
Article:
Gleeson, K. orcid.org/0000-0002-3767-3001, Husband, S. orcid.org/0000-0002-2771-1166, Gaffney, J. et al. (1 more author) (2024) Algorithms to mimic human interpretation of turbidity events from drinking water distribution systems. Journal of Hydroinformatics, 26 (1). pp. 143-161. ISSN 1464-7141

https://doi.org/10.2166/hydro.2023.159

Journal of Hydroinformatics

Journal of Hydroinformatics Vol 00 No 0, 1 doi: 10.2166/hydro.2023.159

Algorithms to mimic human interpretation of turbidity events from drinking water distribution systems

Killian Gleeson [a,*], Stewart Husband [a], John Gaffney [b] and Joby Boxall [a]

[a] The Department of Civil and Structural Engineering, The University of Sheffield, Sheffield, S10 2TN, UK
[b] Siemens UK, Manchester, M20 2UR, UK
*Corresponding author. E-mail: kgleeson1@sheffield.ac.uk

KG, 0000-0002-3767-3001; SH, 0000-0002-2771-1166; JB, 0000-0002-4681-6895

ABSTRACT

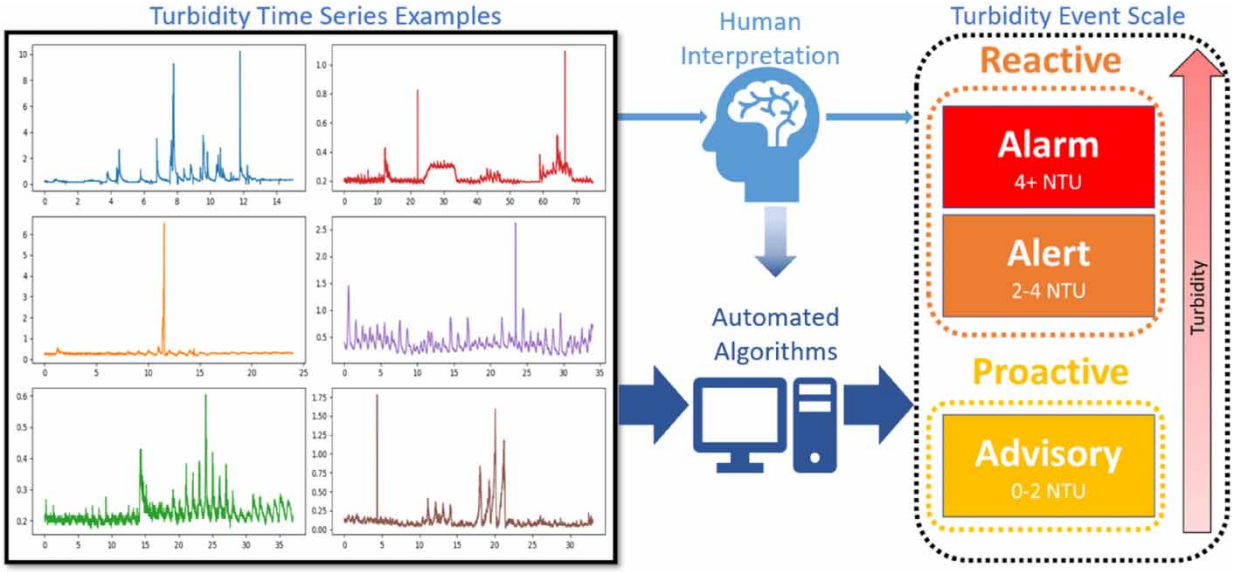
Deriving insight from the increasing volume of water quality time series data from drinking water distribution systems is complex and is usually situation- and individual-specific. This research used crowd-sourcing exercises involving groups of domain experts to identify features of interest within turbidity time series data from operational systems. The resulting labels provide insight and a novel benchmark against which algorithmic approaches to mimic the human interpretation could be evaluated. Reflection of the results of the labelling exercises resulted in the proposal of a turbidity event scale consisting of advisory <2 NTU, alert $2 <$ NTU < 4, and alarm >4 NTU levels to inform utility response. Automation, for scale up, was designed to enable event detection within these categories, with the <2NTU category being the most challenging. A time-based averaging approach, based on data at the same time of day, was found to be most effective for identifying these advisory events. The automation of event detection and categorisation presented here provides the opportunity to gain actionable insight to safeguard drinking water quality from ageing infrastructure.

Key words: discolouration, drinking water distribution systems, drinking water quality, event detection, turbidity, water quality time series

HIGHLIGHTS

- Crowd-sourcing captured domain expert interpretations of events within turbidity time series.
- A turbidity event scale is proposed to inform reactive and proactive interventions.
- Automated algorithms were developed to recreate and scale up domain expertise.

GRAPHICAL ABSTRACT



1. INTRODUCTION

Continuous water quality monitoring within drinking water distribution systems (DWDSs) enables network events to be captured and understood at a level of spatial and temporal detail that regulatory periodic discrete sampling cannot achieve. Causes of post-treatment DWDS water quality events range from hydraulic-induced mobilisation of pipe wall material (Husband *et al.* 2008), infrastructure failures allowing contaminant ingress (LeChevallier *et al.* 2003), to bulk water transformation such as excessive chlorine decay leaving no residual protection against contamination (Speight *et al.* 2019). A primary source of water quality-related customer contacts is discoloured water (DWI 2022) with turbidity sensors, using optics to measure the light scattering of water, considered a proxy measurement (Boxall & Saul 2005). Turbidity has also been shown to provide network specific correlation with iron and manganese (Cook *et al.* 2016), so also providing some insight into these parameters. Time series turbidity data taken from within DWDS are, therefore, of particular interest to operators who wish to understand and hence reduce the likelihood of discolouration events and customer contacts. Utilities are increasingly deploying turbidity sensors within DWDS, with the resulting datasets currently relying on manual interpretation that is reactive, subjective, situation specific, and time consuming.

Data visualisation and interpretation is a powerful human skill due to the brains ability to subconsciously process visual information in as little as 13 ms (Potter *et al.* 2014), significantly faster than text or numbers. An expert analyst can quickly identify and label periods of data of interest from interpreting graphical representations, yet the subjective nature limits the ability to cross compare. The sheer volume of data now being collected, along with the 24/7 nature of DWDS, makes reliance on such subjective manual assessment unviable, particularly as human brains can only accurately and quickly comprehend up to four variables at once (Halford *et al.* 2005). There is, therefore, a need to better understand the human process and to develop computing algorithms that can automate aspects of the interpretation and analysis of turbidity data to rapidly provide actionable information for operational decisions. The IWA's recent series of white papers on digital transformation (IWA 2022) stresses the need to move to more proactive infrastructure management and analysis of DWDS. Higher frequency turbidity time series data has the potential to enable this and improve our understanding of discolouration processes that will aid sustainable and safe delivery of high-quality drinking water.

1.1. Background

Detection of interesting, undesired, or anomalous events in datasets is a widely studied and varied topic. The most common form is in detecting rare or unusual data points, often termed outliers or anomalies, by seeking deviations from assumed or modelled normality (Aggarwal 2016). Successful examples are found in network hacking, credit card fraud, and medical diagnostics (Aggarwal 2016). A review of anomaly detection techniques by Chandola *et al.* (2009) identified the nature of the

available data and the type of event detection required as two key factors that dictate what methods are suitable. The availability of labelled data, an agreed designation where one or more labels identify properties, characteristics or classifications, opens an array of supervised machine learning approaches. These include support vector machines (SVMs) and artificial neural networks (ANNs) that can be more effective than unsupervised techniques as they use knowledge of known previous examples (Aggarwal 2016). Another important factor is the number of variables in a dataset, with significant research being done to detect anomalies in applications where high-dimensional datasets are the norm such as financial records and online interactions (Thudumu *et al.* 2020). When the data are in a time series, the temporal context of each dataset requires consideration, and detection methods rely either on a statistical or forecasted expected value, from which the real values are compared and some sort of outlier score is determined (Gupta *et al.* 2014; Blázquez-García *et al.* 2020). The field of time series forecasting is of direct importance here, with ARIMA (Autoregressive Integrated Moving Average) and exponential smoothing two of the most popular approaches (Hyndman & Athanasopoulos 2021). Important considerations are the quantity of data used to determine a forecast and the forecast horizon. ARIMA models utilise the autocorrelations in a time series in order to make forecasts (Hillmer & Tiao 1982), while exponential smoothing gives greater importance to more recent data and has been adapted to account for trend and seasonality (Blázquez-García *et al.* 2020). Seasonality in time series data can refer to patterns occurring on a repeated periodic basis, such as yearly, monthly, weekly, or daily. Seasonality is relevant to DWDS time series due to the strong links to seasonal weather and human behaviour patterns. SARIMA (seasonal ARIMA) is a modification of ARIMA that is capable of accounting for seasonality while VAR (vector autoregression) and ARIMAX (X representing exogenous variables) models are adaptations that can consider additional variables. The ETS (error, trend, seasonal) framework describes nine exponential smoothing variations, based on how the error, trend and seasonal components are calculated and combined (Hyndman & Athanasopoulos 2021). Recent advances in time series forecasting include the use of neural networks, with LSTM (long short-term memory) particularly popular for supervised multi-variate time series forecasting (Hochreiter & Schmidhuber 1997), and Prophet, which can be applied automatically and considers holiday effects (Taylor & Letham 2018).

Research on detecting events in DWDS has been dominated by leakage detection methods, most commonly looking for unusual patterns in acoustic or pressure sensor data (El-Zahab & Zayed 2019). Detection of water quality events within DWDS has not attracted as much attention, but research has been done to detect intentional contamination of DWDS by the US Environmental Protection Agency (EPA), who produced an open-source event detection software package called CANARY, which consists of various different statistical algorithms to detect outlier values based on rolling window statistics, from which an event probability is calculated for each window using a Binomial Event Discriminator (BED) (McKenna *et al.* 2007). The use of rolling windows is a common way to account for the temporal context in time series. CANARY has been applied to DWDS data in the UK, where it has shown promise in detecting multi-parameter events (Mounce *et al.* 2012). The difficulty of linking detected events to confirmed real-world actions is highlighted by this research, where only 28% of detected events could be linked either to customer contacts or hydraulic disturbances. Labelled data in DWDS are uncommon and the process of linking data to information from network operations or customer interactions is time-consuming. Additionally, deciding what constitutes a water quality event is not clear cut, meaning any labels cannot be considered ground truth. Crowd-sourced labels are commonly used in machine learning and research has been done to understand how to deal with inevitable human error (Ustalov *et al.* 2021) with strategies that include multiple labellers per example. However, the labels used are generally definitive, such as whether a picture contains a cat or a dog, and little research has been done to understand how labels can be combined in cases where the question posed is highly subjective.

When developing methods for analysing events in turbidity time series, it is important to first understand the nature of turbidity data and the desired events to detect and study. This is not a trivial challenge. Depending on turbidity *event* definition, these may occur frequently or as unique incidents and are linked to network and sensor installation location. In the UK, legislation dictates that the water at customers taps should not exceed 4 NTU (nephelometric turbidity unit), nor 1 NTU exiting treatment works (DWI 2018). Therefore, network turbidity sensors recording values more than 1 NTU are evidence of in-transit deterioration, and this may represent actionable information. In reality, turbidity levels leaving treatment works are generally much lower than 1 NTU, with less than 0.01% of regulatory turbidity samples exiting treatment works exceeding 1 NTU in 2021 (DWI 2022). Therefore, even turbidity events occurring below 1 NTU may relate to variations in discolouration risk and also be worthy of identification and study. Yet analysis of DWDS turbidity time series data has tended to focus on reacting to larger events, meaning the information at lower turbidity levels has remained unused. Computing and modern

analysis techniques, however, offer the potential to rapidly analyse lower-level turbidity data but require specific instructions that are currently not well understood.

The aim of this research was to explore and to improve understanding of what constitutes an event worthy of further consideration in turbidity time series data and then to develop and assess automated computing algorithms that can rapidly review and identify such events, mimicking human judgements and intuitive extrapolation to inform both reactive and proactive utility responses.

2. METHOD

2.1. Methodology

The difficulty and subjectivity of linking turbidity data with real-world evidence of water quality deterioration led to a crowdsourcing approach being taken that involved a time series labelling exercise, with domain experts being tasked to label what they considered to be events of interest within turbidity time series examples. This approach takes advantage of human brain power, which computer algorithms can only approximate when given specific instructions. To overcome the problem of bias and subjectivity, the same time series examples were shown to different groups at different meetings, with each of the resulting Boolean labelled time series combined and averaged. This averaging of results for each turbidity datapoint returned an associated 'label average' score, between 0 and 1. This value could then be used as a benchmark to evaluate the suitability of algorithmic approaches such as flat-line detection and the calculation of event score time series of similar form to the averaged labels. The labelled data would also inform whether a single approach can handle different event types or whether a combination of approaches is more suitable.

2.2. Event labelling exercise

An interactive labelling exercise was compiled using the open-source browser-based time series labelling tool Trainset (Geocene 2020), to enable users to label six turbidity time series examples. The examples were selected after reviewing approximately 100 turbidity sensors across four different UK DWDS, which combine to a rough total of 150 years' worth of turbidity time series data. As this research is interested in both reactive and proactive aspects of turbidity events, six turbidity examples were selected to represent a range of different event types and magnitudes that were observed in the wider datasets. Each turbidity example had been quality assured according to the procedures set out in Gleeson *et al.* (2023). Different durations were used to represent realistic but manageable sensor deployment time frames, ranging from 2 to 10 weeks. Care was taken to ensure the participants were not aware of the reasoning behind the examples. Six time series of 16–75 days in duration was considered to be a safe limit to ensure the human experts maintained a high level of focus and attention to detail. Additionally, limiting this exercise to just examples meant it took roughly 10 min to compete, which was considered a realistic expectation of participants. A screenshot of the event labelling software with example number 4 is shown in Figure 1 where the pink highlighted data are an example of what user-labelled data looks like. Example 4 is unique out of the six time series turbidity datasets in that it was artificially constructed (by splicing and combining data) to represent some different theoretical types of turbidity event; (1) a hydraulic-induced material mobilisation event, (2) a single point event, (3) a baseline change event, and (4) an increase in diurnal turbidity event. These event types are marked in Figure 1, where event (5) is a combination of the first four. The other examples are all unedited turbidity time series from different UK DWDS. To ensure consistency between data from multiple sources and reporting intervals, all examples were resampled to a 15-min sampling interval.

The labelling exercise was run across multiple sessions with anonymity retained and participant consent required to confirm they understood what they were taking part in before they could proceed to the labelling interface. Upon completing the labelling exercise, users were directed to an upload page on a dedicated webpage, which had an upload button that anonymously uploaded the labelled data to a dedicated server folder. For the labelling sessions of the exercise, users were simply given the instruction to 'label events' with the following provided as an event definition:

'*An event is described as a noteworthy period of data to be flagged for further consideration*'

2.3. Event score calculation'

Each algorithmic approach involved first making a forecast (other than flat-line), which is then subtracted from the turbidity data to obtain a residual time series. The next step was to transform the residual into an event score time series, which could

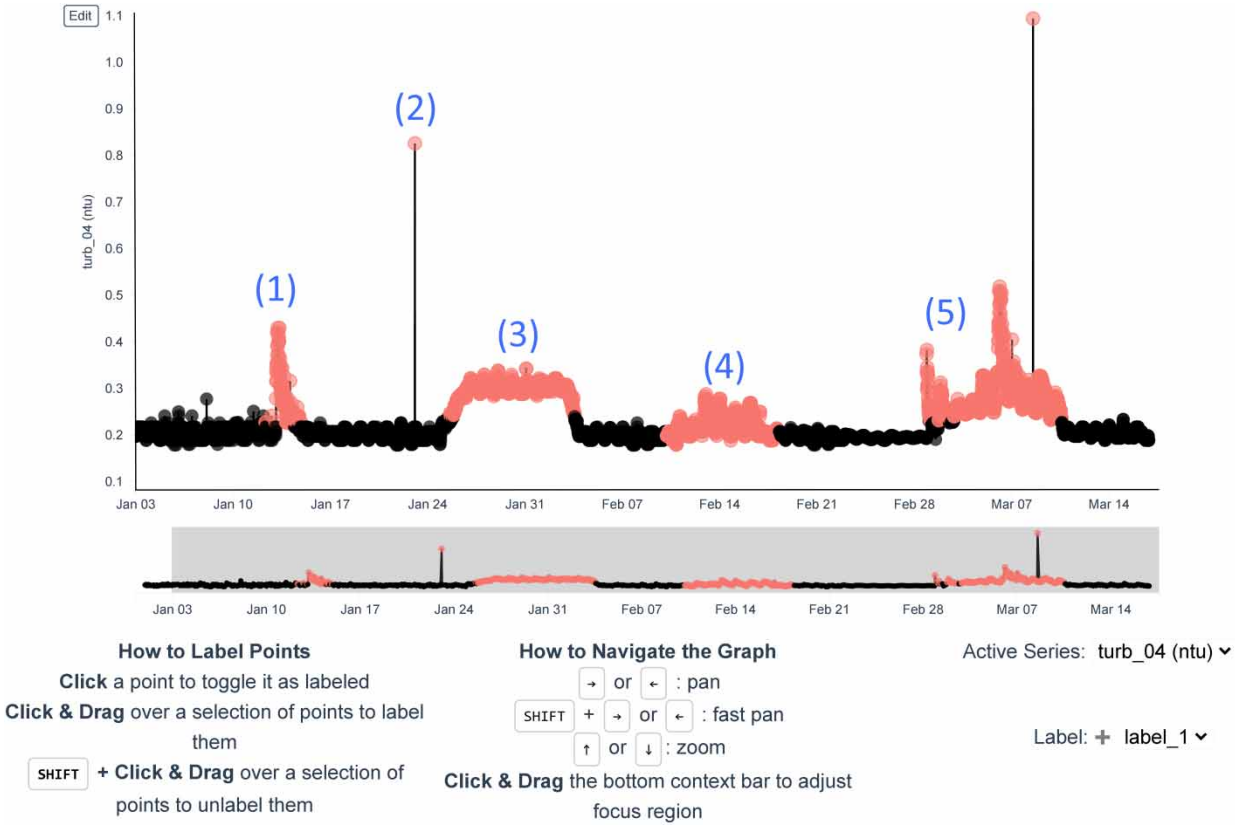


Figure 1 | Screenshot of turbidity Example 4 in event labelling tool, with the theoretical event types highlighted.

be compared to the averaged-out labels. This transformation was achieved using a sigmoid function. To compare different time series forecasting methods, the sigmoid function was optimised to find the lowest error against the labels, for each residual calculated. Each approach involved adjustable parameters, which were investigated in a sensitivity analysis with the goal of determining the combination that most closely captured the information gained from the labelled datasets.

2.3.1. Forecasting methods

The different methods used to make a forecast, from which a residual was calculated, are listed in Table 1. All approaches were employed for sliding windows of 24, 48, and 72 h, as well as expanding windows for forecast horizons of single point and between 2 and 72 h ahead. CANARY was an exception as it only produced next step forecasts and does not include

Table 1 | Forecasting methods and associated adjusted parameters

Approach	Variants (number of adjusted parameters)	Adjusted Parameters
Averaging	Mean (0), median (0), quantiles (1)	quantile value
Time-based Averaging	Mean (1), median (1), quantiles (2)	Window size (h), averaging method, quantile value
ARIMA-based	ARIMA (3)	p, d, q (ARIMA)
	SARIMA (6)	p, d, q, P, D, Q (SARIMA)
Exponential Smoothing	ETS (4)	Error, trend, seasonal, damped trend (ETS)
	EWM (1)	Alpha (EWM)
Prophet	Auto and with settings (3)	Growth method, growth cap (if method is logistic), seasonality mode
CANARY	CANARY LPCF (1)	Outlier threshold

expanding windows. However, the remaining approaches all share window and forecast horizon parameters, with all method-specific adjustable parameters listed in Table 1. Averaging methods were based on using data within the specified window, with different quantile levels examined, as well as mean values. The time-based average method represented a deviation from the typical sliding window approach. Instead of using a window directly preceding each datapoint, this method looks at previous data at the same time of day, accounting for the diurnal patterns often seen in DWDS data that is heavily linked to human behaviour. New adjustable parameters were introduced here, the size of window to include each day (e.g. for a datapoint at 8:30 am, a 2-h window would mean any data between 7:30 am and 9:30 am would be included) and the averaging method used. The averaging and time-based averaging approaches were developed in Python using the Pandas (McKinney 2010) library.

ARIMA has three input parameters: the lag order (p), the degree of differencing (d), and the order of moving average (q) (Hillmer & Tiao 1982). These parameters make up the order, often shown in the form: (p, d, q). SARIMA also has seasonal ordering parameters P, D, Q, and m, where m is the seasonal period. Wherever the seasonal period was a possible option, 96 represent the diurnal patterns that turbidity time series can exhibit as this is how many samples were in a day (at 15-min sampling rates). Exponential smoothing methods were explored using the ETS framework that looked at the impacts of different error, trend, and seasonal component calculations. Each component can be either additive or multiplicative. An exponential weighted mean (EWM) approach was also included, which requires the decay to be specified, either in terms of centre of mass, span, half-life, or as a smoothing factor. Other methods investigated were Prophet and CANARY. Prophet was run using both its automatic functionality, and for different growth methods and seasonality modes. The ARIMA, SARIMA, ETS, and Prophet approaches were developed using the machine learning for time series interface library sktime (Loning *et al.* 2019). The CANARY software was run and included in this analysis, using the linear prediction correction filter (LPCF) method and BED. The alternative multi-variate nearest neighbour (MVNN) method is not applicable to this univariate problem. LPCF uses the MATLAB filter and lpc functions to estimate the next value based on weighted filter applied to a window of normalised data proceeding each datapoint (Murray & Haxton 2010). The user needs to specify window size and the threshold, in standard deviations, above which is considered an outlier. Window sizes between 1 and 72 h were included in the sensitivity analysis while standard deviation thresholds were looked at between 0.5 and 1.5. Parameters event timeout, the number of timesteps after an event is found before alarm is silenced automatically, and event window save, a parameter related to plotting identified events, were not adjusted as this research was more interested in the residual calculated. Table 1 lists the adjusted parameters for each forecasting approach, aside from the window and forecast horizons.

2.3.2. Event score and comparison to labels

After residuals were calculated for each of the forecast methods, the sigmoid function was used to transform the residuals into event scores. The sigmoid function maps inputs to outputs between 0 and 1 using the following Equation (1).

$$y = \frac{1}{(1 + e^{-c * (x-b)})} \tag{1}$$

where x = input data, b = sigmoid centre point, c = sigmoidal width. The two sigmoid parameters were optimised to minimise the error against the labelled data using SciPy's optimisation function (Virtanen *et al.* 2020). The optimisation was done with all examples and using just 3–6, so without Examples 1 and 2. Examples 1 and 2 contain significant large-scale events exceeding 4 NTU and were excluded from one optimisation to investigate what approaches work best for lower-level turbidity events, in this case with all data below 2 NTU. RMSE (root mean squared error) was used to evaluate each approach. The root mean squared error (RMSE) is a commonly used forecasting metric, and is represented by Equation (2):

$$\text{RMSE} = \sqrt{\frac{\sum_{j=1}^{n} (x_j - y_j)^2}{n}} \tag{2}$$

As all event score values are between 0 and 1, the largest possible error is when the labels are 1 and the event detection system is 0 (or vice versa). For this research, it is desirable to punish these outcomes. Although there are different error

metrics available such as mean absolute error (MAE) or mean squared error (MSE), RMSE is known for being sensitive to outliers (Chai & Draxler 2014) and as such it is considered ideally suited for this task. To include more forecasting methods, windows, and horizons, the first 3 days of each example was omitted when calculating the RMSE. Methods such as ETS require two full cycles of data to account for seasonality. Other approaches such as the time-based averaging require at least 1 day of data, while some methods worked best for forecast horizons of 24–48 h. This also handles the 'cold start' problem many forecasting methods have, where it is very difficult to make predictions without any prior data.

For the CANARY LPCF, the BED approach was used in addition to the sigmoidal approach already outlined. Since BED requires a Boolean input, this could not be used for other residuals without adding an additional outlier threshold step, which risks losing complexity and adds an unnecessary additional input. BED uses probability theory to estimate event probability for each datapoint, based on the number of outliers present within a specified window. BED takes two input parameters, window size and outlier probability. Outlier probability is a probability threshold above which events are counted. Since this research is only interested in the probability score, the probability threshold is not needed. The CANARY manual recommends using BED windows between 4 and 18 timesteps (between 1 and 4.5 h for 15-min data) so this was the range examined in the sensitivity analysis.

3. RESULTS

3.1. Labelled results

The turbidity time series labelling exercise was run four times during different academic and industry events, with a total of 48 participants returning complete labelled data. Session 1 took place during an online meeting by a university research group who focus on DWDS. Session 2 took place during a water utility-academia event, focusing on discolouration in DWDS, with 12 UK utilities represented. Session 3 was run independently by a water utility's network modelling team. The final session was run during a separate water utility-supply chain-academic (industry dominated) event, which focuses specifically on water quality within DWDS, with 16 different water utilities and at least eight supply chain companies present. Aside from session 3, these exercises were run during academic and industry meetings that focussed on discolouration and water quality issues in distribution systems, meaning the participants were not selected specifically for the purpose of participating in this exercise. The labelling exercise sessions are summarised in Table 2. Figure 2 shows boxplots of the total percentage labelled data points from each session, illustrating the variety in responses within and across sessions. Session 3 stands out as having the lowest amount of labelled data. This session was run externally, without the authors of this research in attendance. A stricter definition of what constitutes an event was used, with attendees focussing on significant events, potentially of regulatory concern. This highlights the challenges in defining what constitutes an event of interest, or with respect to this research, what information is required to inform what decisions from the data and the impact of differing instructions and perspectives.

Figure 3 plots each turbidity time series example along with the labelling results, averaged out for each datapoint. The value of each datapoint indicates the fraction of participants that deemed it to be noteworthy. These event score time series provide a useful way to interpret the labelling results and a benchmark to evaluate algorithmic approaches against. One of the challenges of human analysis is inherent bias towards higher turbidity events. In these time series, Examples 3–6 had little data above 2 NTU, meaning that the lower-level events were more easily visualised than in Examples 1 and 2. Figure 4 is a scatter plot showing the average absolute turbidity values for each averaged-out label, divided into those from Examples 1 and 2 with significant higher turbidity events, and those from Examples 3 to 6. This illustrates the impact of bias on the presence of higher turbidity events and how the human participants then interpreted the lower-level turbidity data present in the same dataset. In order to examine automated analysis methods that would work well at analysis of lower-level turbidity events,

Table 2 | Event labelling sessions

Session	Event	Valid Labelled Datasets
1	University Research Group	8
2	Water Utility-Academia Event	9
3	UK Water Utility	12
4	Water Utility-Supply Chain-Academia Event	19

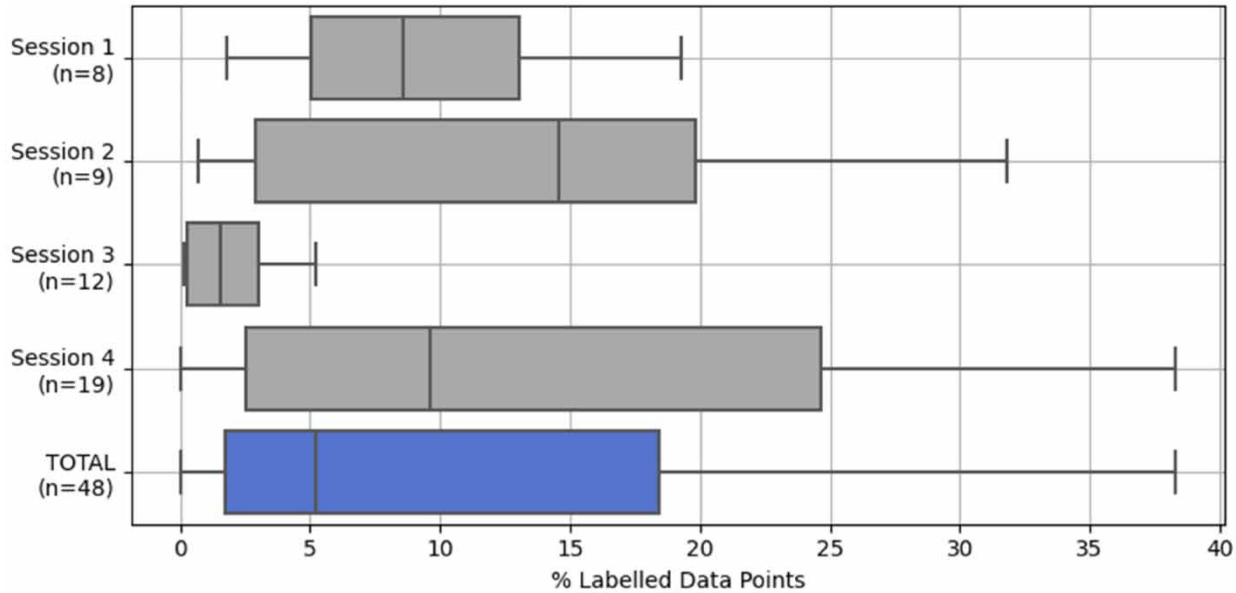


Figure 2 | Boxplot of total percentage labelled data across all six example datasets per session.

the data from Examples 3 to 6 were treated separately. To distinguish the low-level events, a threshold of 2 NTU was identified and analysis of events below this are termed 'advisory'. At the same time events exceeding the regulatory value at customer taps of 4 NTU, and, therefore, typically requiring immediate attention, are considered 'alarm'. Between these levels, events are considered 'alert', representing significant deterioration compared to the maximum permitted turbidity of 1 NTU leaving a treatment works. This turbidity event scale naming convention and the boundaries are summarised in Table 3.

3.2. Event analysis results

Using the event scaling outlined in Table 3, this research explored whether a single algorithm could deliver all three levels, or whether combinations were required. Such as the use of simple flat-line approaches to identify and separate alert and alarm events for reactive response, and tune more sensitive algorithms with the ability to accurately identify lower-level events that could inform proactive measures. Methods that output event score time series between 0 and 1, like that of the averaged-out labels, were, therefore, developed and tuned for all six examples (all three event categories), and to advisory events only, Examples 3–6.

3.2.1. Flat-lines

Flat-line algorithms to detect alert and alarm events are considered separately here as they do not require the process of prediction and residual calculation. These flat-line thresholds are shown applied to Example 1 in Figure 5. This simple approach was effective at identifying both alarm and alert events. Attempts to use a flat-line approach for identification of advisory events were very poor. Unlike alert and alarm events, applying a flat-line at lower turbidity levels would make detection strongly dependent on background turbidity levels. For example, applying a 0.5 NTU flat-line would result in one advisory event in Example 3 and 39 advisory events in Example 5, due to Example 5 having higher background turbidity. Therefore, analysis of advisory events required consideration of background turbidity which the calculation of residual values achieved, prior to conversion to event scores.

3.2.2. Calculating event scores

Figure 6 illustrates the application of the ARIMA algorithm as an example of the approach adopted. The top plot shows an expanding window ARIMA forecast for Example 1, while the bottom plot shows an expanding time-based averaging forecast for Example 3. Figure 7 illustrates how the obtained residuals were then transformed into an event score time series. A sigmoid function was used to understand and optimise the relationship between the residual values and the labelled data. The

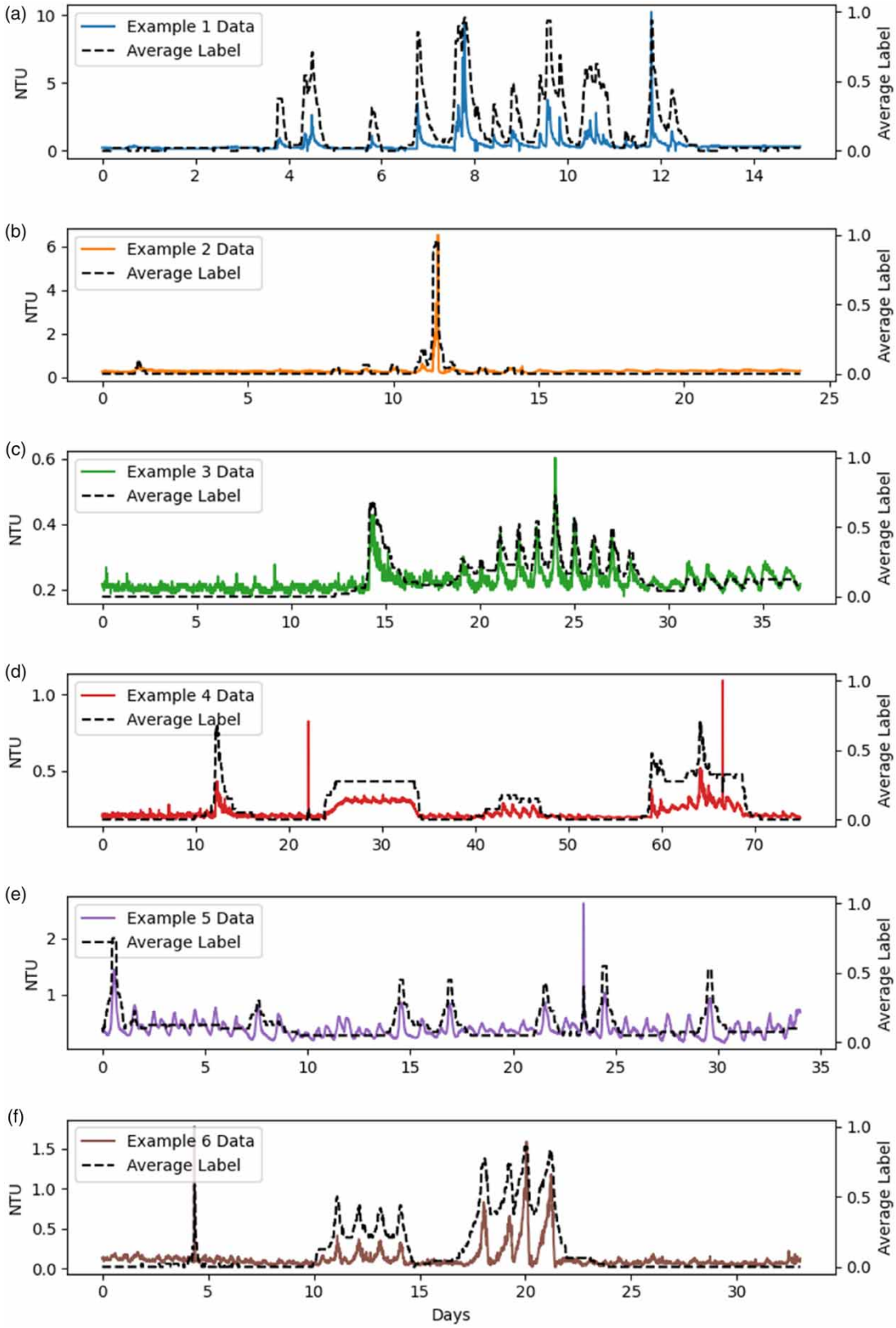


Figure 3 | Six turbidity time series examples, in plots (a)–(f), respectively, along with corresponding average label results (the *x*-axis and NTU *y*-axis ranges are scaled for each dataset).

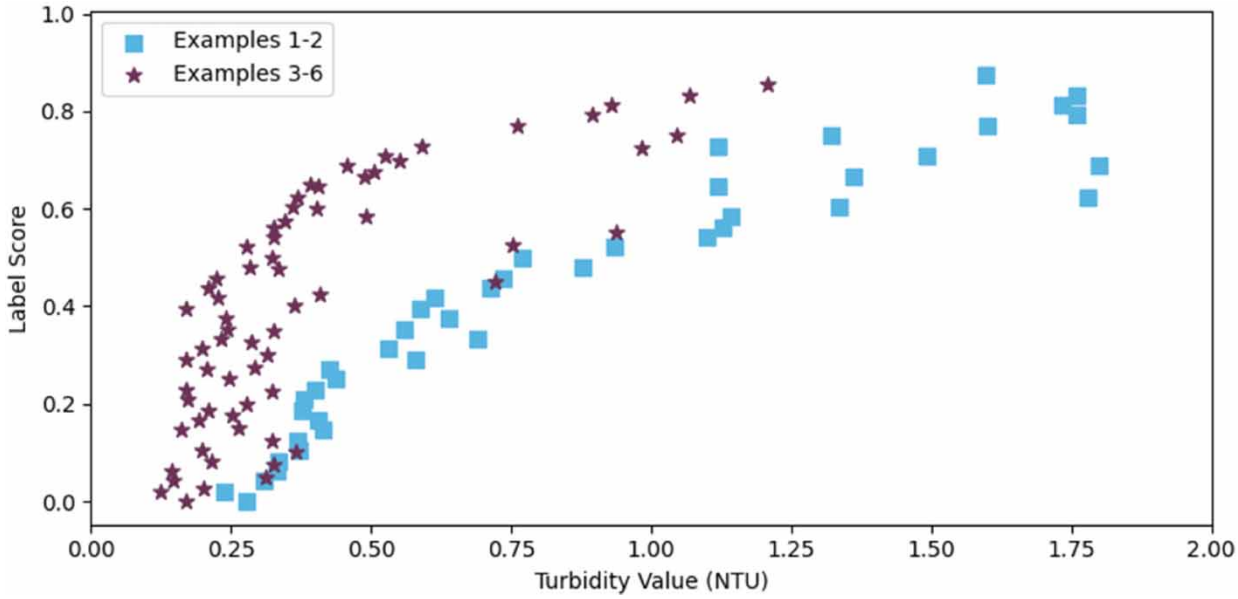


Figure 4 | Mean absolute turbidity value for each average label value; for Examples 1 and 2 (blue squares) and Examples 3–6 (purple stars).

Table 3 | Turbidity event scale

Event Type	Turbidity Limits
Advisory	< 2 NTU
Alert	$2 <$ NTU < 4
Alarm	>4 NTU

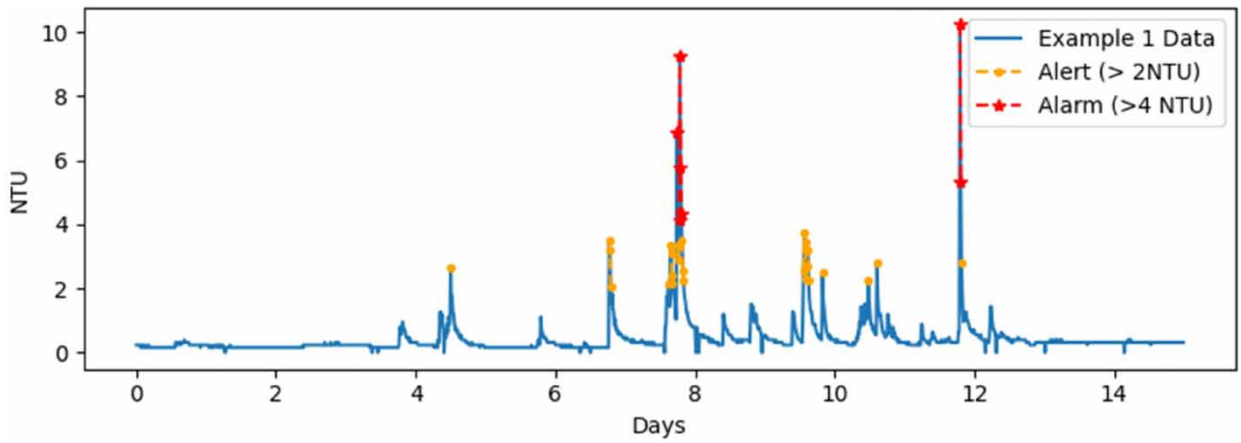


Figure 5 | Flat-line thresholds differentiating Alert (2–4 NTU) and Alarm (>4 NTU) events in Example 1.

RHS plots show the outputs of the sigmoid function compared to the labels when averaged, demonstrating how the obtained human interpretation can be mimicked using this sigmoidal function.

3.2.2.1. Forecasting methods. Each approach (other than flat-lines) included in this research had its calculated residual optimised to minimise errors against the averaged-out labels, hence the forecasting methods can be compared to each

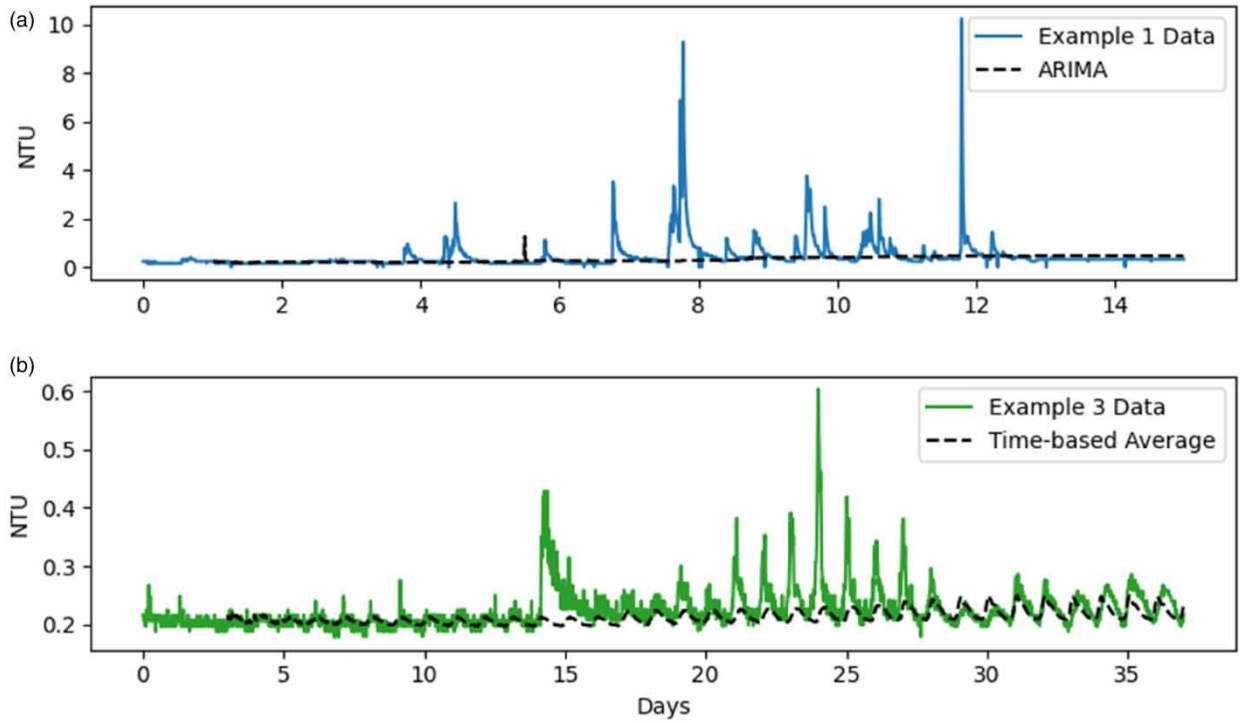


Figure 6 | Example 1 with expanding ARIMA forecast values (a) and Example 3 with expanding time-based average forecast values (b).

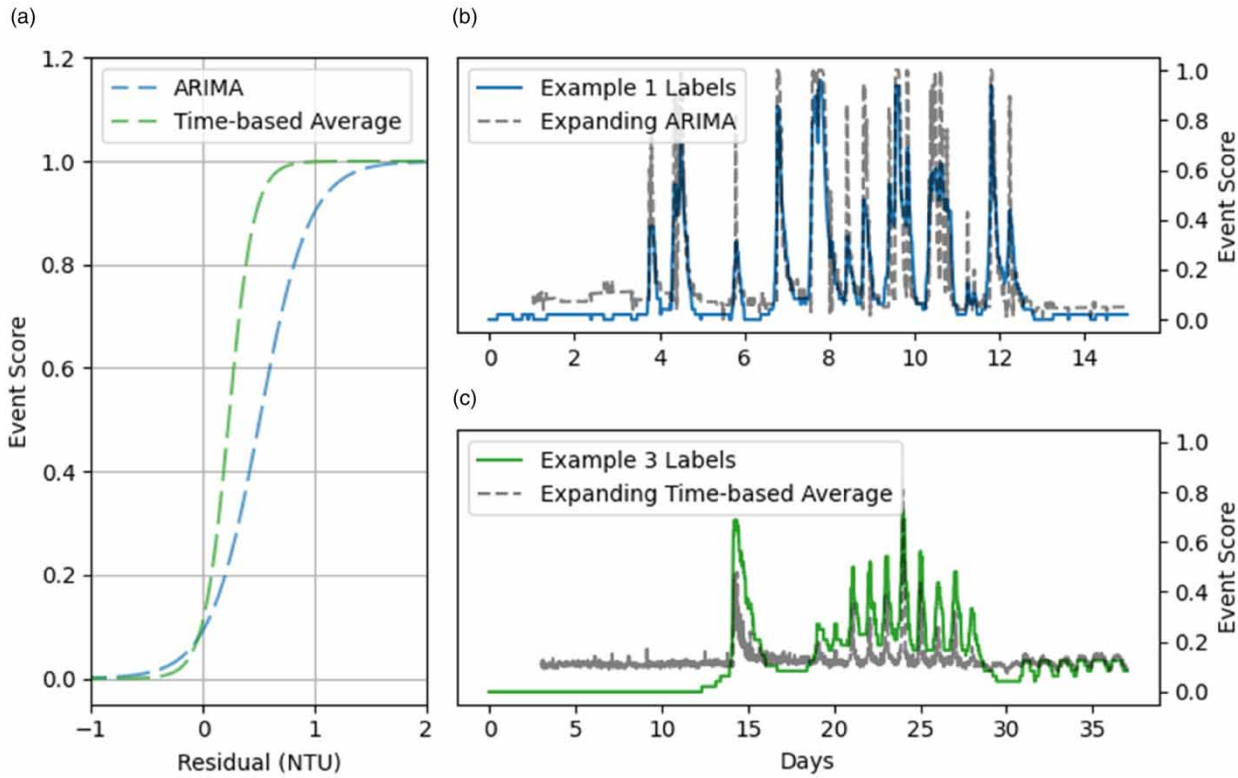


Figure 7 | Optimised sigmoid function for ARIMA tuned on all events and time-based (tb) average method tuned on advisory events (a), with corresponding outputs compared to Example 1 and 3 labels, (b) and (c), respectively.

other using the optimised error values. The solution with the lowest RMSE for both 'all events' and for 'advisory only events', for averaging and time-based averaging approaches, is shown in Table 4. An expanding window had better results than any sliding window approach, while forecast horizons of 12 and 24 h were best for averaging, for 'all events' and 'advisory only events', respectively. For time-based averaging, forecast horizons can only be in multiples of days, with 1 and 3 days found to work well for 'all events' and 'advisory only events', respectively. Daily window sizes of 6 and 3 h were found to work better than exact time-based values, showing that including data before and after each timestamp was useful.

The optimal parameters for ARIMA and SARIMA are displayed in Table 5. An expanding-type window in combination with 24-h forecast horizon worked the best. For ARIMA the most useful order was (1,0,0) which represents a first-order auto-regressive model without any differencing or moving averaging. SARIMA was found to the most time-consuming approach, meaning not all possible combinations were completed. Of those that were, the best approach had an order of (1,0,0) × (0,0,0) meaning no seasonality terms were employed, suggesting the (1,0,0) order ARIMA was adequate. The optimal parameters for the exponential smoothing approaches are listed in Table 6. Using a half-life of 14 days worked well for EWM, while the

Table 4 | Optimal parameters found for averaging and time-based averaging

	Averaging		Time-based averaging	
	Tuned on all events	Tuned on advisory only events	Tuned on all events	Tuned on advisory only events
Window	expanding	expanding	expanding	expanding
Forecast Horizon	12-h	24-h	1 day	3 day
Other Parameters	averaging method = mean	quantile = 0.8	daily window size = 6-h, averaging method = median	daily window size = 3-h, averaging method = mean
Sigmoid Parameters	b = 0.51, c = 4.47	b = 0.24, c = 7.72	b = 0.59, c = 4.04	b = 0.23, c = 8.98

Table 5 | Optimal parameters found for ARIMA and SARIMA

	ARIMA		SARIMA	
	Tuned on all events	Tuned on advisory only events	Tuned on all events	Tuned on advisory only events
Window	expanding	expanding	expanding	expanding
Forecast Horizon	24-h	24-h	24-h	24-h
Other Parameters	order = (1,0,0)	order = (1,0,0)	order = (1,0,0) × (0,0,0)	order = (1,0,0) × (0,0,0)
Sigmoid Parameters	b = 0.51, c = 4.50	b = 0.32, c = 6.55	b = 0.51, c = 4.49	b = 0.31, c = 6.57

Table 6 | Optimal parameters found for EWM and ETS

	EWM		ETS	
	Tuned on all events	Tuned on advisory only events	Tuned on all events	Tuned on advisory only events
Window	expanding	expanding	24-h	24-h
Forecast Horizon	24-h	24-h	12-h	48-h
Other Parameters	half-life = 14 days	half-life = 14 days	error = additive, trend = None, damped = False, seasonal = None	error = multiplicative, trend = None, damped = False, seasonal = None
Sigmoid Parameters	b = 0.52, c = 4.42	b = 0.34, c = 5.98	b = 0.63, c = 3.89	b = 0.41, c = 5.25

optimal solutions found using ETS both involved no trend or seasonal components, meaning simple exponential smoothing was found to work best. The optimal parameters for Prophet are shown in Table 7. For Prophet the same residual was found to be the best solution for all events and for only advisory events. Even the sigmoid function parameters are like each other. For LPCF, parameters are shown in Table 8 shows solutions using both the optimised sigmoid approach and CANARY's BED function. CANARY does not allow expanding windows, nor does it include forecast horizons other than single point. The optimised sigmoid parameters have noticeable different parameters to other methods due to the larger magnitudes seen in CANARY LPCF residual time series. When using BED, the best solution for all event levels and advisory event levels were identical.

3.2.2.2. Comparison to Labels. Figure 8 plots the lowest RMSE found for each of the forecasting methods investigated. The methods were tuned for lowest RMSE and assessed both for all event levels (using all six examples) and for advisory events only (using Examples 3–6). The residuals from CANARY's LPCF algorithm were passed through the same sigmoidal function, as well as using CANARY's in-built BED function (though this was not tuned in the same way as the sigmoid approach). An expanding ARIMA approach with a (1,0,0) order and 24-h forecast horizon resulted in the lowest RMSE of 0.1137 across all examples. The second-best approach across all examples was the simple averaging (RMSE of 0.1140) though there were ten different ARIMA combinations that resulted in an overall RMSE of 0.1140 or less, including a (1,0,0) order at a 48-h horizon, (0,0,0) order, which represents white noise, with shorter 6 and 12-h horizons, and orders (0,0,q) with q = 1,2,3 for forecast horizons of 6 and 12 h. This suggests ARIMA has strong applicability to this research and that including autoregressive or moving average terms can be useful for calculating the residual and subsequent event score time series. The time-based averaging approach worked the best for advisory events with a RMSE of 0.1095, but this approach did not work effectively at generalising across all event types and performed worse than the averaging approach for all examples. Figure 9 shows this approach applied to data from Example 2, before and after a larger alarm type event, and compared to the averaged-out labels in Example 2, with the turbidity data clipped to exclude the larger 6 NTU event. The averaged-out label, shown in the context of this NTU *y*-axis scale, did not have any score above 0.2 outside of this larger alarm event. By contrast the time-based average approach, tuned on advisory data, is not biased by the presence of the alarm event, and returned event scores of increasing magnitude in the days leading up to the alarm event, with a value of 0.7

Table 7 | Optimal parameters found for prophet

	Prophet	
	Tuned on all events	**Tuned on advisory only events**
Window	expanding	expanding
Forecast Horizon	24-h	24-h
Other Parameters	growth = logistic, growth cap = 0.5, daily seasonality = False, seasonality mode = additive	growth = logistic, growth cap = 0.5, daily seasonality = False, seasonality mode = additive
Sigmoid Parameters	b = 0.55, c = 4.17	b = 0.48, c = 4.42

Table 8 | Optimal parameters for LPCF using sigmoid function and BED

	LPCF		LPCF + BED	
	Tuned on all events	**Tuned on advisory only events**	**Best for all events**	**Best for advisory only events**
Window	48-h	9-h	72-h	72-h
Forecast Horizon	N/A	N/A	N/A	N/A
Other Parameters	outlier threshold = 0.5	outlier threshold = 1.0	outlier threshold = 1.5	outlier threshold = 1.5
Sigmoid (or BED) Parameters	b = 6.41, c = 0.36	b = 11.30, c = 0.19	BED window = 4	BED window = 4

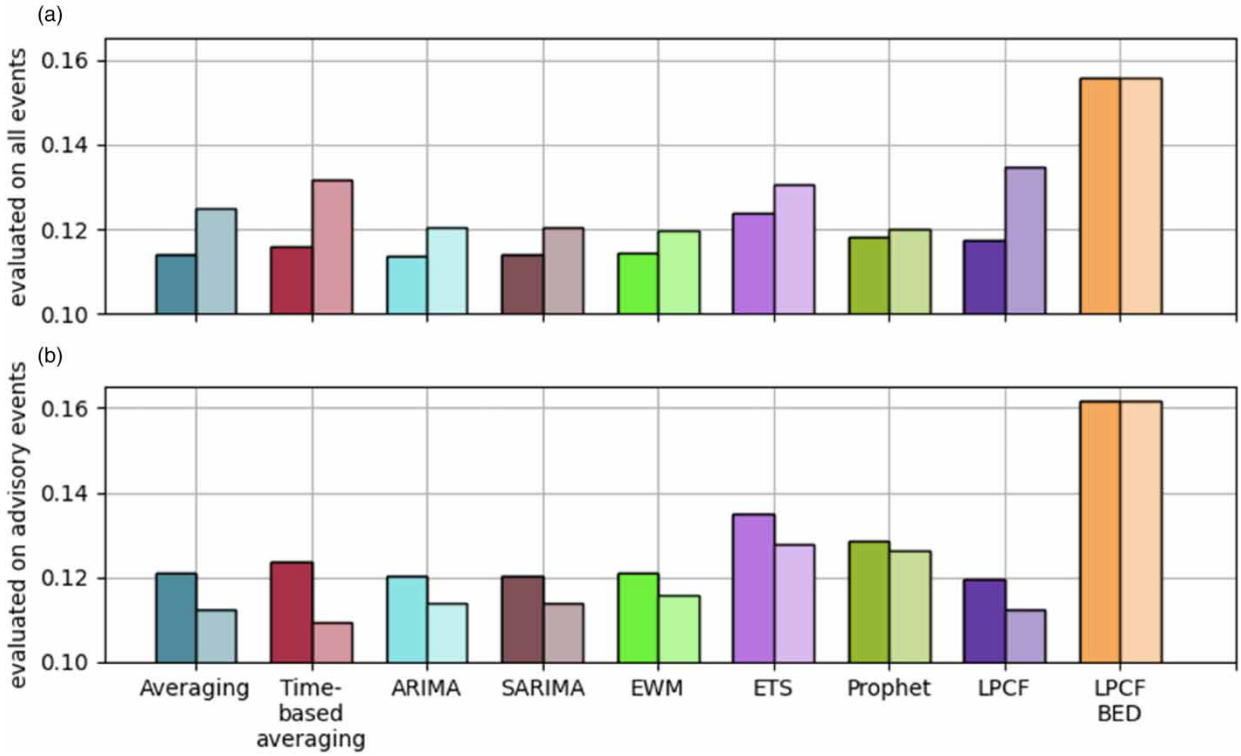


Figure 8 | Lowest RMSE for each type of forecasting methods, divided into tuning for all event types (darker shades) and advisory only events (lighter shades) and evaluated on all event types (a) and advisory events (b). Note that the *y*-axis is clipped from 0.1 for visual interpretation.

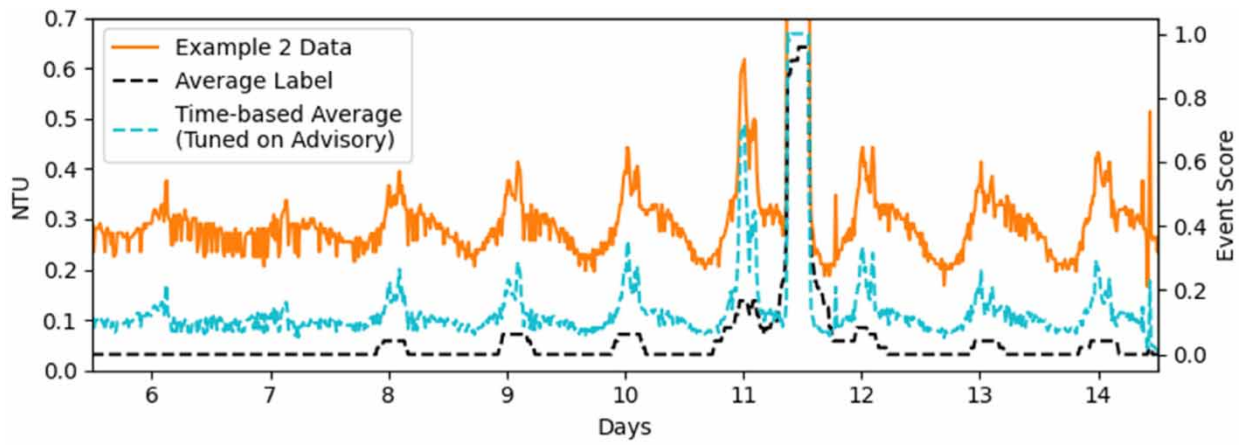


Figure 9 | Example 2 turbidity data (clipped exclude the alarm event on day 11) and averaged-out labels versus sigmoid output using time-based averaging method, optimised for advisory events.

seen about half a day before the alarm event occurred. This demonstrates the promise of this approach in being part of a proactive management system.

4. DISCUSSION

This research presents an evaluation of approaches to analyse and understand events in DWDS turbidity time series data and uses a crowd-sourced labelled dataset as a benchmark. This evaluation process presents an alternative approach to overcome

the difficulty of linking turbidity data with confirmed real-world events. With six turbidity examples and 48 participants in total, but covering a wide range of companies/organisations, conclusions should be considered with this relatively small sample size in mind. The number of examples included was limited by how long these sessions could be run, while obtaining more participants would be challenging without reducing the level of domain expertise. Following reflection on the results of the four labelling exercises, a three-level turbidity event scale was defined as advisory, alert, and alarm (Table 3). The presence of alert and alarm turbidity events in Examples 1 and 2 impacted labellers interpretation of advisory events. Such advisory events were easier to interpret in Examples 3–6, with participants considering many noteworthy. As highlighted in the background section, even smaller turbidity events with responses <1 NTU can suggest in-network deterioration and may provide valuable precursor information about levels of discolouration risk within DWDS. This research identified and was subsequently able to focus on proactive discolouration approaches as well as reactive measures by differentiating through consideration of alert and alarm events. Flat-line alert and alarms at 2 and 4 NTU are reliant on the turbidity sensor accuracy and overall data quality. Lower flat-line approaches would be too dependent on sensor calibration accuracy and background turbidity. The residual calculations performed in this research are however agnostic to turbidity baseline values, instead looking for increases compared to recent data. Therefore, these approaches are not reliant on sensor calibration accuracy. However, a prior data quality assessment is required, particularly to remove sensor errors such as drift that may occur in the data which, if left, could interfere with the residual calculations. This was performed on the examples in this research using a set of data quality assessment rules previously developed (Gleeson *et al.* 2023).

The labelling and evaluation process to mimic human interpretation is something that could be repeated for different parameters, or for turbidity with additional contextual information, such as flow data or customer contacts. Additionally, it could be run for specific teams or companies, with the aim to develop a solution that best matches their requirements and collective intuition. This highlights the need to clearly understand what information or insight is required prior to developing automation techniques, and the need to match such techniques to the data and insight sought. The results show that many residual calculation approaches, using both statistical averaging and time series forecasting, can be used in combination with a sigmoid function to produce an event score time series. Method selection may, therefore, come down to decisions about computational power, processing time and number of adjustable parameters. The event score time series output can form an event detection system or can be used to understand network conditions or performance. This understanding could then be applied to compare performance across networks or allow temporal analysis to detect changing performance if mobile monitors are deployed on a shorter term but repeating basis.

4.1. Human interpretation of turbidity events

Participants were asked to highlight 'noteworthy periods of data to be flagged for further consideration'. The exercise did not provide any additional contextual information, such as sensor location or other supporting information such as flow data or connected sensors also measuring turbidity or other water quality parameters. Such information is important when further analysing turbidity events in DWDS but were outside of the scope of this specific research for identifying data of interest for such further interpretation. The experience of visualising each example, one by one, within the trainset application told a story and may have influenced the participants interpretation and understanding. However, this is unavoidable with any visualisation of complex data, but by averaging across multiple participants it is hoped this effect was limited. The results from the labelling exercises demonstrate a variety in how domain experts interpret turbidity events. This highlights the complexity of the question of what a turbidity event even is; something that is often incorrectly considered as a black and white problem. Additionally, the responses show that context is everything when it comes to human interpretation. Something is only noteworthy if it stands out in the context in which it is presented. Analysis of how participants interpreted Examples 1–2, compared to Examples 3–6, as shown in the scatter plot in Figure 4, highlights how the presence of larger events impacts interpretation of lower-level data. Larger events seen in Examples 1 and 2 led them to ignore the lower-level events also occurring in these examples which are less visible due to the y-axis scale, yet these are similar in magnitude to those seen in Examples 3–6 that most participants acknowledged as events. This demonstrates that human interpretation is inherently subjective. By contrast, a computer will follow instructions precisely and repeatedly. It also highlights that when presented with these lower-level turbidity events unbiased by larger events, participants tended to consider them noteworthy.

Even when participants provided consistent labelled responses, there is an assumed capability that cannot be proven that a participant working in this domain is sufficiently skilled. Though all labellers are actively working in positions where they

deal with and understand turbidity and discolouration, high-frequency turbidity time series data like the examples presented is relatively newly available. This means even domain experts may not necessarily be very experienced in interpreting such data. Similarly, it is not possible to determine whether participants were influenced by external opinions or factors. Some difficulties were encountered during the labelling exercises, with some participants only labelling one example, or leaving just one unlabelled. Due to the anonymity of the responses (required to meet ethics standards), it was not possible to question participants giving invalid responses. Therefore, such responses were omitted from this research. In total 48 verified labelled responses were included. This included responses from Session 3 that consisted of some unlabelled examples, learned in a post labelling debrief. Session 3 was run externally, and the participants were given a slightly different event definition, where an event was considered anything requiring immediate action, so over 4 NTU or at least two customer contacts. This explains why there were significantly lower levels of labelled data in this session and demonstrates how easily even domain experts are influenced when given specific instruction.

The examples included in the labelling exercise were selected to include different types and magnitudes of turbidity events. However, it is not possible to include all possible scenarios with limits also required to make the labelling exercise practical for participants. The examples were checked to be clear of sensor errors, though in reality differentiating sensor errors from genuine events can be difficult. Example 4 was the only example to have been artificially concatenated, to understand how different theoretical event types may be interpreted. Looking at the corresponding averaged-out labels (Figure 3(d)), the first event, representing a hydraulic-induced mobilisation type event, had the highest event score of 0.68, with the other three event types, the single point event, the baseline change event, and the changing diurnal pattern event, not exceeding a 0.3 event score. The final event in Example 4 is a combination of these four events and the resulting event score shows an increase in interpreted noteworthiness due to this combination, with the start of this combined event exceeding 0.4. Ultimately it was decided that focusing on events at different scales was more useful for this research, though future research could focus more on categorising different event types.

4.2. Event score calculation

The approaches used to analyse events in turbidity in this research were performed with the understanding that turbidity events are not necessarily rare, rendering many outlier or anomaly detection methods developed in other fields unsuitable. Time series forecasting methods were explored, with the aim to obtain residual values that enable noteworthy periods of interest to be sufficiently emphasised for subsequent conversion to event score time series with values between 0 and 1, comparable to the averaged-out labels. Calculating event score time series that matched the averaged-out labels was not trivial, in particular finding a solution that generalised across the different examples. The optimisation and sensitivity analysis allowed for each method, and associated input parameters, to be compared in terms of their suitability for this task. This research did not focus on the most accurate forecasting approach, but instead investigated what approach best enables periods of noteworthy data to be highlighted. For this reason, averaging approaches worked effectively at ignoring periods of increased turbidity and in doing so these periods were revealed in the residuals. Some interesting outcomes about window type and forecast horizons that are useful for analysing turbidity events were uncovered. Expanding window types worked best across multiple methods, meaning more data tended to be beneficial in the time scales examined in this research. Short forecast horizons run into problems during an event where the forecasts start to account for the elevated turbidity. For this research, a forecast that effectively ignores increases is beneficial with a 24-h horizon achieving this. It has the added benefit of accounting for any seasonality present, in this case seasonality referring to repeating daily trends. As turbidity can contain diurnal trends, typically associated to hydraulic demand patterns, several methods that can account for these were included. The ETS and SARIMA methods both required at least two full cycles of data to include seasonality effects. For this reason, errors were calculated excluding the first 3 days. This meant more methods could be included in the comparison and omits potentially spurious forecasts at the very start of the time series, a problem often referred to as the 'cold start' problem. Due to the length of the examples included, between 15 and 75 days, seasonality effects over longer periods such as seasons or annually were not considered.

Modifying expanding average approaches to consider data at the same time of day improved performance, but only when looking at advisory events with no improvement seen across all examples. Advisory events are subtler by nature and more likely to be confused with diurnal fluctuations, which can vary by network and location. The time-based approach accounts for this factor, and matches with human interpretation that repeated diurnal fluctuations, such as those present in Examples 3 and 5, are not noteworthy and that it is changes in patterns that should be the focus. ARIMA approaches resulted in the best overall performance against all six examples. SARIMA was found to be extremely slow when provided with a seasonality

period of 96, meaning not all combinations could be explored but suggesting it is unsuited for this research. The exponential smoothing approaches within the ETS framework did not perform as well as other methods, perhaps due to putting too much weight on recent data, though EWM performed better on both advisory and across all events. Future research could include additional parameters to be used as exogenous variables to improve turbidity event analysis. The CANARY LPCF and BED output was more binary than the averaged labels were, with some complexity lost due to the additional step of determining whether each datapoint is an outlier, before counting the outliers to determine event probability. Therefore, the BED output was not well-suited to this research. By instead applying a sigmoidal fuzzy logic membership directly to the residuals, the complexity of the labels averaged out could be better approximated and a better solution was found. This reinforces how well-suited the sigmoidal approach is to this research due to its ability to transform residual time series into output that matches the complexity and fuzziness found in the averages labels.

4.3. Proactive versus reactive events

Supporting the approach to mimic human interpretation of turbidity events using event score calculations, this research also examined flat-line detection methods at different turbidity limits. Any event exceeding 4 NTU is exceeding regulatory limits for end-users, meaning these warrant the highest level of response, regardless of human judgement. Therefore, these events fall naturally into being categorised as alarm events. Alert events are a step lower, but represent significant deterioration compared to the 1 NTU limit at treatments works exit. For the purposes of this research, and to create a convenient division between the first two examples and the final four turbidity datasets used, a limit of 2 NTU was selected, though other values such as 1 NTU could be selected. A system that is only reactive does not prevent events from occurring and as drinking water is required to be below 1 NTU when leaving a treatment works, even low-level turbidity events are evidence of deteriorating water quality during transit. Such low-level advisory events do not come close to breaching regulatory limits and, as such, are generally ignored. However, capturing these events digitally enables whole networks and multiple sensors to be analysed automatically, meaning extra information can be used to improve strategic management of these assets.

The next step for this research is, therefore, incorporation into an automated event analysis system consisting of reactive alert and alarm event detection as well as novel proactive advisory alarms based on calculated event score time series.

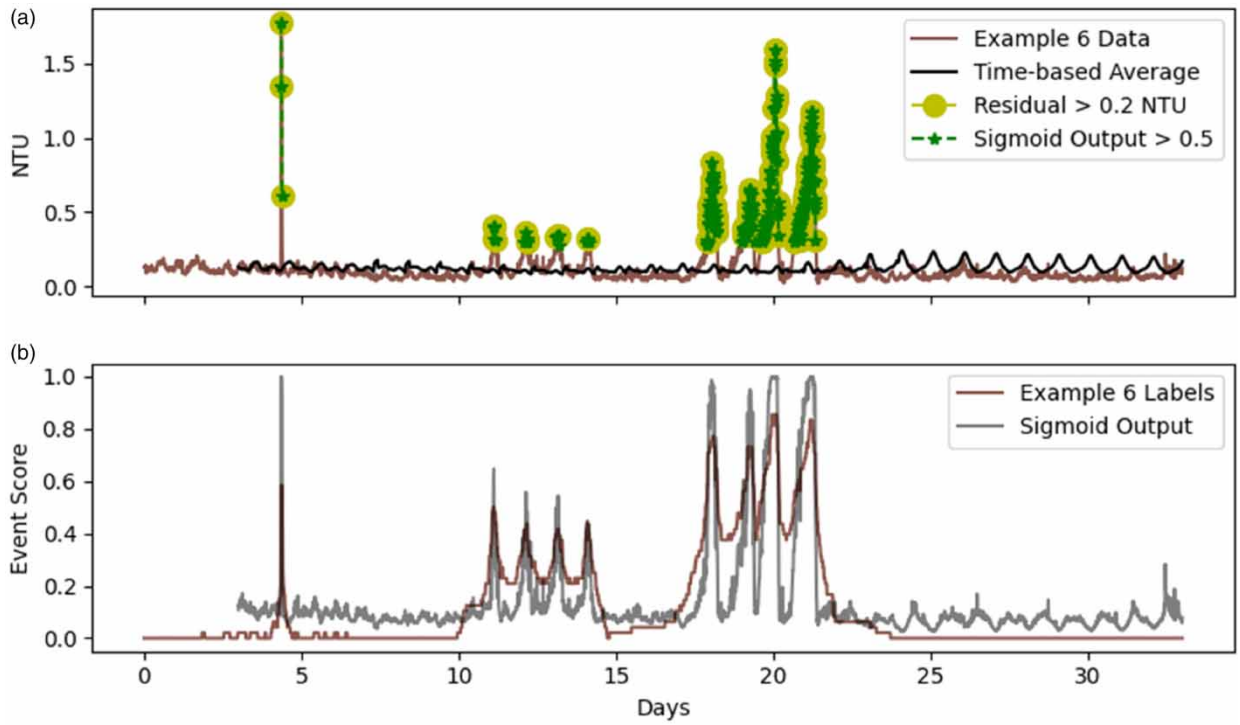


Figure 10 | Advisory event detection on Example 6 using a 0.2 NTU limit applied to the residual, and a 0.5 threshold using the sigmoid output (a), with equivalent average labels and sigmoid output (b).

One approach to converting the event score time series into proactive advisory alarms is to simply take a threshold and report any exceedances. This in effect would be similar to applying a flat-line threshold on the residual time series, though instead it would be applied to the more easily understandable event score with threshold values between 0 and 1. The similarity between these two approaches is further demonstrated and applied to Example 6 in Figure 10, which shows how a residual-flat-line threshold looking for any residuals >0.2 NTU results in a similar outcome to a threshold of 0.5 applied to the event score time series. Note the flat-line is applied to the residual, not to the turbidity data. This shows that practical application of this research may not necessarily require the sigmoidal function, though its use was essential in determining the approach that best approximated the gained insight from the labelling sessions. This research provides a platform from which such a system could be built, but the specific details of how this could be used to issue advisory alerts that aide strategic management require further understanding of what is desired.

5. CONCLUSIONS

This research shows how complex and time-consuming human interpretation of turbidity time series data from DWDS can be mimicked in real-time by computing algorithms. Automating such interpretation provides a rapid and more extensive capability to understand network performance, allowing for focussed strategic and operational decisions to manage in-network discolouration. The crowd-sourced labelling exercises undertaken represents a novel approach that addresses the difficulty in obtaining confirmed real-world events, while also highlighting the need to fully understand what is wanted from the data before developing analytic methods. These exercises informed a turbidity event scale that considers reactive alarm (>4 NTU) and alert (>2 NTU) events, alongside proactive advisory (<2 NTU) events. For alert and alarm events, a flat-line approach is considered best, assuming quality assured data are available. A time-based averaging approach was found to work best at identifying advisory events. These approaches require little computational power and could be applied in real-time.

ACKNOWLEDGEMENTS

This research has been supported by an Engineering and Physical Sciences Research Council (EPSRC) studentship as part of the Centre for Doctoral Training in Water Infrastructure and Resilience (EP/S023666/1) with support from industrial sponsor Siemens UK. We express our gratitude to the water companies that provided data and all the participants in the labelling exercises.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Aggarwal, C. C. 2016 *Outlier Analysis*, 2nd edn. Springer New York, New York, NY. Available from: http://link.springer.com/10.1007/978-1-4614-6396-2.

Blázquez-García, A., Conde, A., Mori, U. & Lozano, J. A. 2020 A review on outlier/anomaly detection in time series data. Available from: http://arxiv.org/abs/2002.04236.

Boxall, J. B. & Saul, A. J. 2005 Modeling discoloration in potable water distribution systems. *J Environ Eng* **131** (5), 716–725. Available from: https://ascelibrary.org/doi/10.1061/%28ASCE%290733-9372%282005%29131%3A5%28716%29.

Chai, T. & Draxler, R. R. 2014 Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geosci Model Dev* **7** (3), 1247–1250. Available from: https://gmd.copernicus.org/articles/7/1247/2014/.

Chandola, V., Banerjee, A. & Kumar, V. 2009 Anomaly detection. *ACM Comput Surv* **41** (3), 1–58. Available from: https://dl.acm.org/doi/10.1145/1541880.1541882.

Cook, D. M., Husband, P. S. & Boxall, J. B. 2016 Operational management of trunk main discolouration risk. *Urban Water J* **13** (4), 382–395. Available from: http://www.tandfonline.com/doi/full/10.1080/1573062X.2014.993994.

DWI 2018 The Water Supply (Water Quality) (Amendment) Regulations 2018 SI No. 706.

DWI 2022 *Drinking Water 2021: The Chief Inspector's Report for Drinking Water in England*. Available from: www.dwi.gov.uk.

El-Zahab, S. & Zayed, T. 2019 Leak detection in water distribution networks: An introductory overview. *Smart Water* **4** (1), 5. Available from: https://link.springer.com/10.1186/s40713-019-0017-x.

Geocene. 2020 Trainset. Geocene. Available from: https://github.com/Geocene/trainset.

Gleeson, K., Husband, S., Gaffney, J. & Boxall, J. 2023 A data quality assessment framework for drinking water distribution system water quality time series datasets. *J Water Supply Res Technol.* Available from: https://iwaponline.com/aqua/article/doi/10.2166/aqua.2023.228/93861/A-data-quality-assessment-framework-for-drinking.

Gupta, M., Gao, J., Aggarwal, C. C. & Han, J. 2014 Outlier detection for temporal data: a survey. *IEEE Trans Knowl Data Eng* **26** (9), 2250–2267. Available from: http://ieeexplore.ieee.org/document/6684530/.

Halford, G. S., Baker, R., McCredden, J. E. & Bain, J. D. 2005 How many variables can humans process? *Psychol Sci* **16** (1), 70–76. Available from: http://journals.sagepub.com/doi/10.1111/j.0956-7976.2005.00782.x.

Hillmer, S. & Tiao, G. 1982 An ARIMA-model-based approach to seasonal adjustment. *J Am Stat Assoc.* **77**, 63–70.

Hochreiter, S. & Schmidhuber, J. 1997 Long short-term memory. *Neural Comput* **9** (8), 1735–1780. Available from: https://direct.mit.edu/neco/article/9/8/1735-1780/6109.

Husband, P. S., Boxall, J. B. & Saul, A. J. 2008 Laboratory studies investigating the processes leading to discolouration in water distribution networks. *Water Res.* **42** (16), 4309–4318.

Hyndman, R. J. & Athanasopoulos, G., 2021 In: *Forecasting: Principles and Practice*, 3rd edn (Texts, O. ed.). Melbourne, Australia. Available from: OTexts.com/fpp3

IWA. 2022 In: *A Strategic Digital Transformation for the Water Industry* (Grievson, O., Holloway, T. & Johnson, B. eds.). IWA Publishing, OTexts, Melbourne, Australia. Available from: https://iwaponline.com/ebooks/book/860/A-Strategic-Digital-Transformation-for-the-Water.

LeChevallier, M. W., Gullick, R. W., Karim, M. R., Friedman, M. & Funk, J. E. 2003 The potential for health risks from intrusion of contaminants into the distribution system from pressure transients. *J Water Health* **1** (1), 3–14. Available from: https://iwaponline.com/jwh/article/1/1/3/1789/The-potential-for-health-risks-from-intrusion-of.

Loning, M., Bagnall, A., Ganesh, S., Kazakov, V., Lines, J. & Kiraly, F. 2019 sktime: A Unified Interface for Machine Learning with Time Series. In *33rd Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada.

McKenna, S. A., Hart, D., Klise, K., Cruz, V. & Wilson, M. 2007 Event Detection from Water Quality Time Series. In: *World Environmental and Water Resources Congress 2007*. American Society of Civil Engineers, Reston, VA, pp. 1–12. Available from: http://ascelibrary.org/doi/10.1061/40927%28243%29518.

McKinney, W. 2010 Data Structures for Statistical Computing in Python. In *Proceedings of the Python in Science Conferences*, Austin, Texas. SciPy.

Mounce, S., Machell, J. & Boxall, J. 2012 Water quality event detection and customer complaint clustering analysis in distribution systems. *Water Sci Technol Water Supply.* **12** (5), 580–587.

Murray, R. & Haxton, T. 2010 *Water Quality Event Detection Systems for Drinking Water Contamination Warning Systems – Development, Testing, and Application of CANARY*. Cincinnati, OH. [cited 2022 May 5]. Available from: https://www.researchgate.net/publication/216301147_Water_Quality_Event_Detection_Systems_for_Drinking_Water_Contamination_Warning_Systems--Development_Testing_and_Application_of_CANARY.

Potter, M. C., Wyble, B., Hagmann, C. E. & McCourt, E. S. 2014 Detecting meaning in RSVP at 13ms per picture. *Attention, Perception, Psychophys* **76** (2), 270–279. Available from: https://link.springer.com/10.3758/s13414-013-0605-z.

Speight, V. L., Mounce, S. R. & Boxall, J. B. 2019 Identification of the causes of drinking water discolouration from machine learning analysis of historical datasets. *Environ Sci Water Res Technol.* **5** (4), 747–755. Available from: http://xlink.rsc.org/?DOI=C8EW00733K.

Taylor, S. J. & Letham, B. 2018 Forecasting at scale. *Am Stat.* **72** (1), 37–45. Available from: https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1380080.

Thudumu, S., Branch, P., Jin, J. & Singh, J. 2020 A comprehensive survey of anomaly detection techniques for high dimensional big data. *J Big Data* **7** (1), 42. Available from: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00320-x.

Ustalov, D., Pavlichenko, N., Losev, V., Giliazev, I. & Tulin, E. 2021 A General-Purpose Crowdsourcing Computational Quality Control Toolkit for Python. Available from: http://arxiv.org/abs/2109.08584.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., Vijaykumar, A., Bardelli, A. P., Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C. N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D. A., Hagen, D. R., Pasechnik, D. V., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G. A., Ingold, G.-L., Allen, G. E., Lee, G. R., Audren, H., Probst, I., Dietrich, J. P., Silterra, J., Webber, J. T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J. L., de Miranda Cardoso, J. V., Reimer, J., Harrington, J., Rodríguez, J. L. C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N. J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, Per. A., Lee, P., McGibbon, R. T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T. J., Robitaille, T. P., Spura, T., Jones, T. R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y. O. & Vázquez-Baeza, Y. 2020 Scipy 1.0: Fundamental algorithms for scientific computing in Python. *Nat Methods* **17** (3), 261–272. Available from: http://www.nature.com/articles/s41592-019-0686-2.

First received 23 June 2023; accepted in revised form 16 November 2023. Available online 27 November 2023