Advanced turbidity prediction for operational water supply planning *

Matthew Stevenson[1] and Cristián Bravo[2,3]

[1]Southampton Business School, University of Southampton
[2]Department of Decision Analytics and Risk, Southampton Business School, University of Southampton
[3]Centre for Operational Research, Management Science and Information Systems, University of Southampton

Abstract

Turbidity is an optical quality of water caused by suspended solids that give the appearance of 'cloudiness'. While turbidity itself does not directly present a hazard to human health, it can be an indication of poor water quality and mask the presence of parasites such as Cryptosporidium. It is, therefore, a recommendation of the World Health Organisation (WHO) that turbidity should not exceed a level of 1 Nephelometric Turbidity Unit (NTU) before chlorination. For a drinking water supplier, turbidity peaks can be highly disruptive requiring the temporary shutdown of a water treatment works. Such events must be carefully managed to ensure continued supply; to recover the supply deficit, water stores must be depleted or alternative works utilised. Machine learning techniques have been shown to be effective for the modelling of complex environmental systems, often used to help shape environmental policy. We contribute to the literature by adopting such techniques for operational purposes, developing

*NOTICE: this is the author's version of the work accepted for publication in Decision Support Systems on February 26, 2019, published online as a self-archive copy after the 18-month embargo period. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. Please cite this paper as follows: M. Stevenson and C. Bravo, Advanced turbidity prediction for operational water supply planning, Decision Support Systems, 2019. https://doi.org/10.1016/j.dss.2019.02.009

a decision support tool that predicts >1 NTU turbidity events up to seven days in advance allowing water supply managers to make proactive interventions. We apply a Generalised Linear Model (GLM) and a Random Forest (RF) model for the prediction of >1 NTU events. AUROC scores of over 0.80 at five of six sites suggest that machine learning techniques are suitable for predicting turbidity peaking events. Furthermore, we find that the RF model can provide a modest performance boost due to its stronger capacity to capture nonlinear interactions in the data.

Keywords: Analytics; Water Quality; Turbidity Prediction

1 Introduction

Turbidity can be defined as the "optical quality [of water] that causes light to be scattered and absorbed rather than transmitted in straight lines through a sample" [1, p.200]. It can also be understood to be "the cloudiness of water caused by suspended particles such as clay and silts, chemical precipitates such as manganese and iron, and organic particles such as plant debris and organisms" [2, p.3].

While turbidity itself does not present a hazard to human health, it can be an indication of poor water quality. Furthermore, high levels of turbidity present during the treatment of raw water can limit the effectiveness of filtration and chlorination processes designed to remove dangerous bacteria and parasites such as Cryptosporidium [1]. It is therefore recommended by the World Health Organisation (WHO) that turbidity should not exceed a level of 1 Nephelometric Turbidity Unit (NTU) before chlorination [3].

Turbidity (NTU) levels can change slowly over time due to changes in water catchments as part of an underlying trend, but it can also rapidly peak over shorter periods, sometimes appearing random. Peaks in turbidity are linked to environmental events such as heavy rainfall but can also be a result of operational activities like pumping. The inherent solution features at the site such as fissures within the aquifer can also lead to turbidity events [2].

Peaks in turbidity (NTU) present a significant challenge to the operation of a drinking water company. Turbidity is a naturally occurring a phenomena and somewhat inevitable, however, for a drinking water supplier there are many operational interventions which impact its ability to continue to

supply potable water. Depending on the treatment works, there may be varying degrees of treatment activities used to reduce turbidity. Those most resilient to turbidity will likely include a system of filters and settling tanks that remove sediment before chlorination, but even so, these sites with more complex processes will have a limited capacity before treatment must be suspended for cleaning and maintenance. In response to short-term outages, a water supplier may rely on storage reservoirs, alternative treatment works or likely a combination of both. The challenge, however, is that turbidity peaks can occur rapidly and therefore these mitigating activities must be actionable immediately; storage reservoirs will need sufficient supplies, and alternative sources will need to be able to meet the new additional requirement caused by outages. Failure to do so will result in the company unable to meet its demand, or, water entering supply that is not fit for consumption; either instance would be damaging for a water supplier and its customers.

We propose a decision support system that provides drinking water suppliers 7-days notice of a turbidity event, allowing time for remedial actions to be prepared in advance of short term outages caused by turbidity peaking events.

Our first objective is to explore the cause of daily turbidity (NTU) peaking events by identifying candidate predictor variables which we test across each of the six sites. We use this to confirm relevant variables from the literature, but also further explore how operational features such as pumping activity impact on turbidity levels at a treatment works. We apply a static correlation analysis and a dynamic cross-correlation analysis which considers time lags of some variables.

Our second objective is to assess the effectiveness of models from the field of machine learning for the prediction of >1 NTU events. We use a Generalised Linear Model (GLM) and a non-linear Random Forest (RF) model to predict turbidity across the six treatment works. We use a linear and non-linear model to assess the impact of any non-linearity that may exist in the data; furthermore, they represent different aspects of the trade-off between complexity and predictive capability [4]. We use the AUCROC score to asses the performance of the models, models with a score of greater than 0.70 are considered satisfactory. The causation analysis is complemented using the Variable Importance outputs from the GLM and the RF. We also review the cut-off probability points for event classification.

To address the research problem, we first review how machine learning has been applied to predict a range of other water quality parameters in the

literature, and, identify causal factors in turbidity peaking. We then the illustrate the behaviour of turbidity peaking across the six sites and use the static and dynamic correlation analysis to determine candidate variables for the models. We consider the results in three parts: 1) the model performance of the GLM and RF models are reviewed using the AUROC metric, 2) the Variable Importance outputs of the models are examined to understand the multivariate nature of turbidity prediction, complementing the earlier correlation analysis and, 3) we then use a cost-based approach to define the cut-off probability points for event classification for each of the sites. We conclude by reflecting on the general findings for turbidity causation compared to that of the literature and assess the viability of an operation turbidity prediction model as a decision support system.

2 Background

Techniques from the field of Machine Learning have been applied to solve a wide range of event prediction problems [5, 6, 7, 8]. In this section, we seek to understand how statistical and machine learning tools and techniques have been applied to understand and solve a variety of challenges surrounding water quality. Some of the research is focused upon causal analysis while other research attempts to predict and model systems to be tested under different conditions, this, in turn, can be used to direct the management policy of natural systems including lakes, rivers and dams. We note that this differs from our research which is of the perspective of the operational management of a drinking water supplier. We seek to understand how both natural and operational features can be used in combination to predict future turbidity events that can be implemented within a live operational system. That said, there are many insights that we can draw from the field of hydroinformatics that we exploit in our modelling.

Linear models have been used to predict water quality parameters which have provided useful insights into the behaviour of a natural system, however, have demonstrated relatively modest results. LeChevallier et al. [9], undertook a comprehensive review of 66 water treatment plants across Canada and the United States to understand the occurrence of Giardia and Cryptosporidium organisms in the raw water supplies. A linear regression analysis was applied to review which individual characteristics of raw water sources such as turbidity levels, coliform levels, faecal coliform, temperature, pH and total

coliforms were influencing Giardia and Cryptosporidium levels in the source waters. The linear regression on single variables showed a significant relationship between Turbidity and Total Coliforms for Giardia. A significant relationship between Turbidity (NTU) and Cryptosporidium levels was also present. The final multiple linear regression (MLR) models had corrected R^2 scores of 0.491 and 0.468 for Giardia and Cryptosporidium respectively. In this instance a linear model has demonstrated what seems relatively modest results, we intend to use a GLM to confirm if a linear model can provide us with an adequate decision support system, or if more advanced nonlinear models can provide a superior result.

Other research has applied correlation analysis to identify if factors such as land use can influence water quality parameters. Tong and Chen [10] undertook a comprehensive study to look at how land use in conjunction with hydrological factors could be used to explain the variance in many water quality parameters. The study focused on surface water quality influenced by surface 'run-off' and determined that, depending on the land use, the runoff may contain many different contaminants. The research considered five land use types; Urban, Forrest, Agriculture, Barren Land and Water[body]. Using Spearman's rank correlation analysis for watersheds in Ohio, the results demonstrated that Faecal coliform levels have a strong positive correlation with the commercial, residential, and agricultural land. Agricultural landuse was strongly correlated with conductivity and PH. The residential and commercial land was related to sodium, cadmium, lead, conductivity biochemical oxygen demands (BOD) and zinc. These findings provide evidence there can be significant differences between geographical areas, and since the six sites included in our study are not co-located, it also suggests we will require the development of separate models for our sites.

Random Forests (RF) have been shown to be suitable predictors for nonlinear data while providing useful insights. Read et al. [11] sought to apply RF models to understand how lake water quality across the US is affected by both regional factors and lake specific drivers. The models attempted to predict several water quality metrics including total phosphorus, total nitrogen, dissolved organic carbon, turbidity and conductivity. The inputs to the model included eleven predictor variables in total, including regional, basin and lake specific variables. Regional features included land type (forest/crops/agriculture), basin features included elevation and land coverage (forest, conifer) and lake specific variables included maximum depth, sediment to volume ratio, latitude, longitude and elevation. The models pro-

duced what was considered promising results with the combined model of all variables explaining 61%-66% of the variance. The model using a subset of just the lake specific variables explained 54%-60% of the variance for four of five of the predictions. The RF models were deemed to be successful by Read et al. [11], not just for their predictive success, but also for the practical insights provided by examining the Variable Importance metric of the models. Furthermore, the Random Forest models were considered robust to the multi-collinearity and non-normal distributions of the predictor variables and additionally effective for controlling overfitting.

Like Random Forests, Artificial Neural Networks (ANNs) have been shown to be effective at predicting water quality parameters where the data is non-linear, noisy and the statistical relationships between inputs and outputs are not well understood. Muttil and Chau [12] applied an ANN model to predict harmful algal blooms in Tolo Harbour, Hong Kong. The models used eight input variables consisting of hydrogeological and weather factors including; chlorophyll-a, Chl-a (µg/L), total inorganic nitrogen, TIN (mg/L); phosphorus, PO4 (mg/L); dissolved oxygen, DO (mg/L); secchi-disc depth, SD (m); water temperature, Temp (°C); daily rainfall, Rain (mm); daily solar radiation; SR (MJ/m^2) and daily average wind speed; WS (m/s). The variables were lagged seven days creating 63 input variables in total to capture the dynamic element of algal blooms varying over time. A study of the variable significance of both models suggested that chlorophyll-a, the measurement of algal blooms itself, was sufficiently significant that it could be used to model future events independently due to its autocorrelative nature. The ANN model was considered suitable for predicting long-term trends of algal blooms however the performance was relatively weak at short-term prediction. The authors attributed the weakness in predicting short-term trends to the fact only twice-weekly data was available. This research demonstrates how more advanced analytical techniques were successful for prediction with noisy and non-linear data. Furthermore, it demonstrates how the dynamic nature of the data could be captured using lagged predictor variables, interestingly including lags of the dependent variable which may also be relevant for turbidity prediction.

In addition to algal bloom prediction, ANN models have also been used to predict a broader array of water quality parameters. Najah et al.[13] developed six ANNs to predict three water quality parameters for the Johor River and its tributary river. These parameters included electrical conductivity, dissolved solids and turbidity; however, this study does not consider

turbidity peak prediction or the application of dynamic time-dependent variables. The ANN models could predict the water quality parameters to an accuracy of 10% average mean percentage error (AMPE). Other studies such as that by Elhatip and Kömür [14] have also found ANN models to be accurate for water quality prediction. In their study, ANNs were used to predict changes in electrical conductivity (EC) and dissolved oxygen (DO) with the aim of creating a model for surface water management of the Mamasin dam in Turkey. The models produced a Mean Average Percentage Error (MAPE)of 6.46% and 4.72% for EC and DO respectively. These two studies have shown again that advanced non-linear techniques such as ANNs can be used to predict an array of water quality parameters. Both these studies, however, are used to model a natural system, seeking to mimic these systems under different conditions and identify environmental management policies, they are not decision support systems to be used as part of day-to-day operations for a drinking water supplier as our study aims to.

The literature presented so far has only considered water quality parameters for surface water sources such as rivers, reservoirs and lakes. Furthermore, all have modelled water quality at a static snapshot in time or as annual averages rather than as a time series, except for the research undertaken by Muttil and Chau [12]. The purpose of this paper is to model turbidity (NTU) events seven days in advance to a daily frequency for groundwater sources. The majority of the sites we model are located within a karst landscape whereby karst is defined as "distinctive landforms that develop on rock types such as limestones, gypsum and halite that are readily dissolved by water... typically characterised by a lack of permanent surface streams and the presence of swallow holes and enclosed depressions" [15, p.42].

Perhaps the most comprehensive time series study into turbidity (NTU) within a karst aquifer was undertaken by Massei et al.[16]. The research sought to investigate particle transportation using cross-correlation analysis, spectral analysis and wavelet analysis for a spring source within a chalk aquifer in the lower Seine Valley, France. They reflected on the behaviour of turbidity (NTU) in the natural spring source highlighting that the system behaves differently in the long-term and short-term. In the long-term, the system behaves linearly as the aquifer is slowly infiltrated on mass through the 'karstic system' with surface water. In the short-term, however, turbidity (NTU) behaves 'extremely' non-linearly due to hydrogeological 'quick-flow' features in the aquifer such as sinkholes providing a fast route for turbidity (NTU) to reach the source. The study considered three lagged input vari-

ables in the analysis; rainfall, water level and conductivity and their 'short-term' and 'long-term' memory effects on turbidity (NTU). The relationship between these three variables and turbidity was found to be a complex one. Rainfall had the greatest effect on turbidity in the short-term which is caused by soil erosion and surface runoff entering 'quick-flow' features of the aquifer. Rainfall was also shown to have some effect on water level and turbidity in the longer term (up to 35 days), this was attributed to rainfall flooding the karstic system and re-suspending particles. The wavelet analysis included the deconstruction of water level into a long-term smoothed trend element capturing a small amount of the variance and a short-term element capturing a more substantial amount of the variation in water level. The longer-term water level element had a smaller influence on turbidity but reflected the slower rising and falling of turbidity levels as particles work through the karstic system. The short-term/higher frequency element of water level had a clear connection with the short-term fluctuations in turbidity (NTU) attributed to surface water passing through quick-flow features. The study demonstrates that the interactions between the rainfall, water level and turbidity are complicated, the data is noisy, non-linear and time-dependent which we consider for the build of a turbidity classification model.

3 Methodology

3.1 Data Description

Turbidity, the dependent variable, is obtained for each of the six sites from the telemetry system of the water company. Turbidity (NTU) level is recorded at least every 15 minutes using apparatus located at the water treatment works, for this paper, only the daily maximum NTU level is required as it always reveals whether turbidity has exceeded 1 NTU in each day. The record for most sites extends back to at least 1 November 2007 up to the point of extraction on to 15 September 2017. Therefore, there are 3606 days/data points per site before the treatment of outliers and combination with other data sets.

Groundwater levels have been shown to influence turbidity levels [16]. Like turbidity, the water company also records the water level in the borehole sources a minimum of every 15 minutes. In this paper two sources of level data are included in the analysis; 1) level in metres (m) at the source

and 2) level Metres Above Ordnance Datum (mAOD) at the company level monitoring borehole at the centre of the company resource zone. Water recorded at the source is measured 'top-down' whereby the measurement is the distance between the water level and the borehole head pit. Since the relationship with turbidity is most likely related to the lowest level of the water in the borehole in each day, we use the daily maximum for local borehole data representing the maximum air gap between the head pit and the water level in each day. The water company also has a central monitoring borehole specifically for recording level which has a long and consistent record. Unlike the local level (m) data, level (mAOD) at the central monitoring borehole is measured 'bottom-up' from sea level to the borehole water level; therefore, we use the minimum water level (mAOD) instead. We use both the central level (mAOD) and site-specific level (m) in the data exploration as candidate variables where we use the more significant of the two variables in the modelling phase. For the two spring sites Site-D and Site-E, the central monitoring borehole is local and therefore we only use the central level (mAOD) in these two instances.

In addition to level data, we also utilise other data relating to the operation of each of the sites. Each site has at least one pump used to lift the water from the borehole or spring. For the daily pump record, we use a binary 1/0 variable $pumpX$ to represent if the pump had been running. Flow (l/s) is also recorded by the water company at each of the sites and represents the rate at which a sourceworks has abstracted water. We obtain flow (l/s) in daily maximums/minimums. We anticipate that flow (l/s) acts as a proxy for how aggressive the pumping regime had been in each day which is anecdotally linked to increased turbidity (NTU) at some sites. Flow and pump data is limited or erroneous at several of the sites, and therefore it is only included it if a sufficiently long and reliable record is available.

We obtained Rainfall (mm) data from the Centre for Ecology & Hydrology [17]. The full CEH-GEAR dataset includes 1 km-gridded estimates of daily rainfall (mm) for Great Britain and Northern Ireland which are derived from the Met Office national database of observed precipitation, the most recent copy of the file at the time of writing extends from 1890 to the end of 2015. We extract the rainfall (mm) data from the national data set for the location of each site.

The daily telemetry and rainfall (mm) data are combined for each site to create a record extending from 2007 to the end of 2015, typically 2982 data points per site.

Data	Units	Summary	Type	Source
Turbidity	NTU	Max	Continuous	Water company telemetry
Water Level (Local)	m	Max	Continuous	Water company telemetry
Water Level (Central)	mAOD	Min	Continuous	Water company telemetry
Flow	l/s	Min/Max	Continuous	Water company telemetry
Pump Operation	1/0	NA	Categorical	Water company telemetry
Rainfall	mm	Total	Continuous	Centre for Ecology & Hydrology (CEH)

Table 1: Summary of raw data

3.2 Data Treatment

Turbidity (NTU), Water Level (m, mAOD), and Flow (l/s) data is reliant upon instrumentation at the abstraction site and can be subject to false readings. Furthermore, the equipment can be recalibrated over time causing step changes in the data thus inducing errors. We use a distribution based analysis to identify and remove erroneous data. Based on the distribution, we remove extreme outliers from the dataset as they likely indicate false readings while we cap near outliers at an appropriate level for the site.

Flow (l/s), Level (m, mAOD) and Turbidity (NTU) data are time dependent therefore outliers with temporal change are detected, removed and imputed using time-series based methods. While the level (mAOD) data for the central monitoring borehole is well maintained with few errors, the local level (m) recordings contain both erroneous and missing data. Where local level (m) data is erroneous, we remove the data. If the level (m) data is missing for more than seven consecutive days, a linear regression model is used to impute the data rather than the time series based analysis as the time series imputation produces unreliable results over large sections of missing data. The linear regression models use flow (l/s) and the central level (mAOD) to predict the local level, typically these models work suitably well and achieve R^2 scores over 0.70. Across the sites, level was imputed in less than 5% of all days.

For missing data relating to pump operation ($pumpX$), we apply a Multivariate Imputation by Chained Equations (MICE) approach [18]. In cases where more than 10% of the pump operation data was unavailable, the variable was screened out.

3.3 Descriptive Statistics

Table 2 and Table 3 show a summary of the extent of the levels of turbidity (NTU) at each site. Site-D and Site-E experience the highest recorded levels of turbidity with recorded levels of 8 NTU and 9 NTU respectively. These two sites also have by far the highest frequency of events where turbidity is greater than 1 NTU with Site-E experiencing such levels in 33% of all days. The water company manages the high turbidity levels at Site-D and Site-E by applying additional processing when the turbidity exceeds 0.7 NTU at a subsequent treatment plant to reduce the level of turbidity in the water before chlorination.

There are some notable differences between the mean and median statistics with the mean larger than the median in all instances. The turbidity data is therefore non-normally distributed and positively skewed presenting a challenge for the development of a predictive model.

	Site-A	Site-B	Site-C	Site-D	Site-E	Site-F
Min	0.03	0.00	0.02	0.01	0.09	0.02
Q1	0.07	0.04	0.04	0.34	0.35	0.08
Median	0.09	0.08	0.06	0.58	0.62	0.15
Mean	0.19	0.31	0.10	1.14	1.33	0.28
Q3	0.18	0.39	0.08	1.24	1.38	0.32
Max	1.30	2.40	1.20	8.00	9.00	2.20

Table 2: Summary Statistics of Turbidity (NTU)

Site	Non-Event	Events	%
Site-A	2838	154	5%
Site-B	2735	257	9%
Site-C	2867	54	2%
Site-D	2123	933	31%
Site-E	2056	1000	33%
Site-F	2448	108	4%

Table 3: Event frequency by site

Behind Site-D and Site-E, Site-B has the most significant recordings of turbidity of up to 2.4 NTU and experiences a turbidity event in 9% of all

days.

Site-A, Site-C and Site-F have a >1 NTU turbidity event in less than 5% of all days over the recorded period and have relatively low maximum recordings compared to other sites with values of less than 2.2 NTU.

Higher event frequencies at Site-D and Site-E can likely be attributed to the location of these sites within the aquifer. The sites sit closer to the coastal plain and are within an area deemed more vulnerable to sediment which can easier permeate the surface and work its way to these treatment works. As a result, we might expect rainfall to have a more significant impact on turbidity as it is rainfall that transports the sediment to the treatment works. Site-A also has similar geological properties however seems to experience fewer events.

Figure 1 provides some insight into the behaviour of turbidity as a time series at each of the six sites with the red line indicating the 1 NTU level of turbidity. All of the sites demonstrate some form of seasonality with turbidity rising and falling over time. Additionally, the spiking events tend to occur close together.

This pattern is most prominent at Site-D and Site-E where the turbidity level is typically rising towards the beginning or at the end of each year, possibly reflecting the rising and falling of groundwater levels. Site-D, Site-E and to a lesser extent Site-A experience turbidity events broadly around the same periods. Site-A, Site-B, Site-C and Site-F have a noisier time series than Site-D and Site-E with long periods where turbidity events >1 1 NTU do not occur.

Notably at Site-B, although there are periods where turbidity (NTU) remain low when turbidity (NTU) exceeds 1 NTU it tends to remain high for an extended period giving signs of autocorrelation that we can exploit in the predictive modelling.

3.4 Static Correlations

We apply a static correlation analysis to explore the relationships between the independent variables and turbidity (NTU) at lag 0 so that we can derive a subset of candidate variables for each site.

For the correlation analysis, turbidity (NTU) is treated as a raw continuous variable rather than in a binary form. As turbidity (NTU) has been shown to have a non-linear relationship to variables such as rainfall (mm)

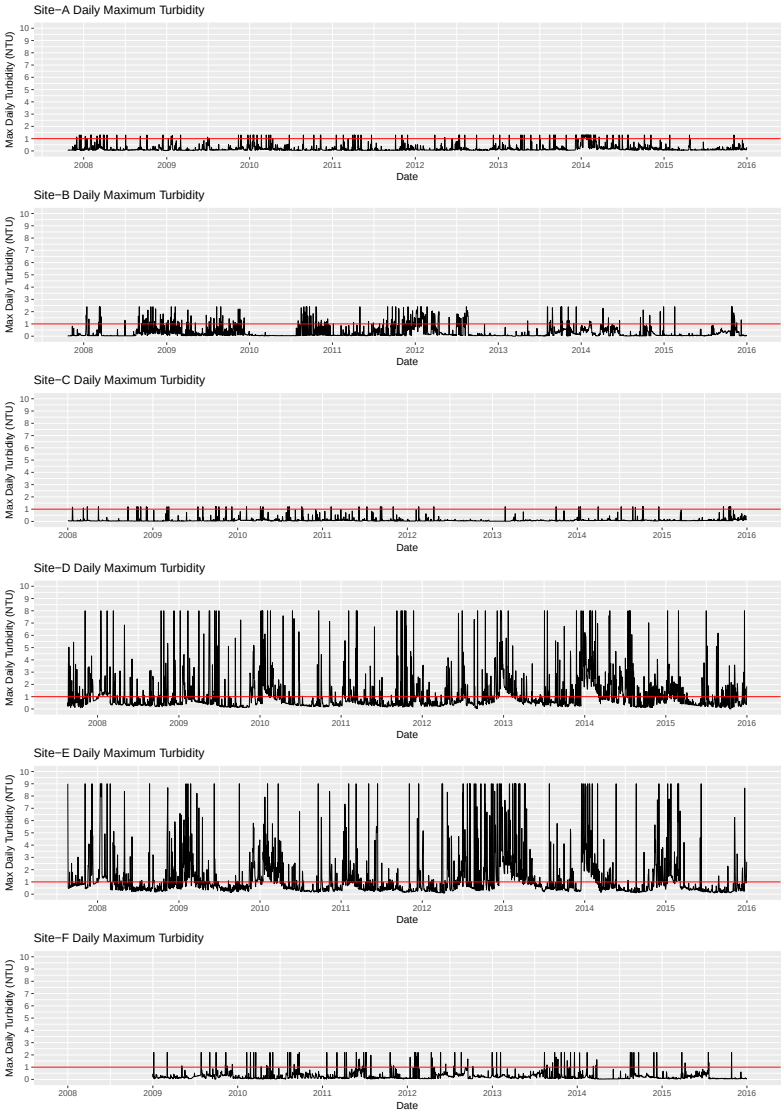

Figure 1: Time series Plots of Turbidity (NTU)

[16], a Spearman's rank correlation coefficient is derived due to the nonparametric nature of the measure [19].

Figure 2 presents the results of the correlation analysis. The blue colouring indicates positive relationships while the red colouring indicates a negative relationship. Crosses presented in the circles can be interpreted to mean

there is no significant relationship at a p=0.05 level.

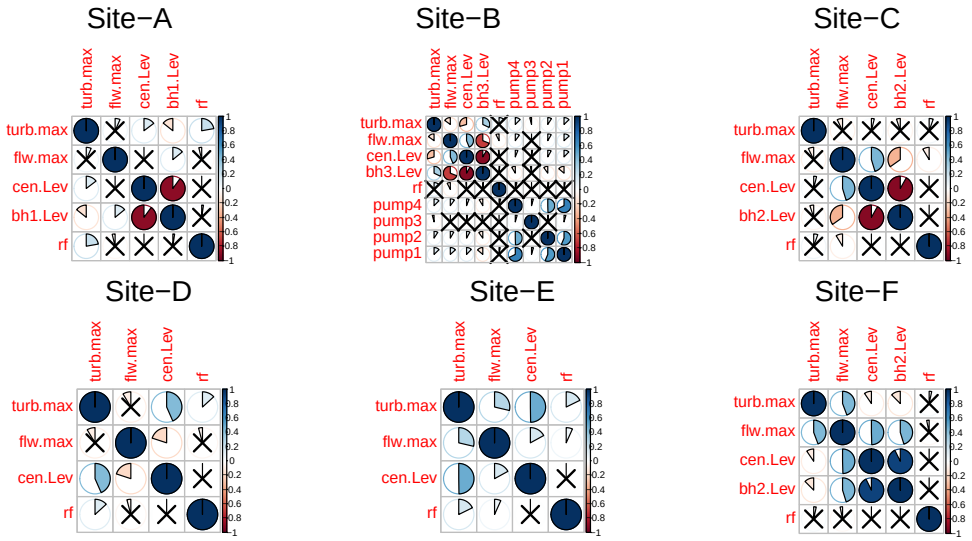

Figure 2: Static Correlations by Site

The static correlation analysis reveals that turbidity (NTU) responds to the candidate predictor variables very differently across the six sourceworks. Additionally, while at some sites specific variables demonstrate a more explicit relationship with turbidity (NTU), at no site does anyone one variable show a strong correlation. At most sites, there tended to be a small relationship with level (m,mAOD) and rainfall (mm) while at site E, we observe a modest relationship with Flow (l/s). Notably, at Site-C, no significant correlations with turbidity (NTU) are present.

3.5 Dynamic Correlations

We consider autocorrelation for turbidity (NTU) and cross-correlation for water level (m, mAOD) and rainfall (mm) with a lag of up to 35 days, the period shown to be significant for rainfall (mm) in other karstic systems [16].

3.5.1 Rainfall and Turbidity

Rainfall can affect turbidity in two ways. In the short term, rainfall events can cause surface waters to reach a site through 'quick-flow' features in the aquifer that cause turbidity to respond non-linearly as particles in the aquifer

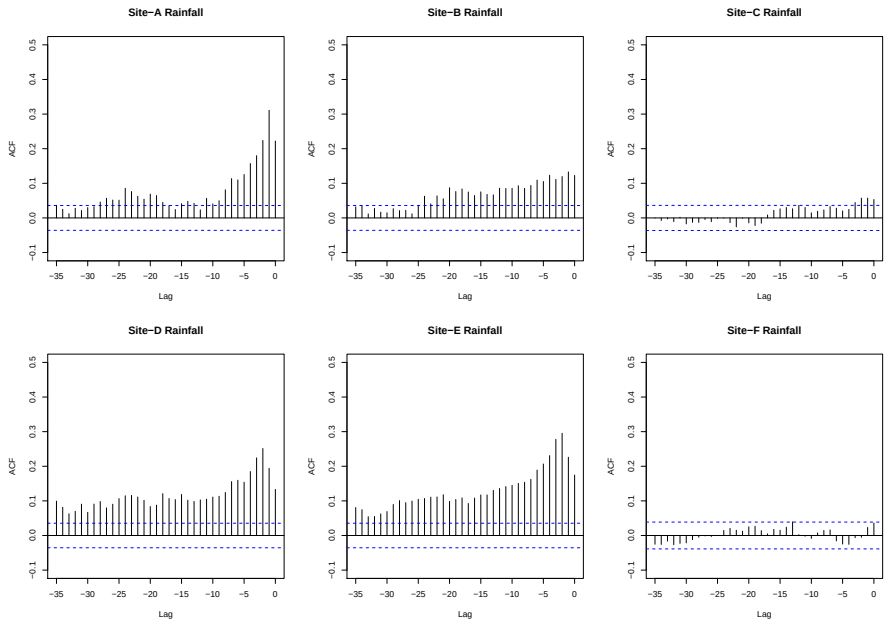

Figure 3: Turbidity (NTU) and Rainfall (mm) Cross Correlation

are re-suspended. Over the longer term, up to 35 days, rainfall reaches a source through turbidity via slower karstic flows resulting in a smooth linear rising and falling of turbidity [16].

Rainfall (mm) was shown to have significant but weak positive lagged correlations with turbidity (NTU) at all sites except for Site-F. The presence of lagged correlations is significant to the modelling process as it suggests that the effect of rainfall (mm) on turbidity (NTU) cannot necessarily be traced to rainfall (mm) on one specific day, but instead it could have a cumulative effect.

Site-D, Site-E and Site-A demonstrate a clear relationship between turbidity (NTU) and rainfall (mm) reflecting the effect of the quick-flow (l/s) features of the karstic aquifer. At Site-D and Site-E the most significant effect of rainfall (mm) was at lag 2 while at Site-A, and to some extent Site-B, lag 1 was most significant.

The sites also respond differently to rainfall (mm) regarding how long the effect of rainfall (mm) on turbidity (NTU) remains significant across the lagged window. At Site-D and Site-E the relationship between rainfall (mm) and turbidity (NTU) exponentially decays after lag 2 but remains significant

throughout the 35-day window. At Site-A the correlation decays linearly and quickly from lag 1 and ceases to be significant at lag 12. Site-B has a weaker correlation but for a relatively extended period where rainfall (mm) remains significant eventually becoming non-significant at lag 25.

At Site-C and Site-F the relationship between turbidity (NTU) and rainfall (mm) is less obvious. Site-C demonstrates a very weak positive correlation which remains significant for just 3 lags while at Site-F rainfall (mm) up to 35 days shows no significant relationship or apparent pattern. Anecdotally, this may be linked to the type of aquifer that these two sources sit within that make them less susceptible to short-term rainfall events.

3.5.2 Level (m, mAOD) and Turbidity (NTU)

The cross-correlation between turbidity and water level is presented in Figure 4. The difference between whether the relationship is positive or negative is a reflection of the measurement of level only; the top down measurement (m) is negative while the bottom-up (mAOD) is positive.

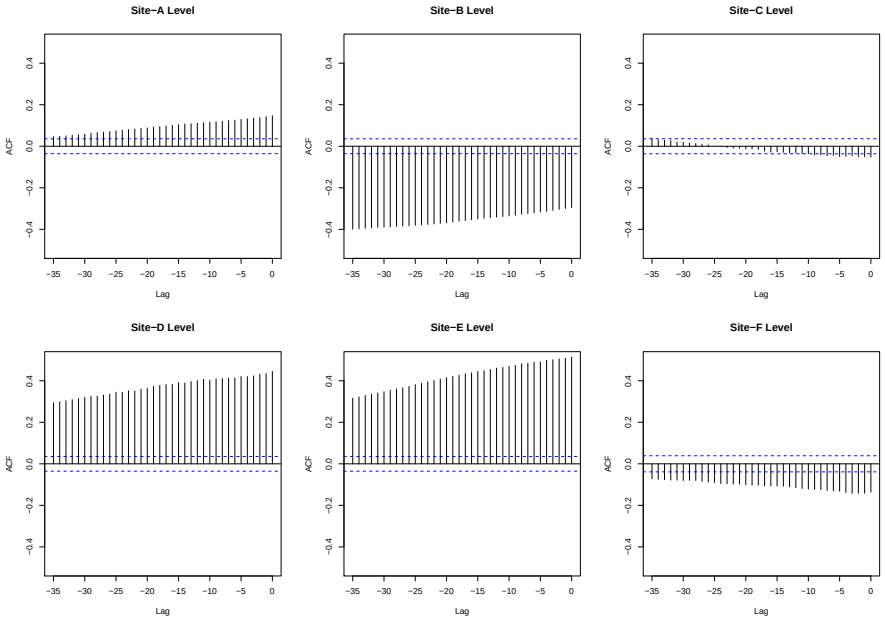

Figure 4: Turbidity (NTU) and Level (m, mAOD) Cross Correlation

Rainfall causes surface water to permeate the karstic system and in turn,

causes the groundwater levels to rise and fall accordingly; level is typically highest at the beginning of the year after a prolonged period of winter rainfall and lowest in the autumn after drier summers. The water level has been shown to influence turbidity as a result of surface water containing particles moving slowly through the karstic system [16].

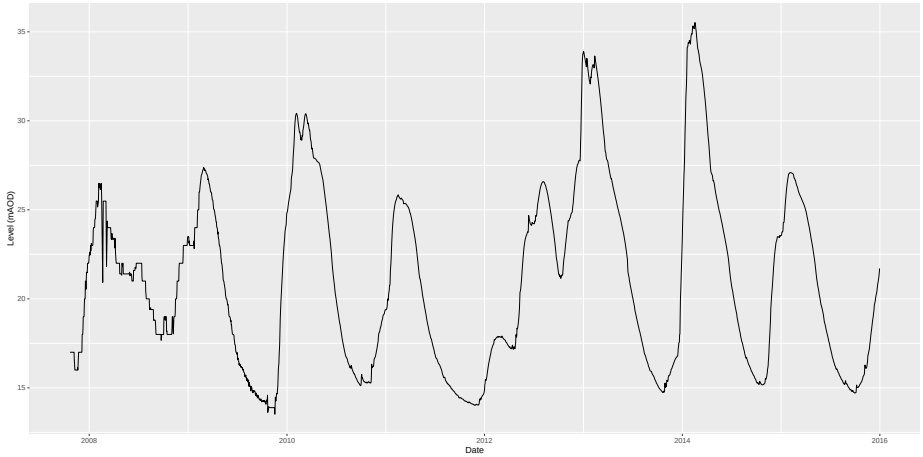

Figure 5: Central Level (cenLev) 2008 to 2016

All sites demonstrate a significant relationship between lagged level (m) and turbidity (NTU) up to the full 35-day period except Site-C which has a weak correlation with a lag of up to 10 days.

Site-D and Site-E both demonstrate the strongest correlation with level (m) and turbidity (NTU) of the six sites with a smooth linear decay. Site-B also has a moderate relationship with lagged level (m) and interestingly has a reversed effect to the other sites where further lags have a more significant response, this could be a result of the difference in distance and geology between the central monitoring borehole compared to that of the site itself. At Site-A and Site-F, the relationship between lagged level (m) and turbidity (NTU) is relatively weak but remains significant up to 35 days.

The smooth linear decay of in the correlation between level (m, mAOD) and turbidity (NTU) at the sites may be an indicator of the 'slow-flow' features of the aquifer causing turbidity (NTU) to rise and fall. It may also indicate that later lags provide no further information than earlier lags.

3.5.3 Auto Correlation of Turbidity

The autocorrelation for turbidity at each site is presented in Figure 6.

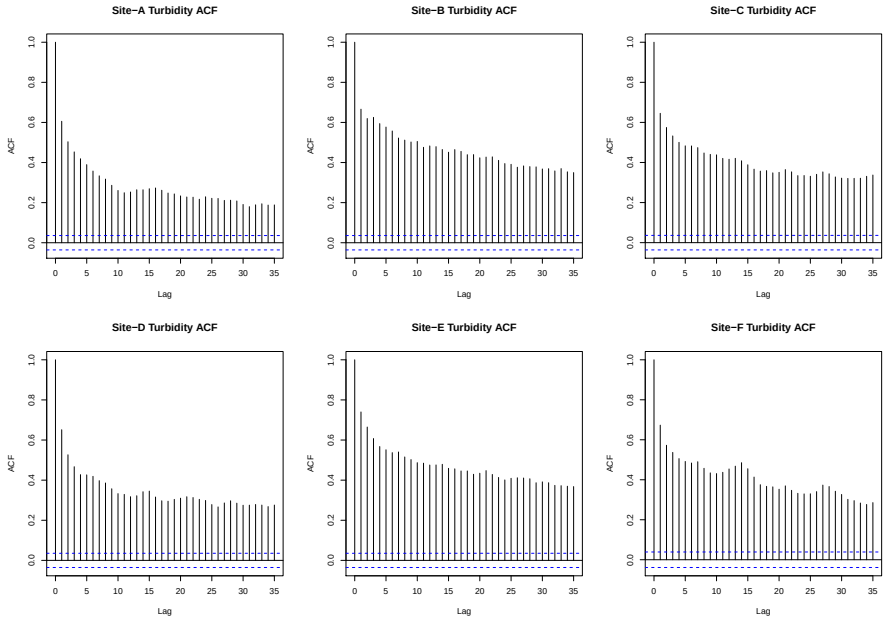

Figure 6: Autocorrelation of Turbidity (NTU)

At all sites turbidity (NTU) is auto-correlated and significant up to 35 days as can be seen at most sites in the time series plots with the rising and falling turbidity (NTU) over time. Each of the sites responds in a similar manner whereby there is an exponential decay in the correlation. The correlation tends to fall away very quickly in the first few lags before we observe a smooth linear decay. Site-A differentiates slightly from the other five sites as there is a comparatively slower decay in the autocorrelation up to lag 10 when the correlation begins to level.

The higher correlations in the first few lags followed by a linear decay may suggest that after turbidity levels have risen initially after a peaking event, it takes several days for turbidity (NTU) level to fall back to background levels.

4 Modelling

For the prediction of >1 NTU turbidity events, we apply a Random Forest (RF) and a Generalised Linear Model (GLM).

Several parameters require tuning in an RF model; *mtry*, *ntree* and a cost rule by which predictions are made in the training of each tree.

As recommended by Breiman [20], we use the square root of the total number of variables *mtry* and this figure halved and doubled.

The number of trees used in the model is referred to as *ntree*. Breiman [21] argues that the more trees there are in the model, the lower the generalisation error. Furthermore, increasing the number of trees does not cause the model to overfit, an advantage that the RF algorithm has over other tree-based methods. We use a value of 1000 for *ntree* as this provides an adequate number of trees with a practical runtime [22].

The turbidity data set at each site is imbalanced whereby 2% to 33% of all data points experience turbidity (NTU) levels of less than 1 NTU depending on the source works. This imbalance can lead to the accuracy prediction metric to bias the more dominant class [23]. We use the Synthetic Minority Oversampling Technique (SMOTE) approach to artificially balance the dataset as this method has been shown to achieve superior AUROC compared to other sample balancing techniques [24].

We also apply a logistic regression alongside the RF model. We use the Glmnet algorithm which uses a penalised maximum likelihood to fit the logistic regression [25]. We make all variables available to the model since the Glmnet is efficient and can effectively screen irrelevant variables. The purpose of the GLM is to provide a baseline performance of discriminative ability for means of comparison against the RF. Furthermore, the variable importance output from the GLM is also used to provide a multivariate insight into the cause of turbidity prediction for each of the sites. We use the same training and test dataset for the GLM and RF. Given that turbidity responds non-linearly to rainfall, we expect the RF to outperform the GLM given the linear constraints of the model, at least where rainfall is shown to be a significant factor leading to a turbidity peaking event.

5 Results

5.1 Model Performance

We present the AUROC performance of the RF and GLM models at each of the six sites in table 4. We use a randomly selected 25%/75% test/train split with a 10 k-fold cross-validation and 10 repeats. For the cross-validation results, we present the mean AUROC across the 100 samples and the standard deviation.

Table 4: AUROC of Generalised Linear Model (GLM) and Random Forest Model (RF)

	Training				Holdout	
	GLM	GLM Std. D	RF	RF Std. D	GLM	RF
Site A	**0.74**	0.07	0.70	0.09	0.81	**0.84**
Site B	0.75	0.05	**0.77**	0.05	0.81	**0.84**
Site C	**0.58**	0.12	0.56	0.13	**0.81**	0.61
Site D	0.77	0.03	**0.81**	0.03	0.79	**0.81**
Site E	0.87	0.02	**0.89**	0.03	**0.86**	**0.86**
Site F	0.61	0.09	**0.69**	0.03	**0.69**	**0.69**

At five of six sites (Site-A, Site-B, Site-C, Site-D, Site-E), we obtained an AUROC score of over 0.80 in the holdout sample suggesting that these models have a 'good' discriminative ability. In all four instances, the most successful model was the Random Forest although the difference in AUROC performance when compared to the GLM is often marginal. The relatively small difference in performance between the RF and GLM may suggest that non-linearity may not be very strong. We note that there are some variations between the cross-validation and holdout samples, especially so at Site-C, though in all instances the result falls within the 95% confidence range.

The RF model provides the highest performance for Site-A in the holdout sample; however, the cross-validation scores suggest that the GLM performs better on average. Coupled with a higher standard deviation for the RF model at Site-A, this indicates that the RF can obtain a good AUCROC, but the model is more sensitive to the sample.

For Site-C the GLM model dramatically outperforms the RF model when

tested on the holdout sample, however, in the cross-validation sample, the performance is similar with both having notably high standard deviations. The significant difference in the holdout result, and, a significant variance in the cross-validation results suggest the models are sensitive to what is included in the sample with the two models responding differently according to the model inputs.

For Site-F, both models have an AUCROC of 0.69 in the holdout sample suggesting that the models have a 'fair' discriminative ability, in the cross-validation samples, however, we note that the RF model provides a larger and more stable AUCROC score.

5.2 Variable Interactions

From both the GLM and the RF model useful insights can be obtained as to the multivariate nature of turbidity causation. While in the static and dynamic correlation analysis turbidity (NTU) was considered in continuous form, in this section we review the predictor variables concerning their association with turbidity (NTU) in binary form. Events are classified as either greater than or equal to 1 NTU (*tEvent*) or below 1 NTU (*nonEvent*).

The multivariate nature of turbidity prediction is considered using the scaled variable importance output of the GLM and Random Forest model. For the GLM the scaled importance is the normalised coefficients of the final model while the Random Forest uses the 'Mean Decrease in Gini' which reflects the average reduction of the impurity within each node in the model when a node is split using a given variable [26]. We present the top 15 most important variables for each model. Since both models are capable of filtering insignificant variables, all variables are made available to the models, variable importance, therefore, provides insight as to which variables remain important, or become more important, in the presence of other predictor variables.

Between the two models, there appears to be some universal agreement which mostly reflects the findings of the correlation analysis. For those sites where a correlation between rainfall (mm) and turbidity (NTU) is present, rainfall (mm) was also shown to be important. However, while rainfall (mm) in the correlation analysis remained significant up to the maximum 35 lags, the variable importance demonstrates that rainfall (mm) is most significant within the first several lags depending on the site. Greater importance in earlier lags likely shows the quick-flow elements of turbidity (NTU) Massei

et al. [16].

We also find that at those sites where both level (m, mAOD) and turbidity (NTU) lags are present, the variable importance top 15 ranking tended to be dominated by lags of turbidity (NTU) rather than level (m, mAOD). Although the dynamic cross-correlation of level (m, mAOD) and turbidity (NTU) showed a moderate relationship to be present at most sites, the variable importance suggests that with the presence of turbidity (NTU), level (m, mAOD) becomes less important. Furthermore, level (m, mAOD) and turbidity (NTU) lags both capture the slow-flow (l/s) features of the karstic system representing the rising and falling of turbidity (NTU) that can contribute towards a turbidity (NTU) peaking event. Furthermore, it could be that while both variables capture the rising and falling of background levels of turbidity (NTU), lagged turbidity (NTU) provides a more significant amount of information than level (m, mAOD). Lagged turbidity (NTU) may provide more information as it is also capturing the residual turbidity (NTU) that exists after an initial peak in turbidity (NTU).

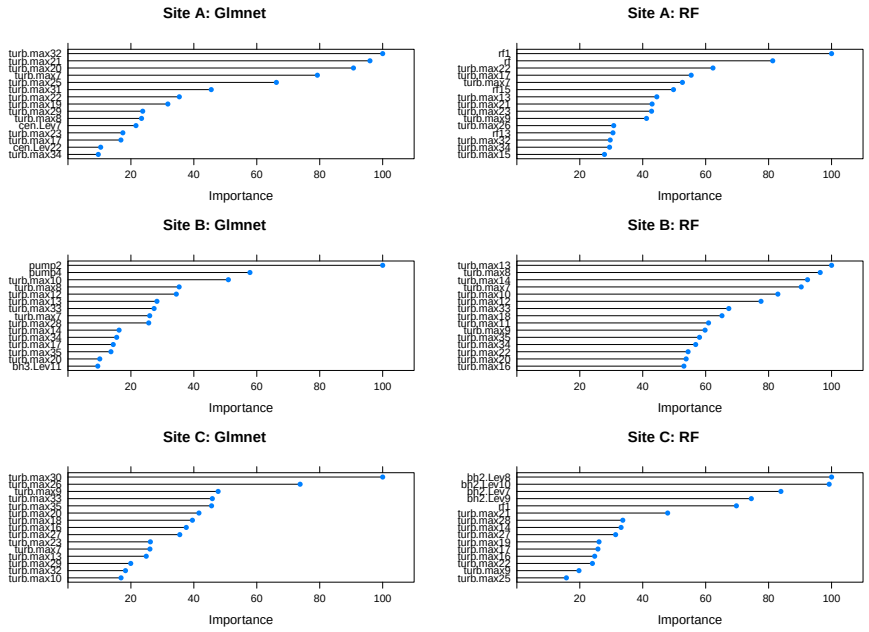

Figure 7: Variables Importance: Sites A-C

The high importance lags of rainfall early in the lag window likely reflect the 'quick-flow' features causing spikes in turbidity (NTU) while level

(m/mAOD) and turbidity (NTU) lags in the longer term causes a smooth rise and fall of turbidity (NTU) more closely representing the karstic flows in the aquifer. We suggest, therefore, that while the seasonal rising and falling turbidity has some contribution towards a turbidity event, in most instances, the event is most likely caused by rainfall events. This is not true at all sites, however, at Site F the RF variable importance is dominated by flow (l/s) while at Site B the GLM importance suggests that the operation of pump 2 and pump 4 have the most significant contribution.

From the variable importance, we can also begin to understand why in some instances the RF model can outperform the GLM. At Site-D and Site-E, which have the most significant occurrence of events, the RF model consistently outperforms the GLM. The RF model puts far more importance on rainfall which is known to interact non-linearly with turbidity in the short term; we, therefore, expected that the RF being the non-linear model would be able to outperform the GLM. At Site-F, the RF performs better in cross-validation. A review of the variable importance shows that while the RF model attributes the most importance on flow (l/s), this variable does not appear in the top 15 most important variables for the GLM suggesting there is also a nonlinear relationship between flow (l/s) and turbidity (NTU).

6 Implementation of the Decision Support System

So far we have considered the performance of the models across the six sites in terms of AUC performance. We now consider at which probability, from 0.00 to 1.00, that the decision support system positively classifies an event and the water company takes mitigating steps. Mitigating actions might include the decision ensure that the reservoirs are full before an event, or, committing personnel to test the equipment at an alternative site to while pumping at the site in question is temporarily suspended.

To determine an optimal cut-off point for each of the models, we apply a cost-based approach to quantify the trade-off between the False Positive (FP) and False Negative (FN) rate. A cost-based approach assigns a value to the prediction of a FP against the prediction of a FN.

Through consultation with the water supply company, we apply a simple costing rule. We assign a cost of predicting a FN as twice as expensive as the

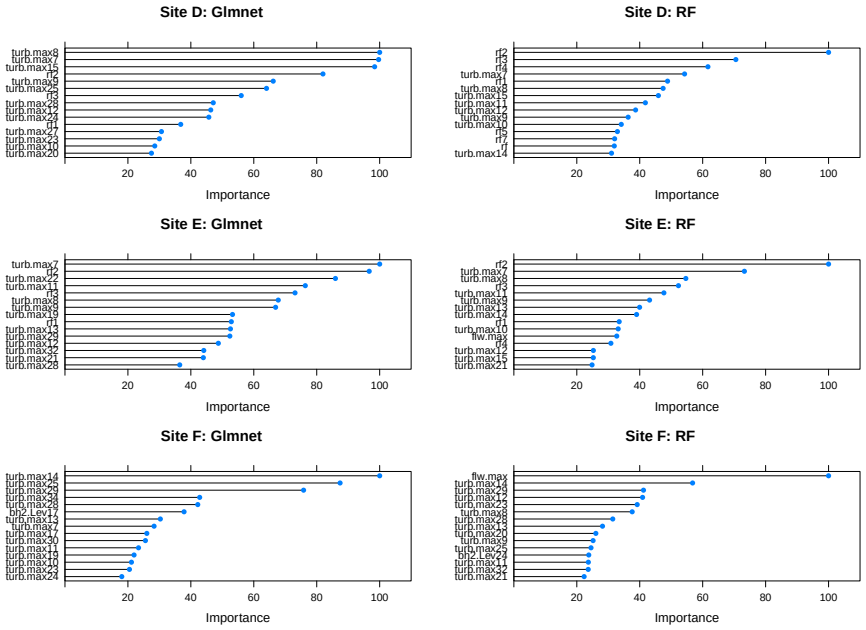

Figure 8: Variables Importance: Sites D-F

cost of a FP, that is to say, the cost of falsely predicting a turbidity (NTU) event that doesn't happen is less costly than falsely predicting a turbidity (NTU) event that does happen but is not predicted. We assign a FP cost of 1 and a FN a cost of 2. We present an example of the application of this approach in Figure 9 where the optimal cut-off point for Site-D is 0.46 with an optimised cost of 187.

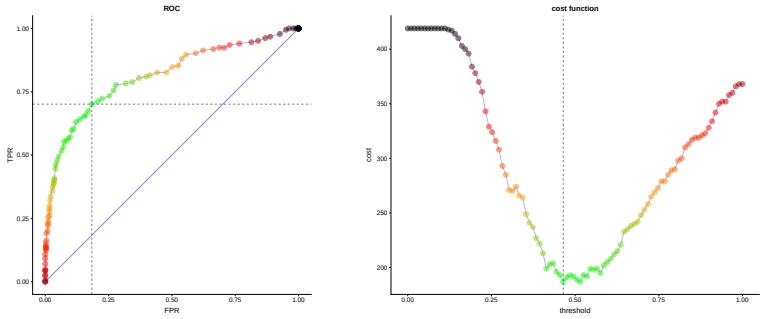

Figure 9: Cut-off Point for Site-D

We apply costs based approach to the final models at each site using the

test dataset only. The cost assumptions are held constant across all sites, and we only consider the model with the highest AUROC.

The baseline cost would reflect the total cost if no positive events were predicted over the test set while the optimised cost is the reduction as a result of the final cut-off probability. The absolute costs between the sites should not be compared however due to the differing record lengths and the rates that which >1 NTU turbidity events occur.

Table 5: Optimised Cut-Off points

	Cut-off	Baseline Cost	Optimised Cost
Site A	0.73	560	50
Site B	0.60	481	89
Site C	0.98	566	20
Site D	0.46	419	187
Site E	0.40	405	166
Site F	0.93	482	42

While the cut-off points for most of the sites seem reasonable, at Site-C and Site-F the optimal cut off point is close to a probability of 1.00, and therefore very few cases would be classed as a >1 NTU event. The reason for this is a function of both the reasonably weak discriminative ability of these two models and, the cost-based approach to defining the cut-off point. While ideally, it would be preferable to develop a stronger model, an alternative approach to determining the cut-off point is required as the current cost-based approach always favours the dominant class '*nonEvent*'. At Site-C and Site-F, because there are fewer positive cases of '*tEvent*' (<=5%), combined with the relatively weak discriminative ability of the model means predicting '*nonEvent*' is always optimal. A revised costing strategy could be applied to overcome the bias to '*nonEvent*' by increasing the cost of predicting False Positives. An alternative measure would be to use an acceptable True Positive rule where the cut-off point based on a minimum allowable percentage of True Positives that must be classified; a strategy commonly applied in credit scoring [27]. For the two sites with unreasonable cut-off points, we use a 70% True Positive rule.

By defining cut-off points for each of the sites, a practical insight is provided as to an appropriate level for establishing when an event is likely to occur, and when implemented in a live environment it indicates at which

Table 6: Revised Cut-Off Points with 70% True Positive Rate rule

	Cut-off
Site C	0.59
Site F	0.28

point operational remedial activities should be applied. Furthermore, the predicted model probabilities can be rescaled for improved interpretation on the likelihood of an event occurring, for example from 0% to 100%.

7 Conclusion

7.1 Causation

The first aim of this paper was to identify the potential variables causing daily turbidity (NTU) peaking events at groundwater sources for a water company operating in the South Coast of England. Several approaches have been used to understand the cause of turbidity (NTU) at six water sources on the South Coast of England; a static correlation analysis, a dynamic correlation analysis and an assessment of variable importance. We sought to confirm the findings of the hydrological literature applied to our research and investigate the impact of operational features. The combined output of the variable analysis has shown that turbidity (NTU) can be noisy, dynamic with association present between predictor variables. Furthermore, the response of the sites to the predictor variables can differ significantly, and in most instances, there is no one clear driver of >1 NTU turbidity (NTU) events.

The static correlation showed mostly weak to moderate relationships between turbidity (NTU) and the predictor variables with level (m, mAOD) and rainfall (mm) showing significance at most sites. At three sites flow (l/s) and pump operation were found to be significant. The significance of these variables is notable as the literature has focused on the response of turbidity to naturally occurring predictor variables in natural systems, but in a water supply context, we find operational variables influence turbidity.

The dynamic cross-correlation demonstrated there is a requirement to consider backward looking lags of level (m, mAOD) and rainfall (mm). Level (m, mAOD) was shown to linearly decay over the 35-day lag window while rainfall was most significant in the first several days followed by a linear decay.

In the dynamic correlation analysis, we also reviewed the autocorrelation of turbidity (NTU). Turbidity was shown to be significant over the 35-day lag window and most significant in the first several lags. We propose that more significance in early lags is the memory effect after a peaking event with further lags capturing the slow rising and falling turbidity levels over the year.

We also use Variable Importance outputs from the GLM and RF models to gain further insight into the multivariate nature of turbidity event prediction. The multivariate analysis mostly confirmed the results of the correlation analyses, with early lags of rainfall (mm) seeming to drive turbidity events. We also found level (m, mAOD) to be less significant in the models where lagged turbidity (NTU) was also present; this suggests that while level and turbidity both capture the slow rising and falling of turbidity over the year, turbidity lags can also capturing the memory effect after turbidity peaking. Though during the significance testing the operational flow variable was shown to be significant, we note that at only one of the sites did the variable remain 'important', though at this site it was the primary variable causing turbidity.

7.2 Prediction

The GLM and RF models were assessed using the AUCROC metric. The performance of the final models ranged from 0.81 to 0.86 in the holdout samples across the six sites leading us to conclude that machine learning models can be used to successfully predict >NTU turbidity events up to 7 days in advance, and therefore suitable as a decision support tool for water supply managers. At three of six sites, the RF model outperformed the GLM. We argue that the outperformance of the RF is due to the model's ability to capture the nonlinearity that is known to exist between rainfall (mm) and turbidity (NTU) [16]. Furthermore, at the sites where rainfall (mm) was less correlated with turbidity, the GLM outperformed or matched the performance RF, although notably achieved a lower AUCROC compared with the other sites where rainfall was a good predictor of a peaking event.

7.3 Implementation

We also considered probability cut-offs for the final selected model for each of the sites for the live implementation of the models. At four of six sites, we

applied a cost-based approach which led to a significant reduction in the baseline costs. At two sites that demonstrated lower predictive performance, in consultation with the water company, we discarded the cost-based approach as the method led to too few events being captured by the system. We instead implemented an alternative rule-based approach which ensured the system would capture a reasonable percentage of turbidity peaking events.

7.4 Further analyses & Recommendations

There are several areas for further research that may provide further insight into both the cause of turbidity (NTU) and the development of turbidity (NTU) prediction models. We have focused on site-specific factors and the development of models that can be used to predict future >1 NTU turbidity events. We modelled the sites separately to overcome the difference in how sites respond to the predictor variables. It may be of interest to explore why the sites respond differently. Research such as that by Tong and Chen [10] has examined broader influences such as land coverage and land usage in the general causation of water quality parameters. Understanding why the sites respond differently may allow the water company to take strategic action, for example, if land coverage was found to be significant then more cover crops nearby could be planted. It also may be of interest for further research to explore extending the methods for prediction used in this paper. Other binary classification models such as Artificial Neural Networks (ANNs) have also been shown to perform well with noisy and non-linear data [13], and have also demonstrated effectiveness with dynamic time series data [12].

Acknowledgements

We would like to acknowledge the anonymous company who provided the data.

This work was supported by the Economic and Social Research Council [grant number ES/P000673/1]; and The Alan Turing Institute under the EPSRC [grant number EP/N510129/1].

References

[1] Alan Charles Twort, FM Law, FW Crowley, DD Ratnayaka, et al. *Water supply.* Number Ed. 4. Edward Arnold (Publisher) Ltd., 1994.

[2] World Health Organization et al. Water quality and health-review of turbidity: information for regulators and water suppliers. 2017.

[3] Drinking Water Inspectorate. Guidance on the implementation of the water supply (water quality) regulations 2000 (as amended) in England. Version 1.1 — March 2012.

[4] Thomas Verbraken, Cristián Bravo, Richard Weber, and Bart Baesens. Development and application of consumer credit scoring models using profit-based classification measures. *European Journal of Operational Research*, 238(2):505–513, 2014.

[5] Miguel Camacho-Collados and Federico Liberatore. A decision support system for predictive police patrolling. *Decision Support Systems*, 75:25–37, 2015.

[6] Željko Deljac, Mirko Randić, and Gordan Krčelić. Early detection of network element outages based on customer trouble calls. *Decision Support Systems*, 73:57–73, 2015.

[7] Nuno Carneiro, Gonçalo Figueira, and Miguel Costa. A data mining based system for credit-card fraud detection in e-tail. *Decision Support Systems*, 95:91–101, 2017.

[8] Véronique Van Vlasselaer, Cristián Bravo, Olivier Caelen, Tina Eliassi-Rad, Leman Akoglu, Monique Snoeck, and Bart Baesens. A novel approach for automated credit card transaction fraud detection using network-based extensions. *Decision Support Systems*, 75:38–48, 2015.

[9] Mark W LeChevallier, William D Norton, and Ramon G Lee. Occurrence of giardia and cryptosporidium spp. in surface water supplies. *Applied and Environmental Microbiology*, 57(9):2610–2616, 1991.

[10] Susanna TY Tong and Wenli Chen. Modeling the relationship between land use and surface water quality. *Journal of environmental management*, 66(4):377–393, 2002.

[11] Emily K Read, Vijay P Patil, Samantha K Oliver, Amy L Hetherington, Jennifer A Brentrup, Jacob A Zwart, Kirsten M Winters, Jessica R Corman, Emily R Nodine, R Iestyn Woolway, et al. The importance of lake-specific characteristics for water quality across the continental United States. *Ecological Applications*, 25(4):943–955, 2015.

[12] Nitin Muttil and Kwok-Wing Chau. Neural network and genetic programming for modelling coastal algal blooms. *International Journal of Environment and Pollution*, 28(3-4):223–238, 2006.

[13] Ali Najah, Ahmed Elshafie, Othman A Karim, and Othman Jaffar. Prediction of johor river water quality parameters using artificial neural networks. *European Journal of Scientific Research*, 28(3):422–435, 2009.

[14] Hatim Elhatip and M Aydin Kömür. Evaluation of water quality parameters for the mamasin dam in aksaray city in the central anatolian part of turkey by means of artificial neural networks. *Environmental geology*, 53(6):1157–1164, 2008.

[15] Kevin M Hiscock. *Hydrogeology: principles and practice*. John Wiley & Sons, 2009.

[16] N Massei, JP Dupont, BJ Mahler, B Laignel, M Fournier, D Valdes, and S Ogier. Investigating transport properties and turbidity dynamics of a karst aquifer using correlation, spectral, and wavelet analyses. *Journal of hydrology*, 329(1-2):244–257, 2006.

[17] M Tanguy, H Dixon, I Prosdocimi, DG Morris, and VDJ Keller. Gridded estimates of daily and monthly areal rainfall for the united kingdom (1890–2015)[ceh-gear]. *NERC Environmental Information Data Centre*, 10, 2014.

[18] S van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, pages 1–68, 2010.

[19] Ronald L Iman and William-Jay Conover. A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics-Simulation and Computation*, 11(3):311–334, 1982.

[20] Leo Breiman. Manual on setting up, using, and understanding random forests v3. 1. *Statistics Department University of California Berkeley, CA, USA*, 1, 2002.

[21] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[22] Philipp Probst and Anne-Laure Boulesteix. To tune or not to tune the number of trees in random forest. *Journal of Machine Learning Research*, 18(181):1–18, 2018.

[23] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[24] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[25] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[26] Philip M Dixon, Jacob Weiner, Thomas Mitchell-Olds, and Robert Woodley. Bootstrapping the gini coefficient of inequality. *Ecology*, 68(5):1548–1551, 1987.

[27] Naeem Siddiqi. *Intelligent credit scoring: Building and implementing better credit risk scorecards*. John Wiley & Sons, 2017.