates:
SS GWR improvement   SS GWR residuals
         116.29415           26.30203
```

Fig. 6. BFC 1999 Test Summary

```
        Brunsdon, Fotheringham & Charlton (2002, pp. 91-2) ANOVA

data:  gwr.fit3
F = 5.4215, df1 = 86.000, df2 = 45.427, p-value = 5.322e-09
alternative hypothesis: greater
sample estimates:
SS OLS residuals SS GWR residuals
       142.59618         26.30203
```

Fig. 7. BFC 2002 Test Summary

```
        Leung et al. (2000) F(1) test

data:  gwr.fit3
F = 0.34919, df1 = 60.166, df2 = 86.000, p-value = 1.446e-05
alternative hypothesis: less
sample estimates:
SS OLS residuals SS GWR residuals
       142.59618         26.30203
```

Fig. 8. LMZ F1 Test Summary

```
        Leung et al. (2000) F(2) test

data:  gwr.fit3
F = 1.7287, df1 = 55.907, df2 = 86.000, p-value = 0.01095
alternative hypothesis: greater
sample estimates:
  SS OLS residuals SS GWR improvement
          142.5962           116.2942
```

Fig. 9. LMZ F2 Test Summary

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :     unknown
Weights matrix      :        None
Dependent Variable  :         toc                Number of Observations:         105
Mean dependent var  :     10.0629                Number of Variables   :          15
S.D. dependent var  :      6.1138                Degrees of Freedom    :          90
R-squared           :      0.8074
Adjusted R-squared  :      0.7774
Sum squared residual:     748.763                F-statistic           :     26.9464
Sigma-square        :       8.320                Prob(F-statistic)     :   3.333e-26
S.E. of regression  :       2.884                Log likelihood        :    -252.123
Sigma-square ML     :       7.131                Akaike info criterion :     534.246
S.E of regression ML:      2.6704                Schwarz criterion     :     574.055
```

Fig. 10. OLS Model Summary using PySAL

```
Model type                                                          Gaussian
Number of observations:                                                105
Number of covariates:                                                   15

Global Regression Results
----------------------------------------------------------------------------
Residual sum of squares:                                            20.225
Log-likelihood:                                                    -62.518
AIC:                                                               155.036
AICc:                                                              163.218
BIC:                                                              -398.632
R2:                                                                  0.807
Adj. R2:                                                             0.777
```

Fig. 11. MGWR Model Summary using PySAL

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :     unknown
Weights matrix      :        None
Dependent Variable  :         toc                Number of Observations:         105
Mean dependent var  :     10.0629                Number of Variables   :          16
S.D. dependent var  :      6.1138                Degrees of Freedom    :          89
R-squared           :      0.8097
Adjusted R-squared  :      0.7777
Sum squared residual:     739.567                F-statistic           :     25.2536
Sigma-square        :       8.310                Prob(F-statistic)     :   1.041e-25
S.E. of regression  :       2.883                Log likelihood        :    -251.474
Sigma-square ML     :       7.043                Akaike info criterion :     534.948
S.E of regression ML:      2.6540                Schwarz criterion     :     577.411
```

Fig. 12. Spatial Lag Model Summary using PySAL

TABLE III
MODEL STATISTICS FOR AUTUMN 2020,2015

	MGWR	Spatial-Lag(Wx)
2020		
Mean dependent var	na	10.0629
S.D dependent var	na	6.1138
R-squared	0.807	0.8097
Adjusted R-squared	0.777	0.7777
Sum squared residual	20.225	739.567
Sigma-square	na	8.31
S.E of regression	na	2.883
Sigma-square ML	na	7.043
S.E of regression ML	na	2.654
F-statistic	na	25.2536
Prob(F-statistic)	na	1.04e-25
Log likelihood	-62.518	-251.474
AIC	155.036	534.948
Schwarz criterion	na	577.411
AICc	163.218	na
BIC	-398.632	na
2015		
Mean dependent var	na	9.9383
S.D dependent var	na	5.9587
R-squared	0.796	0.8047
Adjusted R-squared	0.765	0.7726
Sum squared residual	21.821	734.873
Sigma-square	na	8.076
S.E of regression	na	2.842
Sigma-square ML	na	6.868
S.E of regression ML	na	2.6207
F-statistic	na	25.0036
Prob(F-statistic)	na	7.09e-26
Log likelihood	-66.764	-254.914
AIC	163.527	541.828
Schwarz criterion	na	584.593
AICc	171.572	na
BIC	-408.079	na

3) Results of Linear GAM xGBoost Model: This model was used to predict the TOC concentration in lakes based on mainly the lake chemistry, weather and hydrology data. First we clean the data by imputing the missing values and dropping columns with more than 50% missing values. Then we calculate the correlation of all the features with TOC as shown in figure 44. We only included features with a positive correlation with TOC.

Then we proceeded with calculating the Variance Inflation Factor (VIF) to check the multicollinearity and to ensure the robustness of our model.

Initially we trained a standalone Linear GAM model on our data set which had an accuracy of 69%. Then, to improve the accuracy we looked into using various ensemble techniques in our Linear GAM model. We trained a Random Forest model, Gradient Boosting Model and xGBoost model on the same data set with an accuracy of 66%, 60% and 71% respectively. We also hyper-tuned the Linear GAM model by assigning a smoothing parameter (lam=1.0) to spline terms for each feature instead of a default spline for each feature. This resulted in an accuracy of 70%.

Based on the accuracy scores we combined the hyper-tuned Linear GAM model with XGBoost model to achieve an accuracy of 75% which was a significant increase.

We make ten train-test runs with a 30% test fraction and evaluate average goodness of fit (R^2) [25] and accuracy (1 – sMAPE) [26]

The following table shows the performance comparison of all the models.

Model	R2	sMAPE	Accuracy
GAM	0.9149	30.7009	0.693
Random Forest	0.9442	33.2968	0.667
Gradient Boosting	0.943	39.8309	0.602
XGBoost	0.9625	28.7529	0.7125
Hyper-tuned GAM	0.9149	30.7009	0.693
Combined GAM and XGBoost	0.9584	24.6507	0.7535

TABLE IV
MODEL COMPARISON

Fig. 13 shows the Actual vs Predicted values for the Linear GAM XGBoost Model.

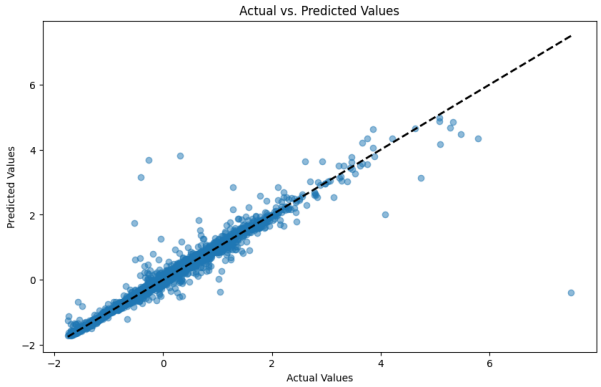

Fig. 13. Actual vs Predicted values : Linear GAM xGBoost Model

Fig. 14 shows the Density plot for the Linear GAM XGBoost Model.

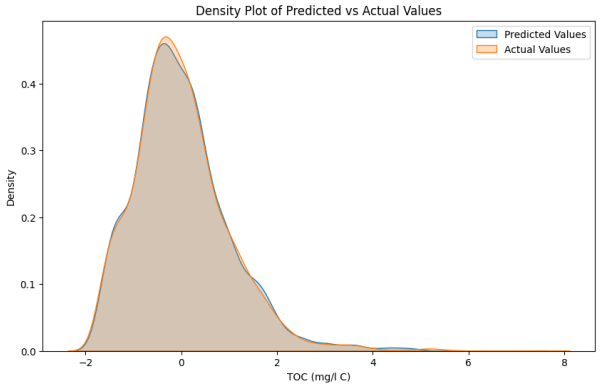

Fig. 14. Density Plot : Linear GAM xGBoost Model

4) Results of Hydrograph Trend Analysis: We created hydrographs per year for all 76 lakes for which we had hydrology data specifically 'Total Waterflow' available from SMHI [24].

The following is a hydrograph for one of the lakes for the year 2012.

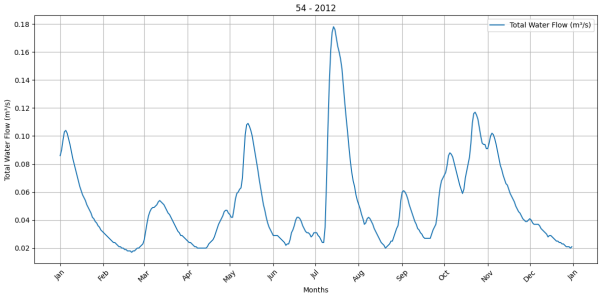

Fig. 15. Hydrograph for lake 54 2012

Then we calculated the 'Shape mean' and 'Shape variance' for all the hydrographs to use them for our trend analysis [5].

We then performed the Kendall Tau Trend test V-F based on the 'Shape Mean' and 'Shape Variance' we calculated from our hydrographs. The trend test clearly showed that for majority of the lakes there is no significant upward or downward trend.

The following scatter plot shows the trends 'Yellow' represents no significant trend, 'Red' represents strong downward trend and no 'Green' dots shows that there is no strong upward trend for any lake.

Hence we can't reach a conclusion to answer one of our research question i.e. Are the lakes and catchments getting wetter or dryer over time?

5) Results of forecasting using ARIMA model: We used ARIMA model [21] to forecast the 'Total Waterflow' from 2022 to 2025 for one of the lakes i.e. MD MVMId 54. Currently we have hydrology data available for 11 years (2010 - 2021) [24].

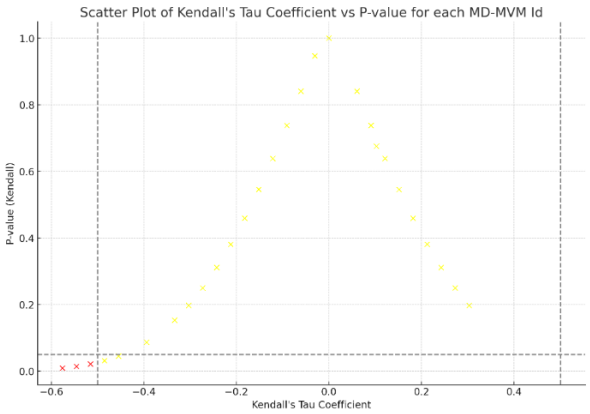

Fig. 16. Kendall Tau Trend Test

Then we perfromed the Augmented Dickey-Fuller (ADF) test [36] to check for the null hypothesis of a unit root (i.e., non-stationarity). This step is important because the ARIMA models assume stationarity in the dataset.

Measure	Value
ADF Statistic	-10.32304667603171
p-value	2.974×10^{-18}
Critical Value (1%)	-3.4318438667946554
Critical Value (5%)	-2.862200107021411
Critical Value (10%)	-2.5671213792977627

TABLE V
ADF TEST RESULTS

The table above shows that since ADF statistic is lower than all the crticial values the null hypothesis can be rejected ensuring we can move on with ARIMA model for forecasting.

Then to determine the possible values for our ARIMA model parameters p, d and q we used Autocorrelation Fuction (ACF) plot and Partial Autocorreation Function (PACF) plot.

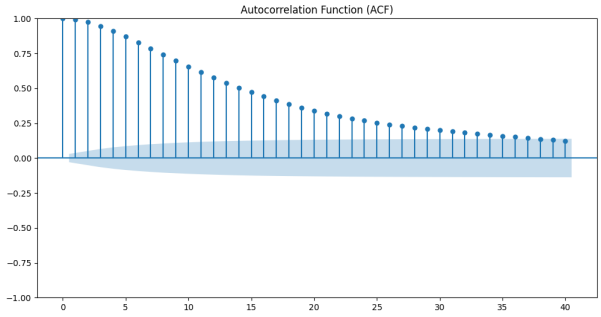

Fig. 17. ACF Plot

We tried various combinations of parameters for the ARIMA model. Below are the results:

We moved forward with ARIMA(3,0,2) model for forecasting 'Total Waterflow' since it had the best fit.

After forecasting 'Total Waterflow' for lake MD MVMId 54 we calculated the variance in the foreasted values to see if the model was overfitting.

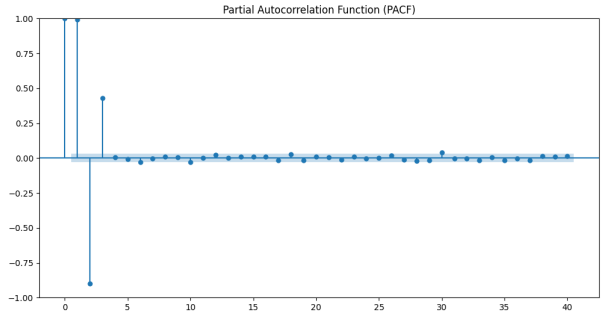

Fig. 18. PACF Plot

Model	Log Likelihood	AIC	BIC
ARIMA(2, 0, 1)	21831.752	-43653.503	-43621.576
ARIMA(2, 0, 2)	21900.589	-43789.179	-43750.866
ARIMA(3, 0, 1)	21541.672	-43071.343	-43033.031
ARIMA(3, 0, 2)	21928.833	-43843.666	-43798.968

TABLE VI
COMPARISON OF ARIMA MODEL RESULTS

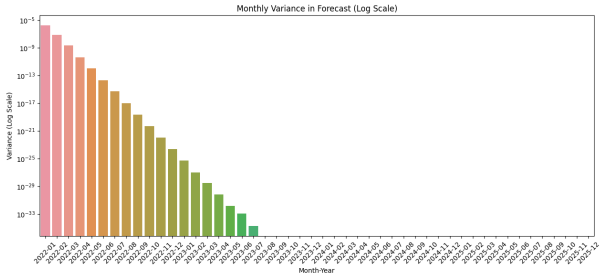

Fig. 19. Variance in Forecasted Values

From the figure above we can clearly see that the forecasted values variance drops significantly as the months go on indicating that our model is overfitting and that there is not enough data to predict three years into the future. Hence, we did not proceed with using this model to forecast 'Total Waterflow' for the remaining lakes.

V. DISCUSSION

The LME models performed quite well, especially the ones based on lake chemistry data. Their relative advantage compared to the model with catchment and weather, is likely explained by the frequency and "proximity" of the data. It is clear that a few chemistry variables explain most of the variation in the data, notably organic N, Mg, Fe, organic P and SO4, as well as share of forest and agriculture in the catchment. Overall, the signs of most of the significant variables were in line with the initial hypotheses. Some results were surprising, however, e.g. that wetland did not come out as significant, and that P had a negative effect on TOC.

There are several improvement avenues for future research. One is to go from static to yearly land use data [37]. Another is collecting lake chemistry data from the start of the current program, in 1989. There is also additional data on catchments available from SMHI, such as soil types. In hindsight, using

the `lme4` package in `R` would have provided more flexibility in specifying error structures. Delving deeper into model comparison (e.g. R^2 alternatives) could lead to further improvements.

The performance of the Spatial models improve if more data is used for modelling. We only had data for 110 lakes that were geographical spread out across Sweden. In a country that has 97,500 lakes larger than 2 acres (8,100 m2) [38], 110 lakes is insignificantly small number to infer spatial patterns at a global scale. A better direction to understand the spatial patterns would be to use image analysis techniques & deep learning on the satellite data.

The Hydrograph Trend Analysis [5] showed no significant upward or downward trend for almost all the lakes which shows a stable level of water flow in lakes and catchments, leading us to conclude that the hydrological ecosystem in lakes and catchments of Sweden is well preserved. The trend analysis was done for only a few lakes this trend can be extended to get a better overall picture.

The ARIMA model [21] used in forecasting did not perform well mainly because of the limited historical data we had at our disposal (2010 - 2021) for the 'Total Waterflow' of lakes. This resulted is low variance in the forecasted values after only an year when we were trying to predict the waterflow for a period of three years (2022 - 2025). The model can perform well if more historical data is used to train.

The Linear GAM model [18] performed significantly better when we combined it with an ensemble model (xG Boost). Still the overall accuracy of the model for predicting TOC remained lower compared to other models used in our research like LME showing that our data has significant random effects and the non-linearities and interactions in the model are not as complex.

REFERENCES

[1] D. N. Kothawala, C. A. Stedmon, R. A. Müller, G. A. Weyhenmeyer, S. J. Köhler, and L. J. Tranvik, "Controls of dissolved organic matter quality: evidence from a large-scale boreal lake survey," *Global Change Biology*, vol. 20, pp. 1101–1114, 2014.

[2] A. M. Kellerman, D. N. Kothawala, T. Dittmar, and L. J. Tranvik, "Persistence of dissolved organic matter in lakes related to its molecular characteristics," *Nature Geoscience*, 2015, published online: 25 May 2015.

[3] K. Toming, J. Kotta, E. Uuemaa, S. Sobek, T. Kutser *et al.*, "Predicting lake dissolved organic carbon at a global scale," *Scientific Reports*, vol. 10, p. 8471, 2020. [Online]. Available: https://www.nature.com/articles/s41598-020-65010-3

[4] J. Fölster, R. Johnson, M. Futter *et al.*, "The swedish monitoring of surface waters: 50 years of adaptive monitoring," *AMBIO*, vol. 43, no. Suppl 1, pp. 3–18, 2014.

[5] R. R. M. B. Simon Tardif, André St-Hilaire and S. Payette, "Statistical properties of hydrographs in minerotrophic fens and small lakes in mid-latitude québec, canada," *Canadian Water Resources Journal / Revue canadienne des ressources hydriques*, vol. 34, no. 4, pp. 365–380, 2009. [Online]. Available: https://doi.org/10.4296/cwrj3404365

[6] X. A. Harrison, L. Donaldson, M. E. Correa-Cano, J. Evans, D. N. Fisher, C. E. D. Goodwin, B. S. Robinson, D. J. Hodgson, and R. Inger, "A brief introduction to mixed effects modelling and multi-model inference in ecology," *PeerJ*, vol. 6, p. e4794, 2018. [Online]. Available: https://peerj.com/articles/4794/

[7] K. Shedden, "Statsmodels-mixedlm," GitHub repository, 2023, last accessed: 2023-11-24. [Online]. Available: https://github.com/kshedden/Statsmodels-MixedLM

[8] Duchesnay, "Linear mixed models," Statistics and Machine Learning in Python 0.5 documentation, 2023, last accessed: 2023-11-24. [Online]. Available: https://duchesnay.github.io/pystatsml/statistics/lmm/lmm.html

[9] J. Perktold, S. Seabold, J. Taylor, and the statsmodels developers, "Linear mixed effects models - statsmodels 0.14.0," Online, May 2023, last update: May 05, 2023, Last accessed: 2023-11-24. [Online]. Available: https://www.statsmodels.org/stable/mixed_linear.html#module-statsmodels.regression.mixed_linear_model

[10] "Geographically weighted regression — gwr — rspatial.r-forge.r-project.org," https://rspatial.r-forge.r-project.org/spgwr/reference/gwr.html, [Accessed 05-01-2024].

[11] P. A. P. Moran, "The Interpretation of Statistical Maps," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 10, no. 2, pp. 243–251, 12 2018. [Online]. Available: https://doi.org/10.1111/j.2517-6161.1948.tb00012.x

[12] A. S. F. C. Brunsdon, "Local forms of spatial analysis," *Geographical Analysis*, vol. 31, pp. 340–358, 1999. [Online]. Available: http://doi.org/10.1111/j.1538-4632.1999.tb00989.x

[13] T. M. Oshan, Z. Li, W. Kang, L. J. Wolf, and A. S. Fotheringham, "mgwr: A python implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale," *ISPRS International Journal of Geo-Information*, vol. 8, no. 6, 2019. [Online]. Available: https://www.mdpi.com/2220-9964/8/6/269

[14] W. Y. A. Stewart Fotheringham and W. Kang, "Multiscale geographically weighted regression (mgwr)," *Annals of the American Association of Geographers*, vol. 107, no. 6, pp. 1247–1265, 2017. [Online]. Available: https://doi.org/10.1080/24694452.2017.1352480

[15] "AM-32 - Spatial Autoregressive Models — GIS&T Body of Knowledge — gistbok.ucgis.org," https://gistbok.ucgis.org/bok-topics/spatial-autoregressive-models, [Accessed 04-01-2024].

[16] "Moran's I - Wikipedia — en.wikipedia.org," https://en.wikipedia.org/wiki/Moran%27s_I, [Accessed 06-01-2024].

[17] "Tobler's first law of geography - Wikipedia — en.wikipedia.org," https://en.wikipedia.org/wiki/Tobler%27s_first_law_of_geography, [Accessed 05-01-2024].

[18] T. J. Hastie and R. J. Tibshirani, "Generalized additive models," *Statistical Science*, vol. 1, no. 3, pp. 297–310, 1986. [Online]. Available: https://projecteuclid.org/euclid.ss/1177013604

[19] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794. [Online]. Available: https://arxiv.org/abs/1603.02754

[20] D. Servén and C. Brummitt, "pygam: Generalized additive models in python," 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1208723

[21] H. AL-Chalabi, Y. Al-Douri, and J. Lundberg, "Time series forecasting using arima model: A case study of mining face drilling rig," 08 2018.

[22] S. Seabold, J. Perktold *et al.*, "statsmodels: Autoregressive integrated moving average (arima) model," 2020. [Online]. Available: https://www.statsmodels.org/stable/generated/statsmodels.tsa.arima.model.ARIMA.html

[23] "SMHI Meteorological Observations API Documentation," https://opendata.smhi.se/apidocs/metobs/index.html, last accessed: 2023-12-19.

[24] "SMHI Water Web," https://www.smhi.se/data/hydrologi/vattenwebb.

[25] "Coefficient of determination - wikipedia," https://en.wikipedia.org/wiki/Coefficient_of_determination, accessed on: [Insert Access Date Here].

[26] "Symmetric mean absolute percentage error - wikipedia," https://en.wikipedia.org/wiki/Symmetric_mean_absolute_percentage_error, accessed on: [Insert Access Date Here].

[27] "Akaike information criterion - wikipedia," https://en.wikipedia.org/wiki/Akaike_information_criterion, last accessed: 2023-12-19.

[28] "Bayesian information criterion - wikipedia," https://en.wikipedia.org/wiki/Bayesian_information_criterion, last accessed: 2023-12-19.

[29] "scipy.stats.shapiro - scipy v1.8.0 reference guide," https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html, accessed on: [Insert Access Date Here].

[30] "PCA documentation! &x2014; pca pca documentation — erdogant.github.io," https://erdogant.github.io/pca/pages/html/index.html, [Accessed 05-01-2024].

[31] "Project 4: Calculating Global Moran's I and the Moran Scatterplot — GEOG 586: Geographic Information Analysis — e-education.psu.edu," https://www.e-education.psu.edu/geog586/node/672, [Accessed 06-01-2024].

[32] "Global tests of geographical weighted regressions — LMZ.F3GWR.test — rspatial.r-forge.r-project.org," https://rspatial.r-forge.r-project.org/spgwr/reference/LMZ.F3GWR.test.html, [Accessed 06-01-2024].
[33] S. A. Fotheringham and C. Brunsdon, "Local forms of spatial analysis," *Geographical Analysis*, vol. 31, pp. 340–358, 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:51766919
[34] "ESDA: Exploratory Spatial Data Analysis &x2014; esda v2.5.1 Manual — pysal.org," https://pysal.org/esda/, [Accessed 06-01-2024].
[35] "PySAL — pysal.org," https://pysal.org/, [Accessed 05-01-2024].
[36] D. A. Dickey and W. A. Fuller, "Likelihood ratio statistics for autoregressive time series with a unit root," *Econometrica*, vol. 49, no. 4, pp. 1057–1072, 1981.
[37] "300 m annual global land cover time series from 1992 to 2015," https://www.esa-landcover-cci.org/?q=node/175, 2017, accessed on: [Insert Access Date Here].
[38] "List of lakes of Sweden - Wikipedia — en.wikipedia.org," https://en.wikipedia.org/wiki/List_of_lakes_of_Sweden, [Accessed 06-01-2024].
[39] A. L. K. A. Luc Anselin, Grant Morrison, "Chapter 6 Contiguity-Based Spatial Weights — Hands-On Spatial Data Science with R — spatialanalysis.github.io," https://spatialanalysis.github.io/handsonspatialdata/contiguity-based-spatial-weights.html, [Accessed 07-01-2024].
[40] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1-2, pp. 81–93, 1938.
[41] B. Beers, "P-value: What it is, how to calculate it, and why it matters," https://www.investopedia.com/terms/p/p-value.asp, 2023.

APPENDIX

A. *Lake chemistry data analysis*

Fig. 20 shows how many measurements there are in total for each lake, in the lake chemistry data set. Most lakes have between 80 and 87 measurements.

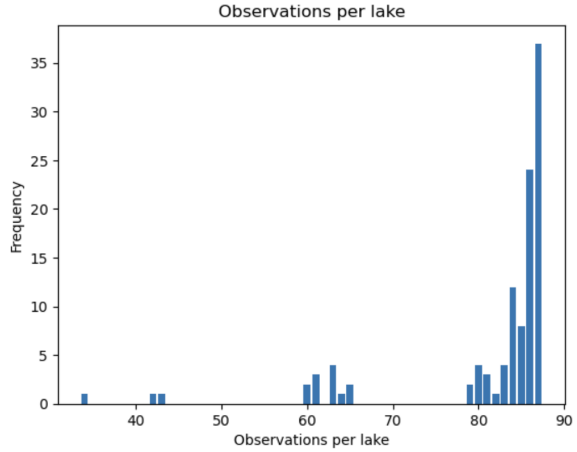

Fig. 20. Measurements per lake in the chemistry data set.

Fig. 21 and Fig. 22 illustrates that there is a high level of autocorrelation between the variables and their lagged counterparts in the lake chemistry data set.

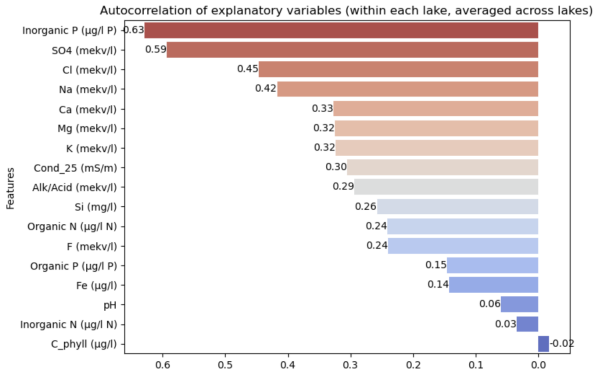

Fig. 21. First order autocorrelation (AR1) between variables.

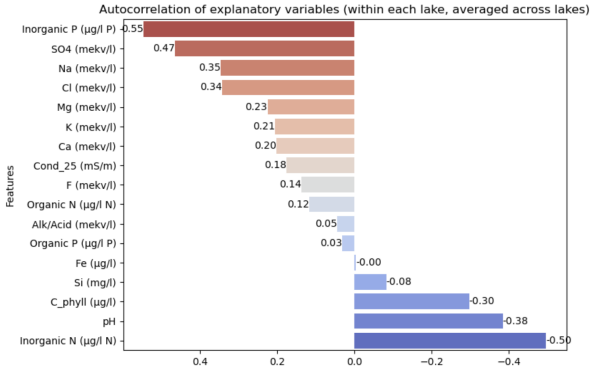

Fig. 22. Second order autocorrelation (AR2) between variables.

B. Meteorology data analysis

Fig. 24 and Fig. 24 shows that weather observations have less coverage than the lake chemistry data, across all lakes and the full time period.

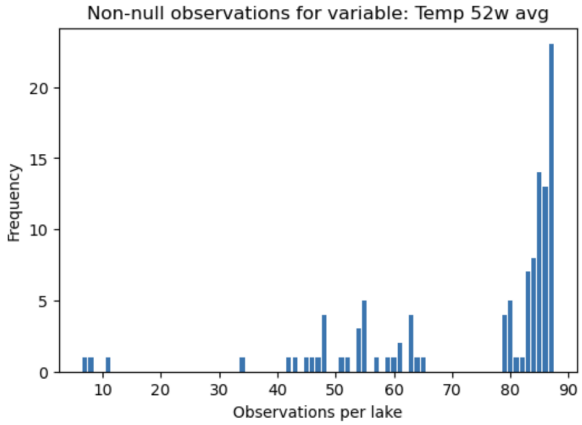

Fig. 23. Rolling average 52w temperature measurements per station / lake.

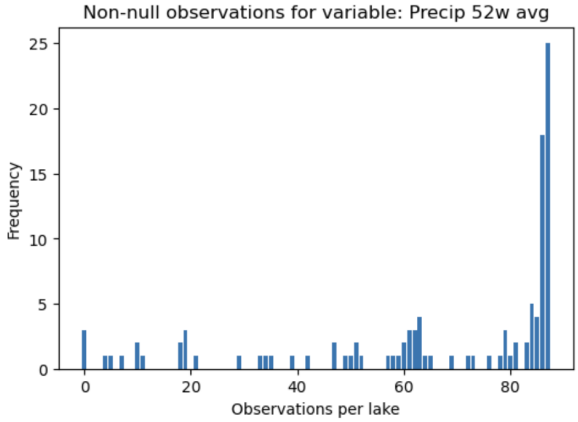

Fig. 24. Rolling average 52w precipitation measurements per station / lake.

Fig. 25 and Fig. 26 shows the distribution of distances between each lake and its closest SMHI stations that report temperature and precipitation. Ideally, these stations should be located as close as possible to the lakes.

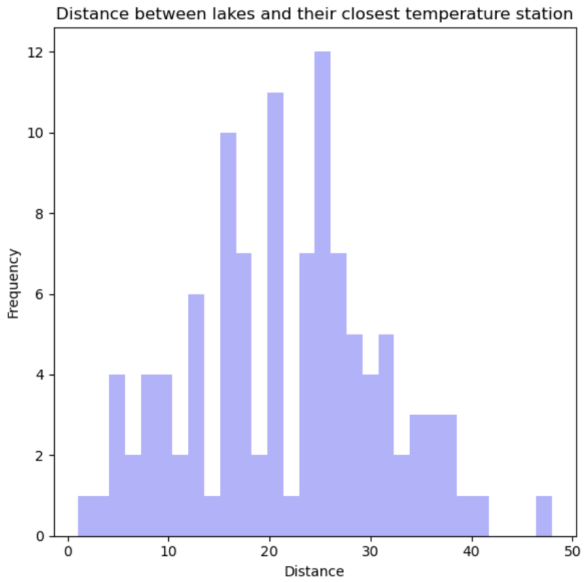

Fig. 25. Distance between each lake and the closest active SMHI station that measures temperature.

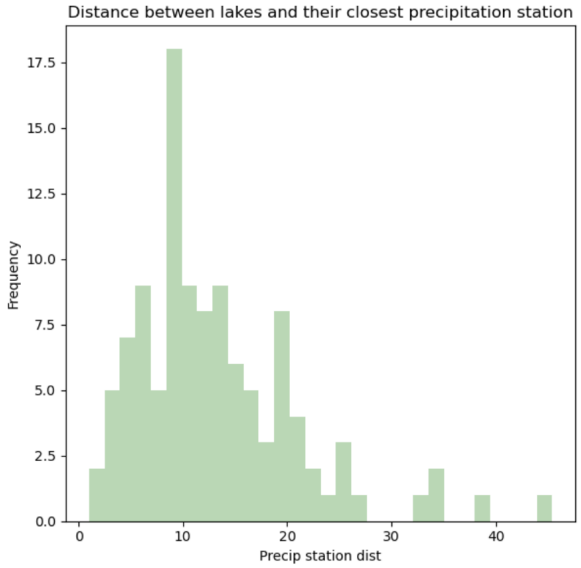

Fig. 26. Distance between each lake and the closest active SMHI station that measures precipitation.

C. Linear mixed effects models

To illustrate what an LME model does, we provide an example here. Let us assume that we are regressing TOC on N concentration, a variable that changes over time, and share of wetland (WL), which is constant over time, the regression function for lake i at time t would be

$$TOC_{i,t} = (\beta_0 + \beta_1 N_{i,t} + \beta_3 WL_i) + ...$$
$$... + (\gamma_{0i} + \gamma_{1i} N_{i,t} + \gamma_{2i} WL_i) + \epsilon_{i,t} =$$
$$= \beta_0 + \gamma_{0i} + (\beta_1 + \gamma_{1i}) N_{i,t} + (\beta_2 + \gamma_{2i}) WL_i + \epsilon_{i,t}$$

where the intercept β_0 and slopes β_1, β_2 are population-level parameters (across all lakes) and γ_{0i}, γ_{1i}, γ_{2i} are group-level parameters (specific to each lake). The errors $\epsilon_{i,t}$ are independent of everything else, identically distributed with mean zero. If only γ_{0i} is included, the LME is a random intercept model (responses in a group are just additively shifted); including γ_{1i}, γ_{2i} makes it a random slope model (responses are shifted and follow a different conditional mean trajectory).

Table VII lists all variables that were excluded from the data sets due to having more than 20% missing values.

TABLE VII
LIST OF LAKE CHEMISTRY VARIABLES

Variable
Turbidity
Al, Al_s
Oxygen
Secchi depth
Secchi depth binoculars
Secchi depth no binoculars

Table VIII shows all variables from the lake chemistry data that form the basis for feature selection (after exclusions e.g. due to large number of missing values). Table IX shows the equivalent for physical attributes, catchment area, and weather.

Fig. 27, 28, 29, 30, 31 show the predicted values compared to actuals for all LME models.

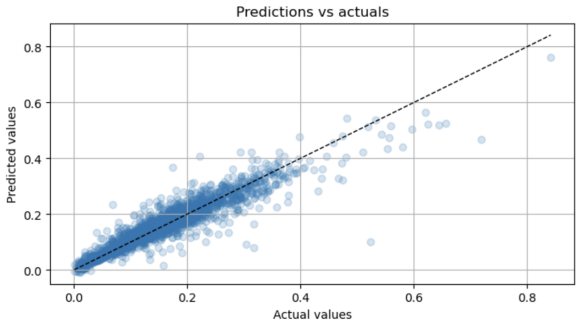

Fig. 27. Predicted TOC values from Model 1.1 compared to actual values.

Table X illustrates the outcome of including different pairs of random effect slopes in the lake chemistry model. Fig. 32

TABLE VIII
LIST OF LAKE CHEMISTRY VARIABLES

Variable
Chlorophyll
pH
Conductivity
Alk/Acid
Ca
Mg
Na
K
SO4
Cl
F
Si
Fe
Water temp
Organic N
Inorganic N
Organic P
Inorganic P

TABLE IX
LIST OF VARIABLES FOR PHYSICAL ATTRIBUTES, CATCHMENT AREA, AND WEATHER

Variable
Altitude
Latitude
Longitude
Forest
Water
Wetland
Clearing
Open land
Agriculture
Mountains
Urban area
WRT
Temp 1w avg
Temp 2w avg
Temp 4w avg
Temp 12w avg
Temp 52w avg
Precip 1w avg
Precip 2w avg
Precip 4w avg
Precip 12w avg
Precip 52w avg

shows the individual slopes for a subset of lakes. These slopes are derived by adding the RE coefficients to the FE coefficient for the whole population.

Fig. 33 shows the residuals from Model 1.1. The distribution is not normal, but still symmetric. This is representative of all the models, with the main difference being the extent of the right tail.

Fig. 34 illustrate the correlations between the independent variables in the lake chemistry models.

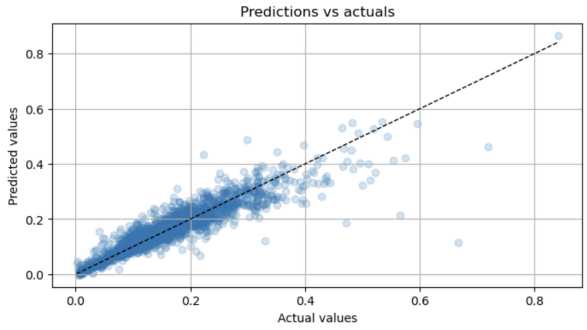

Fig. 28. Predicted TOC values from Model 1.2 compared to actual values.

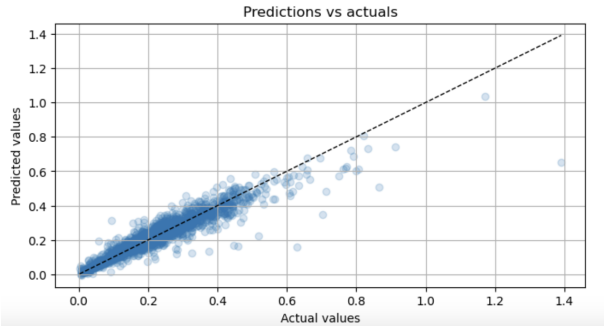

Fig. 29. Predicted TOC values from Model 1.3 compared to actual values.

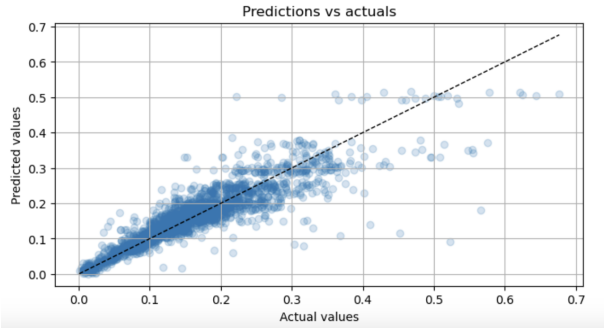

Fig. 30. Predicted TOC values from Model 2 compared to actual values.

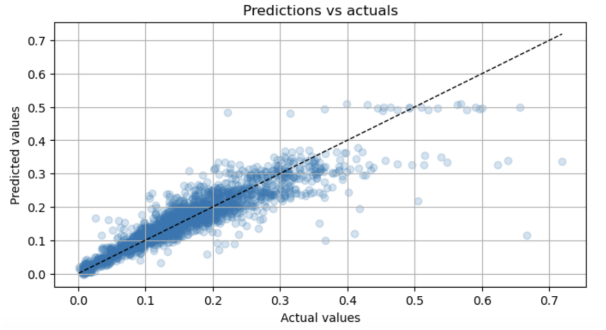

Fig. 31. Predicted TOC values from Model 3 compared to actual values.

TABLE X
CHANGE IN R^2 FROM PAIRS OF RANDOM EFFECT VARIABLES

Variable Pairs	**Change in R^2**
SO4, pH	0.0302
Si, Fe	0.0255
Mg, SO4	0.0253
SO4, Fe	0.0242
Organic N, pH	0.0218
pH, Fe	0.0216
Organic N, SO4	0.0216
SO4, Chlorophyll	0.0205
pH, Si	0.0203
Mg, Fe	0.0190
Mg, Organic N	0.0184
Mg, Si	0.0178
Organic N, Fe	0.0163
Chlorophyll, Si	0.0131
Organic N, Si	0.0122
Mg, Chlorophyll	0.0106
Chlorophyll, Fe	0.0093
Chlorophyll, pH	0.0075
Mg, pH	0.0072
SO4, Si	0.0032

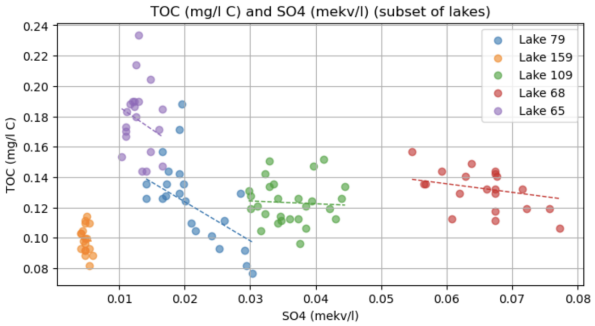

Fig. 32. Illustration of lake-specific slopes (coefficients) from Model 1.2.

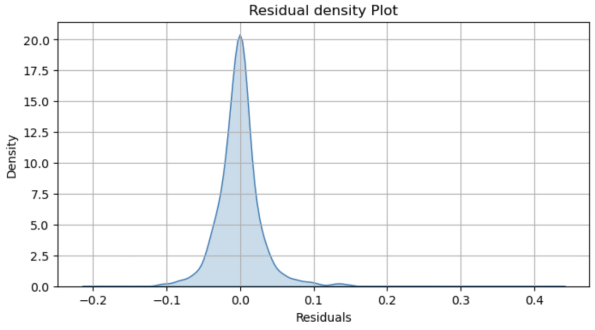

Fig. 33. Residuals (errors) from Model 1.1.

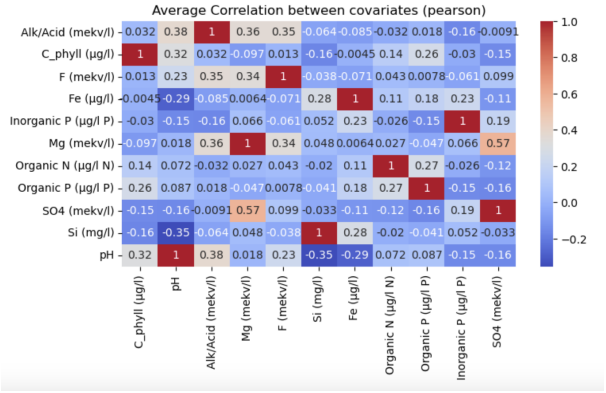

Fig. 34. Correlation coefficients between independent variables in Models 1.1-1.3.

D. Spatial Models

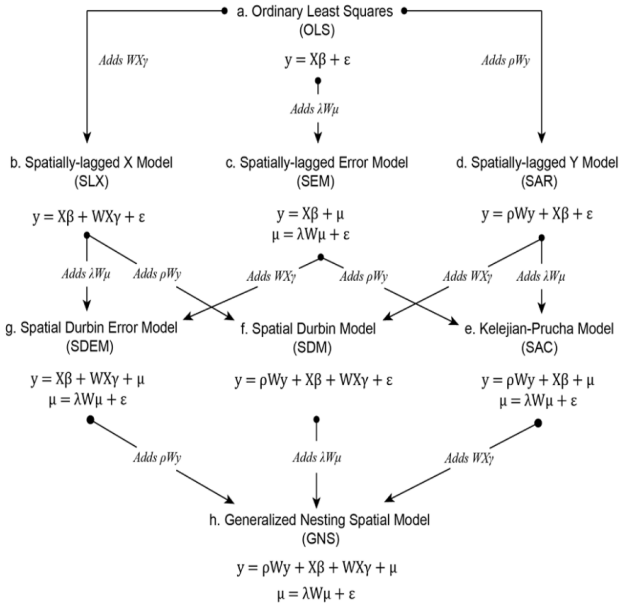

Fig. 35. Autoregressive Model Taxonomy

The GWR estimator [13] in the matrix form for a local estimate at site i is given by:

$$\hat{\beta}_i = \left[X^T W_i X\right]^T X^T W_i y \tag{11}$$

$W\,(i)$ = diag$[w_i, ..., w_n]$ is a n * n diagonal weights matrix that weighs each observation based on its distance from location i, The model inputs X, y, and the geographic co-ordinates (u, v). The kernel density function assigns weights to neighbouring units and the bandwidth h of the function decides the distance decay. Some of the commonly used density functions for calculating the weighting function are:
Gaussian Density Function

$$w_{ij} = exp\left(\frac{-d_{ij}^2}{h^2}\right) \tag{12}$$

where d_{ij} is the distance between locations i & j and h is the bandwidth.
Exponential Weighted Function

$$w_{ij} = exp\left(\frac{-d_{ij}}{h}\right) \tag{13}$$

Bi-Square function

$$w_{ij} = 1 - \left(\frac{d_{ij}^2}{h^2}\right)^2 \tag{14}$$

Selecting a density function also chooses the bandwidth. Some of the commonly used bandwidth selection methods are leave-one-out cross-validation, Akaike Information Criteria, Corrected Akaike Information Criteria (AICc), Bayesian Information Criteria (BIC).

The fig shows 42 spatial weights calculation for deciding who neighbors are based on contiguity spatial weights (Rook, Queen) and distance based spatial weights(KNN) [39]

The fig shows 43 local autocorrelations. Hotspot refers to a cluster of lakes whose TOC values were high in comparison to the global TOC values. Coldspot refers to a cluster of lakes whose TOC values were low in comparison to the global mean TOC values. Doughnut refers to when a lake is surrounded by its neighbors whose values were higher than the lake in focus. Diamond refers to when the lake TOC value is higher than its neighbors. ns refers to non-signficant i.e. their TOC values were in line with the mean global TOC values.

E. Linear GAM XGBoost Model

Fig. 44 shows correlations of all the features with TOC. This is the data that'll be used to train the Linear GAM XGBoost model for predicting TOC.

F. Kendall Tau Trend Test

The Kendall Tau test [40] is a non-parametric hypothesis test that measures the extent of dependency between two variables. In trend analysis, the Kendall Tau test is used to find any statistically significant trend in the time series data. It has two main coefficients p-value and Kendall Tau coefficient.

Given a set of observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, any pair of observations (x_i, y_i) and (x_j, y_j), where $i < j$, are said to be consistent if the ranks for both elements agree i.e. if both $x_i > x_j$ and $y_i > y_j$ or if both $x_i < x_j$ and $y_i < y_j$. They are said to be inconsistent, if $x_i > x_j$ and $y_i < y_j$ or if $x_i < x_j$ and $y_i > y_j$.

The Kendall Tau coefficient is calculated as:

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} \text{sign}(x_i - x_j) \cdot \text{sign}(y_i - y_j) \tag{15}$$

where sign$(\cdot)$ is the sign function.

The p-value in statistical hypothesis testing [41] is the probability of obtaining test results under the assumption that the null hypothesis is correct. A low p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so it is rejected. A high p-value (> 0.05) suggests weak evidence against the null hypothesis, and is not rejected.

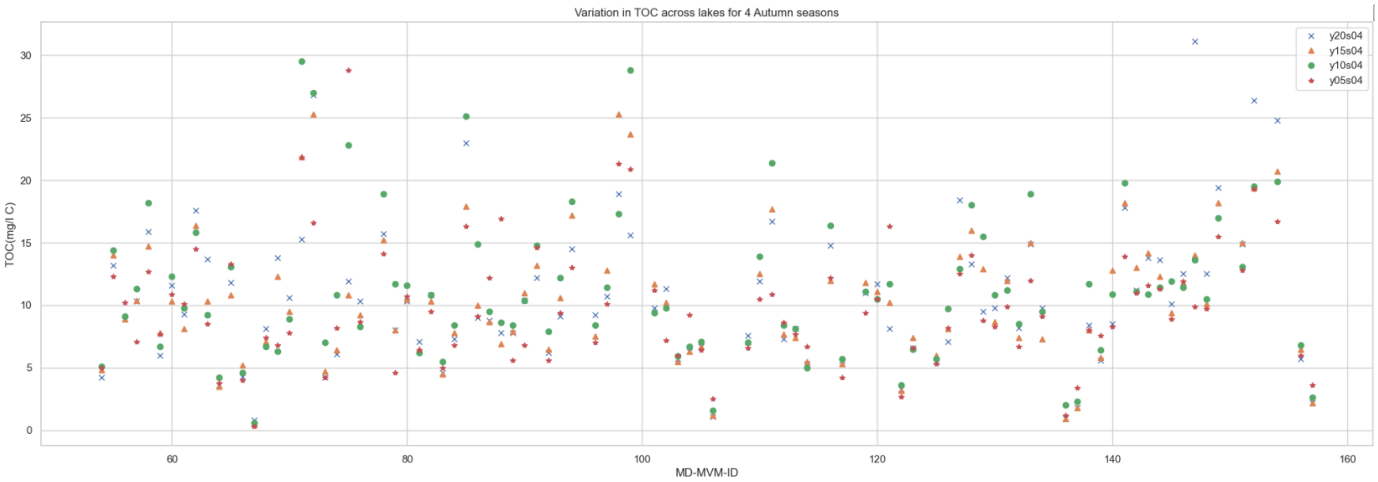

Fig. 36. Variation of TOC across lakes for 4 Autumn seasons

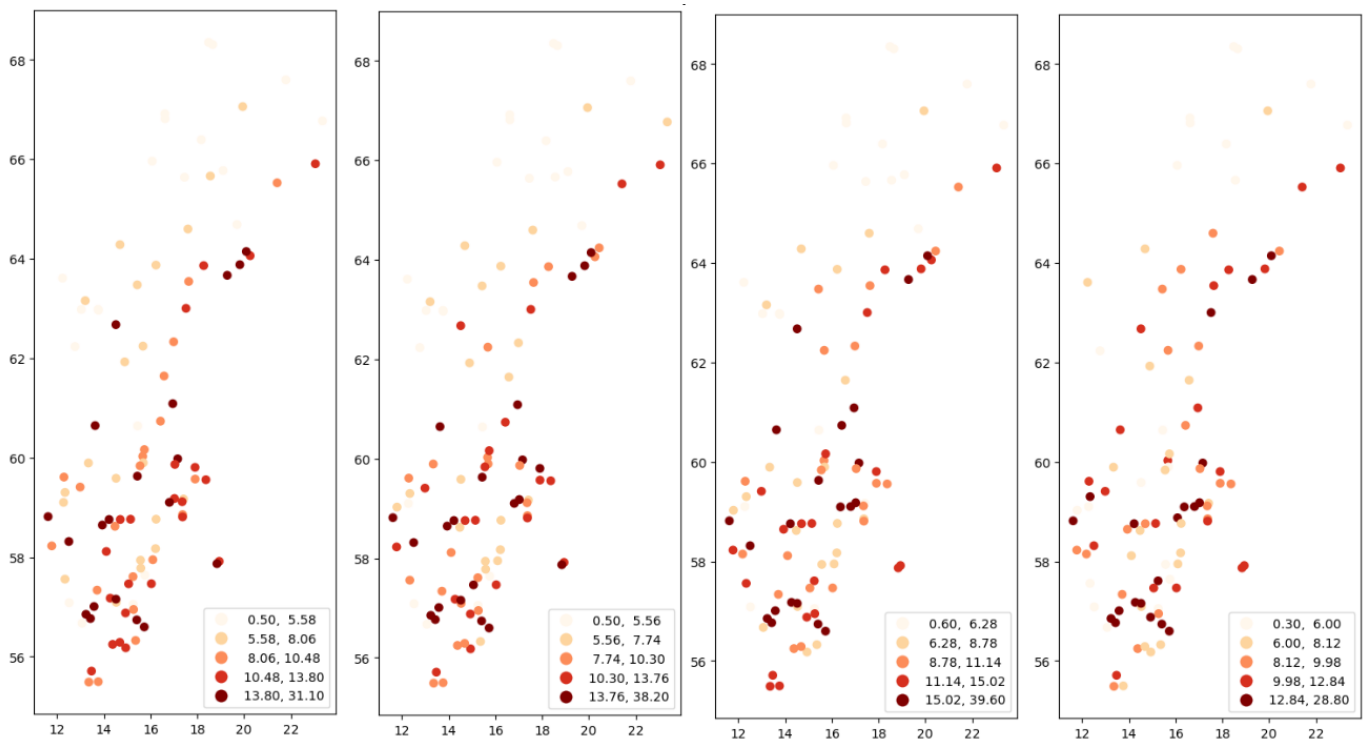

Fig. 37. TOC across lakes in Autumn of 2020, 2015, 2010, 2005

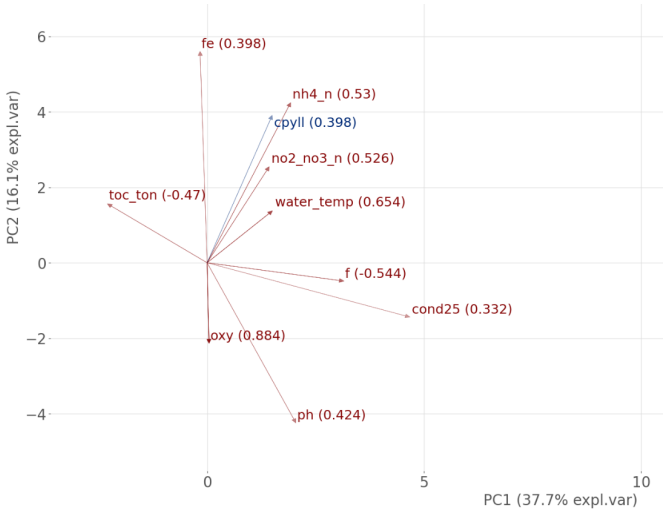

Fig. 38. PCA Biplot

```
(Dispersion parameter for gaussian family taken to be 1.658095)

    Null deviance: 3887.3  on 104  degrees of freedom
Residual deviance:  142.6  on  86  degrees of freedom
AIC: 370.11

Number of Fisher Scoring iterations: 2
```

Fig. 41. GLM Model Summary

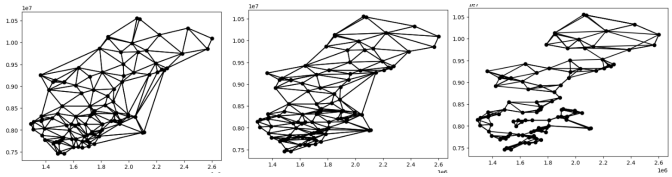

Fig. 42. Spatial weights Contiguity based(Rook, Queen) vs distance based(KNN)

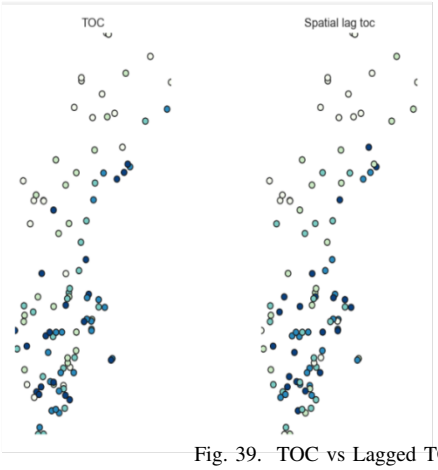

Fig. 39. TOC vs Lagged TOC

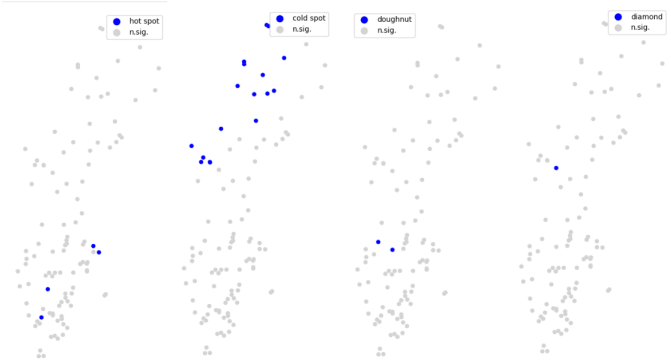

Fig. 43. Local Autocorrelations Hotspot, Coldspot, Doughnut, Diamond

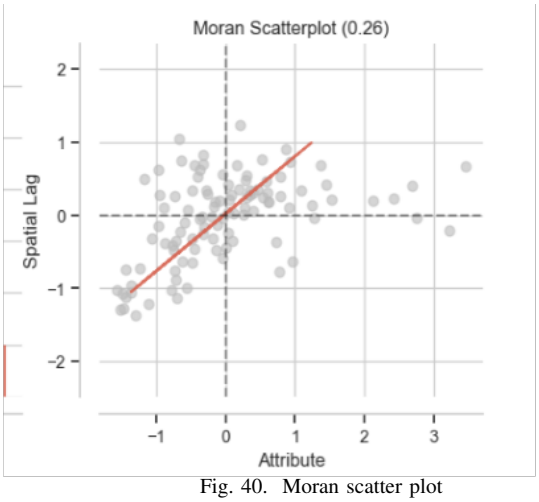

Fig. 40. Moran scatter plot

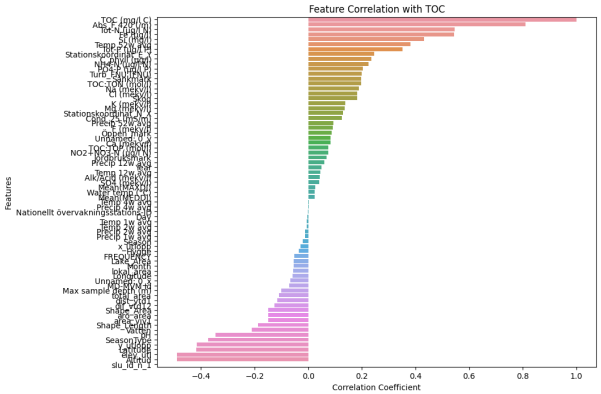

Fig. 44. TOC Correlation