Water Research 124 (2017) 67–76

Contents lists available at ScienceDirect

Water Research

journal homepage: www.elsevier.com/locate/watres

Short-term forecasting of turbidity in trunk main networks

Gregory Meyers*, Zoran Kapelan, Edward Keedwell

College of Engineering, Mathematics and Physical Sciences, University of Exeter, North Park Road, Exeter, EX4 4QF, UK

ARTICLE INFO

Article history:
Received 13 April 2017
Received in revised form
20 June 2017
Accepted 15 July 2017
Available online 17 July 2017

Keywords:
Discolouration
Turbidity
Data-driven
Machine learning
Forecasting
Trunk mains

ABSTRACT

Water discolouration is an increasingly important and expensive issue due to rising customer expectations, tighter regulatory demands and ageing Water Distribution Systems (WDSs) in the UK and abroad. This paper presents a new turbidity forecasting methodology capable of aiding operational staff and enabling proactive management strategies. The turbidity forecasting methodology developed here is completely data-driven and does not require hydraulic or water quality network model that is expensive to build and maintain. The methodology is tested and verified on a real trunk main network with observed turbidity measurement data. Results obtained show that the methodology can detect if discolouration material is mobilised, estimate if sufficient turbidity will be generated to exceed a preselected threshold and approximate how long the material will take to reach the downstream meter. Classification based forecasts of turbidity can be reliably made up to 5 h ahead although at the expense of increased false alarm rates. The methodology presented here could be used as an early warning system that can enable a multitude of cost beneficial proactive management strategies to be implemented as an alternative to expensive trunk mains cleaning programs.

1. Introduction

With the advancements in science and technology, so too have customer expectations risen of the service they receive from their water provider. This has been most recently reflected through regulatory bodies placing heavier incentives, penalties and fines for water quality related issues (OFWAT, 2009). The UK regulatory body OFWAT introduced penalties for water companies that exceed an acceptable number of customer contacts for discolouration back in 2009, yet in 2013 alone over 2 million UK customers were still estimated to have been affected by water discolouration issues (DWI, 2014).

Although improvements have been made to reduce discolouration, it is almost completely dealt with in a reactive way by water companies (Blokker, 2010; Cook et al., 2015). This is done in the form of cleaning parts of the network once a sufficient number of discolouration contacts are reported in that area. With ever increasing regulatory pressures and tighter standards, it is evident that reducing discolouration is a key challenge facing the water industry today and new management methods need to be considered.

Discolouration formation is complex and not completely understood with bulk water quality, temperature, network layout, pipe material and age all believed to be factors (Abe et al., 2012; Husband and Boxall, 2011; Van Thienen et al., 2011; Vreeburg et al., 2008). Discolouration has been seen to vary even between different parts of the same water distribution network and yet is still similarly experienced throughout the world regardless of the wildly differing factors between their Water Distribution Systems (WDS) (Armand et al., 2015; Blokker and Schaap, 2015; Husband et al., 2008; Vreeburg and Boxall, 2007). Discolouration mobilisation is believed to be primarily caused by sufficiently large hydraulic changes in the WDS resulting in the detaching and transportation of the accumulated discolouration material through the network and producing discoloured water at the consumer's tap (Boxall et al., 2003; Prince et al., 2001; Vreeburg et al., 2005).

Trunk mains have been categorised as especially high discolouration risks as their size allows for them to act as a form of a reservoir for discolouration material build up (Cook and Boxall, 2011). Trunk mains can play two roles in the discolouration process, a passive role of slowly sending material downstream to build up in other distribution pipes or an active role of a widespread high consequence discolouration event if the discolouration material is rapidly mobilised.

Cleaning a trunk main can result in the improvement of bulk water quality leaving the trunk main which has been shown to

* Corresponding author.
E-mail address: gmm206@exeter.ac.uk (G. Meyers).

http://dx.doi.org/10.1016/j.watres.2017.07.035
0043-1354/© 2017 Elsevier Ltd. All rights reserved.

reduce the likelihood and magnitude of downstream discolouration events (Blokker and Schaap, 2015). This in turn also reduces the frequency of maintenance and cleaning required in downstream distribution pipes. This is further evidenced by a study in the UK showing that between 30% and 50% of discolouration events seen in the reported District Metered Areas (DMAs) could be linked to imported discolouration material from upstream trunk mains (Cook et al., 2015). The study also found that depending on the DMA, between 0% and 51% of discolouration contacts could be linked to upstream trunk mains. While the mean was only 9%, this shows that some WDS are significantly more susceptible to discolouration events from trunk mains than others.

The benefits of cleaning trunk mains are easy to observe in extreme examples such as where a trunk main supplying 1.75 million customers saw an associated 62% reduction in overall customer contacts after being cleaned (Husband et al., 2010a). However the significant consequences and logistical complexities associated with trunk mains mean that regular cleaning programs are expensive and difficult to implement. This has resulted in very infrequent trunk main cleaning programs usually carried out reactively in situations where the benefits are evident (Husband et al., 2010b; Vreeburg and Boxall, 2007).

According to our best knowledge, the only turbidity prediction model validated on trunk mains in the field is the Prediction of Discolouration in Distribution Systems (PODDS) model (Husband and Boxall, 2016). This model was developed for the cleaning of single pipe stretches with minimal invasive action required. However, due to unknown pipe conditions and discolouration material build up rates, this model requires hydraulic model with onsite model calibration before each use making it unsuitable in the context of continuous (rather than individual event based) turbidity prediction. The Variable Condition Discolouration Model (VCDM) builds upon the PODDS model and is capable of emulating material erosion and regeneration in pipes over time (Furnass et al., 2014). However, the VCDM is currently unverified as it requires repeated site specific turbidity events for model parameter calibration and a calibrated hydraulic model to track the turbidity response. This twin modelling constraint also increases the complexity and therefore potential for error when applied to operational applications.

The requirement of a well calibrated and accurate hydraulic model has been noted as a major limiting factor in many existing water quality models (Machell et al., 2009, 2014; Skipworth et al., 2002; Vreeburg, 2007). This is similarly a common theme found when developing a burst detection model, water demand model or anomaly detection model that must build upon a well calibrated and accurate hydraulic model (Arad et al., 2013; Blokker et al., 2009; Machell et al., 2010; Tao et al., 2014). Whether this has prevented the use of the model in certain areas (Tao et al., 2014) or resulted in a clear decrease in model accuracy away from the calibrated area (Leeder et al., 2012), the dependency of a satisfactorily calibrated hydraulic model limits the application of these model types. These issues stem from the expense of developing and regularly updating hydraulic models and the fact that they are usually calibrated from “24 h” data reflecting an average day in the water distribution system (WDS). This creates an additional problem for accurately forecasting discolouration as discolouration events are thought to be primarily a function of irregular hydraulic disturbances that mobilises the accumulated material and are not therefore part of the “average day”.

To be free from the issues associated with using hydraulic models and to ensure application to almost any WDS that has suitable meters installed, a data driven methodology for short-term forecasting of turbidity was explored and validated on a real trunk main network. While prior work showed that a data driven methodology for forecasting turbidity was possible, the Artificial Neural Network (ANN) based model only forecasted 15 min ahead and with limited accuracy (Meyers et al., 2016). The methodology presented in this paper greatly expands on this by comparing multiple machine learning methods with an improved set of model inputs and forecasting over multiple significantly longer forecast horizons. Additionally, a completely alternative modelling approach is also presented that simplifies the machine learning objective by only predicting if turbidity will exceed a prespecified threshold at a specific time horizon in the future.

2. Site details and data

Flow and turbidity measurements were taken from a section of a trunk main network in the UK over 11 months, starting from 01 September 2013 to 01 August 2014. Hydraulic data was captured from one import flow meter and six export flow meters from the network section while turbidity data was captured from one turbidity meter.

As shown in Fig. 1, in addition to a flow meter placed immediately downstream of the upstream service reservoir, which is the sole inlet for the trunk main network, a flow meter was placed at each water exporting branch. The turbidity meter was placed just before the flow meter at the inlet to a downstream service reservoir so that all turbidity measurements have an associated exact measurement of the flow rate going through the turbidity meter. Between the upstream water inlet and downstream turbidity meter there is 10.5 km of ductile iron trunk main piping.

The flow and turbidity data was recorded at 15 min intervals with flow being logged as the sum of water through the meter during that interval and turbidity being logged as the current turbidity passing through the meter on the 15 min interval mark. Flow was measured in cubic meters per 15 min (m^3/15min) and

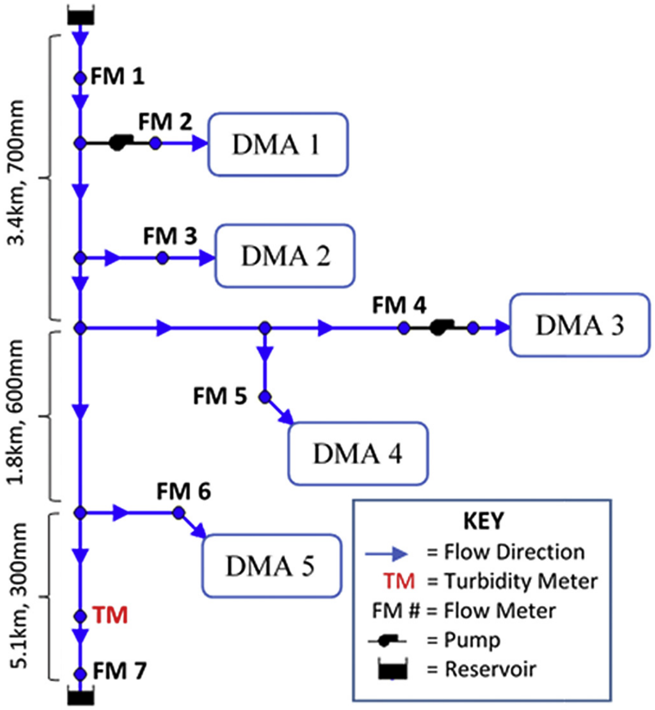


Fig. 1. Trunk main network schematic showing the placement of flow and turbidity meters. Lengths and diameters are shown next to the trunk mains connecting the upstream service reservoir to the downstream turbidity meter.

turbidity was measured in Nephelometric Turbidity Units (NTU).

As the turbidity meter was installed near a flow meter, a constant power supply and connection to the SCADA system was easily achieved. The turbidity meter was cleaned on a monthly basis and no drift was seen in the recorded turbidity data.

The raw turbidity and flow measurement data was drawn directly from a SCADA database. As each meter logged once every 15 min, the 11 months of measurement data resulted in 32,160 15 min timesteps with seven flow and one turbidity measurement per timestep. The first seven months (01 September 2013 to 31 March 2014) of data was used for calibration and the following four months (01 April 2014 to 01 August 2014) of data was used for validation. This split resulted in 20,353 timesteps in the calibration set and 11,807 timesteps in the validation set.

Flow and turbidity measurements were standardised based on the mean and standard deviation calculated from the calibration set only.

3. Turbidity forecasting methodology

3.1. Overview

There are three significantly complex processes that need to be taken into account in order for the model to correctly forecast turbidity at the downstream turbidity meter.

The first process required is that the model must detect if any hydraulic force capable of mobilising discolouration material has occurred. This is a significant challenge to the model because while there are numerous studies showing that flow rates significantly above the daily conditioning values are able to mobilise discolouration material (Husband et al., 2010a; Vreeburg and Boxall, 2007), the scale and extent to which typical flow rates cause naturally occurring discolouration events has not been investigated. Additionally, when looking at the model results it is important to take into consideration that the model only uses flow and turbidity measurements as inputs, thus any turbidity mobilised by other factors like temperature or bio-film shedding may not be picked up.

Linked to the first process, the second process the model needs to do is estimate how much turbidity the mobilisation of discolouration material has resulted in. This second process is based on the concept that if there was a significant discolouration event recently then there may be less or no discolouration material that can be mobilised presently. This process is especially important if the goal is to predict the exact turbidity value.

Finally, the third process also linked to the first process, is estimating where the discolouration material was mobilised from and then the amount of time it will take for that material to reach the downstream turbidity meter. This is important as after the discolouration material has been mobilised it can still require many hours to reach the downstream turbidity meter. Determining which general section of the network the discolouration material is being mobilised from is a complex and difficult problem (Blokker et al., 2011), however the model also has to predict the expected travel time of the discolouration material. This can mean that the model could accurately detect and estimate the size of a discolouration event but not correctly forecast when that increased turbidity will arrive at the downstream meter, resulting in a false alarm and poor overall model performance.

3.2. Turbidity forecasting models

In order to address the significant complexities required by the model, two separate modelling approaches were taken, a regression approach and a classification approach. The regression approach requires the model to predict the actual turbidity measurement at a specified period of time in the future. The classification approach requires the model to predict if turbidity will be above or below a preselected turbidity threshold at a specific period of time in the future.

For the classification approach, all the turbidity measurements are placed into one of two classes, Positive or Negative. If the turbidity measurement is above the preselected threshold then it is counted as being in the positive class, conversely if it is below the threshold then it is labelled as part of the negative class. The classification model then only predicts if the turbidity at a pre-specified future time is positive or negative and not what the actual turbidity will be.

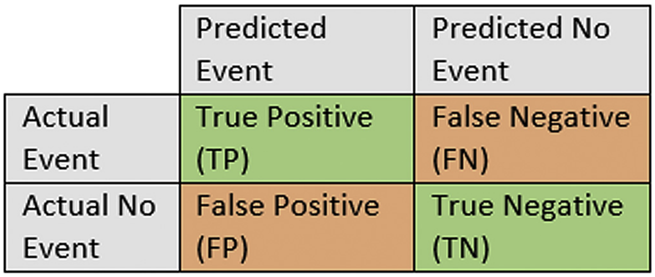
Once the model has made its prediction of positive or negative, those predictions are further divided into True or False depending if the prediction made was correct or not. A True Positive (TP) and True Negative (TN) are predictions made by the model that were correct for their respective classes, while conversely a False Positive (FP) and False Negative (FN) were incorrect predictions. This is illustrated in Fig. 2.

Three identical classification based models with different preselected turbidity threshold values of 1 NTU, 2 NTU and 4 NTU are tested and validated. The 1 NTU threshold was chosen as it is clear measurable response above the background turbidity levels and is the regulatory limit for water leaving water treatment works. The 2 NTU threshold was also chosen as it is indicative of a more serious discolouration event could soon occur and finally the 4 NTU threshold was chosen as it is the UK regulatory turbidity limit at customers' taps. Together these models could be seen as a three-level warning system where operational staff could decide at what level they wish to take action.

3.3. Input variables

The data presented as inputs to the above models is measurement data from the flow and turbidity meters. Recent historical data of turbidity and flow is required for the model to make accurate predictions about the amount of discolouration material that has been mobilised and how long it will take that material to reach the downstream turbidity meter. However, the more inputs given to the model makes the model more likely to overfit on the calibration data. Thus only a set amount of previous lagged meter measurements are given to the model as an input at any one time, this is what is known as time-delay embedding or the sliding window method (Dietterich, 2002; Gershenfeld et al., 1993), as shown in Fig. 3.

When choosing the amount of previous measurements to present to a model, also known as the size of the sliding window, it is important to select the correct size. Too small a window size will not provide the models with enough information to accurately



	Predicted Event	Predicted No Event
Actual Event	True Positive (TP)	False Negative (FN)
Actual No Event	False Positive (FP)	True Negative (TN)

Fig. 2. A confusion matrix diagram used to describe the performance of a classification based model.

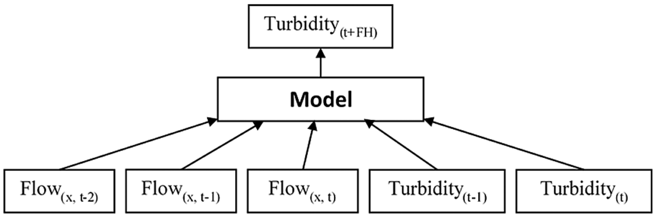


Fig. 3. A simplified diagram of where current and previous flow and turbidity measurements are used as inputs to a model predicting the turbidity value at a future time step. In the diagram "t" denotes the current time step, "x" denotes the flow meter and "FH" is the forecast horizon (i.e. how far into the future the predictions are made).

reproduce the modelled system's dynamics and thus resulting in poor prediction performance. Conversely, too large a window size can result in the model overfitting on the calibration data and requiring increased training times (Abarbanel et al., 1993; Frank et al., 2001). The False Nearest Neighbours (FNN) algorithm with minor adjustments was used to choose the optimal window size for each meter (Rhodes and Morari, 1997).

Previous work showed that providing the peak flow and turbidity values for each meter in the previous days and weeks as separate inputs improves model accuracy with only a relatively minor increase in model complexity (Meyers et al., 2016). Thus in addition to the time lagged meter measurements, three other types of flow inputs were presented to the models. These inputs are functions of current and previous meter readings and are named the *peak x day flow rate,* the *peak flow rate difference* and the *peak turbidity level.*

The peak x day flow rate is the maximum measurement seen by a specific meter in the last x days. This is in order to allow the models to determine if current flows exceed the thresholds of pipe conditioning levels or if a recent high flow rate has flushed away any potential discolouration material.

The peak flow rate difference is the difference between the current flow rate and the daily peak flow rate. While similar to the peak x day flow rate input that acts like a static threshold measurement that is unlikely to change for many days at a time, this input acts as a dynamic threshold measurement that also provides information to the model about how long ago the discolouration material was likely to have been mobilised. This should aid the models in determining when the discolouration material will reach the downstream turbidity meter.

The peak turbidity level aids the models in determining if recent flow events have removed discolouration material or if there is significant potential material build up that could be mobilised.

While too large to list individually here, the window sizes of each meter used in the turbidity prediction models ranged from three to eight time steps of 15 min, corresponding to less than an hour of previous measurements for some meters and up to 3 h of previous measurements for other meters. A daily and weekly version of the peak x day flow rate, peak flow rate difference and peak turbidity level inputs were also used. This resulted in a sum total of 95 inputs presented to each model.

3.4. *Performance metrics*

Performance metrics are an essential part of evaluating and comparing forecasting models and while a visual inspection of a graphical plot of a model's forecast is recommended (Green and Stephenson, 1986; Martinec and Rango, 1989), this can become impractical when comparing and/or presenting multiple models potentially with different parameters and over multiple forecast horizons. As performance metrics inherently tend to have a bias toward particular aspects of one model over another, multiple performance metrics have been used here (ASCE, 1993).

3.4.1. *Regression model metrics*

The *Nash-Sutcliffe Model Efficiency Coefficient (NSE)* and *Naïve Relative Squared Error (NRSE)* Metrics were chosen to assess the performance of the regression based models. These were chosen as, they provide a numerical value of performance and also useful information on how well the forecasts fitted with the observed data (Armstrong, 2001; Hyndman and Koehler, 2006).

The Nash-Sutcliffe Model Efficiency Coefficient is widely adopted by the Hydroinformatics community in measuring model performance. NSE indicates the model performance relative to the mean of all observational data as a forecast at every time step. In a similar way to NSE, the Naïve Relative Squared Error (NRSE) indicates performance relative to the naïve method. The naïve method forecasts each next time step is the same as the last observed value - this is the optimal forecast method if there is no additional information on what the next observation will be (Gershenfeld et al., 1993).

The formulas for NSE and NRSE are as follows:

$$NSE = 1 - \frac{\sum_t (F_t - A_t)^2}{\sum_t (M - A_t)^2} \tag{1}$$

$$1 - NRSE = 1 - \frac{\sum_{t=1} (F_t - A_t)^2}{\frac{t}{t-1}\sum_{t=2} (A_t - A_{t-1})^2} \tag{2}$$

where F_t is the forecast at time step t, A_t is the actual value observed at time step t and M is the mean of all observational data. Note that NRSE is subtracted from one so that a value of one indicates perfect model performance for both NSE and NRSE metrics.

3.4.2. *Classification model metrics*

Due to the infrequent nature of discolouration where little or no discolouration can occur for multiple weeks at a time, less than 1% of the logged data points are above 1 NTU. Because of this significantly disproportionate number of negatives (i.e. non-discoloration events), most error metrics that use the number of negatives as a factor is likely to be misleading. An example of this would be the *Accuracy* metric that is the percentage of observations that are correctly classified. However, a model that simply always predicted negative would have greater than 99% accuracy even though the model predicts no discolouration events. For this reason, the Accuracy metric was not used. In addition, metrics such as the *False Positive Rate (FPR)* and *Receiver Operator Characteristic Area Under the Curve (ROC AUC)* were avoided for the same reason and the *Matthews correlation coefficient* was used instead.

The Matthews correlation coefficient (MCC) is regarded as a good measure of summarising performance even when there is a skew in class sizes (Baldi et al., 2000).

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3}$$

While MCC may give an indication of performance that can be compared with other models, it does not provide significant information about the practical application of the model. Given this, the True Positive Rate and False Detection Rate metrics were additionally used here.

The True Positive Rate (TPR), also known as the probability of detection, is the probability that the model will correctly predict positive class values (i.e. turbidity above pre-specified threshold). For example, a TPR of 0.8 for a model means that of the positive class values, the model will detect 80% and miss 20% of actual events.

$$TPR = \frac{TP}{TP + FN} \tag{4}$$

The False Discovery Rate (FDR) is the probability a model predicts a positive (i.e. a discolouration event, as defined above) when in reality no such event occurred (also known as a false alarm). It is important to keep FDR low to ensure that operational staff maintain confidence in the system when an alarm sounds.

$$FDR = \frac{FP}{FP + TP} \tag{5}$$

3.5. Data driven methods

Building the turbidity prediction models (both regression and classification based) requires the use of a machine learning method. Three mainstream machine learning methods tested here are an Artificial Neural Network (ANN), a Random Forest (RF) and a Support Vector Machine (SVM) (Breiman, 2001; Glorot and Bengio, 2010; Smola and Schölkopf, 2004). These methods were chosen based on three factors: (a) a significant, proven ability to generalise complex non-linear relationships between potentially noisy inputs and outputs; (b) the learning of this non-linear relationship using a significantly different mechanism to the other two methods; (c) that a regression and classification version of the method exists.

ANNs were developed based on models of human brain function and are able to solve complex problems. The ANN used here was a feed-forward multilayer perceptron ANN with hidden layers using the hyperbolic tangent sigmoid transfer function and the output layer employing a linear transfer function. The ANN was trained with a stochastic gradient-based optimizer (Kingma and Ba, 2014). The number of hidden layers and the number of neurons in each hidden layer was set via a hyperparameter tuning process that utilised 5-fold cross-validation on the calibration data only. For the majority of the classification and regression models developed this resulted in a single hidden layer ANN with 60 neurons in the hidden layer.

RFs can essentially be considered an ensemble of decision trees where each individual decision tree is a weak classifier but good performance, scalability and generalisation can be achieved through an ensemble and the ‘bootstrapping’ of data. The RF used here was the Extra Trees RF variant which adds an additional layer of randomness to the model to further decrease the chance of overfitting (Geurts et al., 2006). 1000 weak classifiers and the entropy splitting criterion was used in the RF. The percentage of features to consider when looking for the best tree split and the maximum depth of trees was set via hyperparameter tuning using 5-fold cross-validation on the calibration data only. For the majority of classification and regression models, hyperparameter tuning resulted in only 10% of features being considered at each tree split with the maximum depth of each tree being limited to 2.

SVMs aim to find the maximum margin between decision boundaries in training data with the aim to be less prone to overfitting and thus have a lower generalisation error when tested. The kernel, gamma and the regularisation penalty used by the SVM was set via hyperparameter tuning using 5-fold cross-validation on the calibration data only. For the majority of regression models, hyperparameter tuning resulted in the use of the radial basis function (RBF) kernel with a gamma of 0.01 and 0.1 for the regularisation penalty. For the majority of classification models, hyperparameter tuning also resulted in RBF kernel but with a gamma of 0.001 and 1.25 for the regularisation penalty.

Each machine learning method required less than five hours to calibrate on an Intel Core i7-5600U CPU and 16 GB of RAM. Once calibrated each machine learning method was able to produce forecasts for the four months of unseen validation data in seconds. This indicates that regular recalibrations to include new data is feasible.

4. Results

All results and figures shown below have been calculated from the unseen (i.e. validation or testing) data sets by using models calibrated on seen (i.e. training) data sets.

4.1. Regression based turbidity forecasting models

Fig. 4 shows the performances of each of the regression based turbidity machine learning prediction models for different forecasting horizons up to 2 h. While it is not clear from this figure if the ANN or RF model perform best overall, it is clear that they both do outperform the SVM model. The SVM model's poor performance indicates the SVM was unable to generalise well on the unseen (i.e. validation) data. The ANN and RF model perform well initially before rapidly degrading in performance as the forecast horizon increases. By the 20–35 min forecast horizon region it is questionable as to how reliable the models would be. Forecasts made after 50 min forecast horizon are only slightly better than forecasting the mean of the measurement.

Fig. 5 shows the forecasts of the regression based ANN turbidity prediction model forecasting 20 min ahead against the actual turbidity measurements of a typical event. While the model can be seen to accurately forecast the start and shape of the turbidity event, it under-predicts the magnitude of the event. This is the typical behaviour of the ANN model across all events.

The regression based ANN and RF turbidity prediction models have potential to aid operational staff in the immediate future with

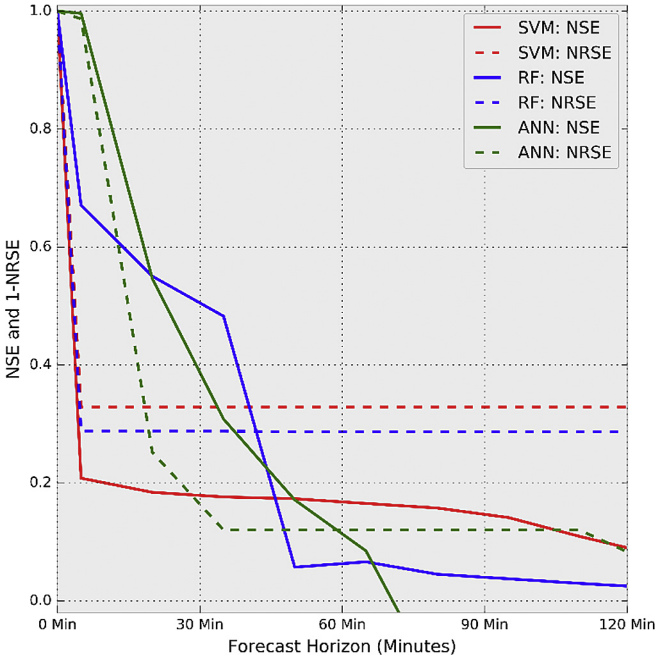


Fig. 4. Nash-Sutcliffe Model Efficiency (NSE) and the 1-Naïve Relative Squared Error (1-NRSE) performance metrics (lower values indicate worse model performance in both cases).

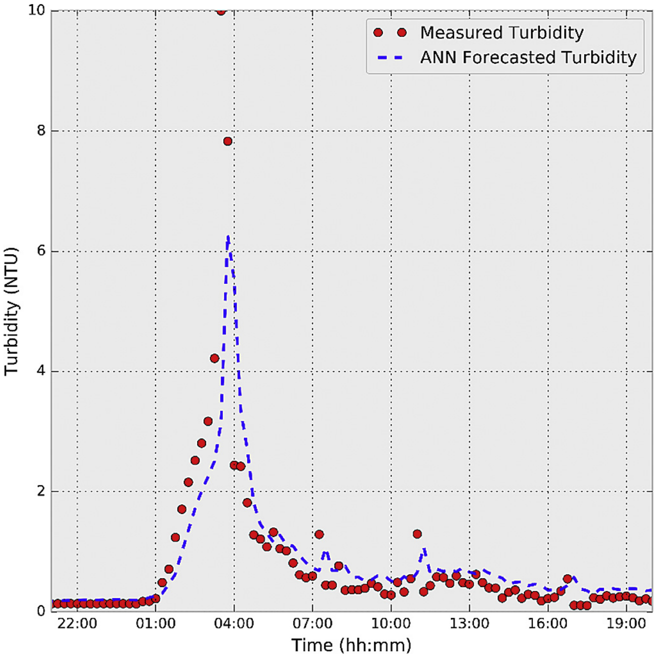


Fig. 5. The regression based ANN turbidity prediction model forecasting 20 min ahead over a typical event in the validation set.

relatively accurate forecasts up to approximately 20 min, however that is unlikely to be sufficient lead time for operational staff to act upon in most cases.

4.2. Classification based turbidity forecasting models

Fig. 6 shows the mean Matthew's correlation coefficient (MCC) calculated from the classification based turbidity prediction models for each of the three machine learning methods over a range of forecast horizons up to 8 h ahead. As it can be seen from this figure, the ANN and RF based models clearly outperform the SVM based model. While the ANN based model initially outperforms the RF based model, it quickly becomes worse performing than the RF based model as the forecast horizon increases.

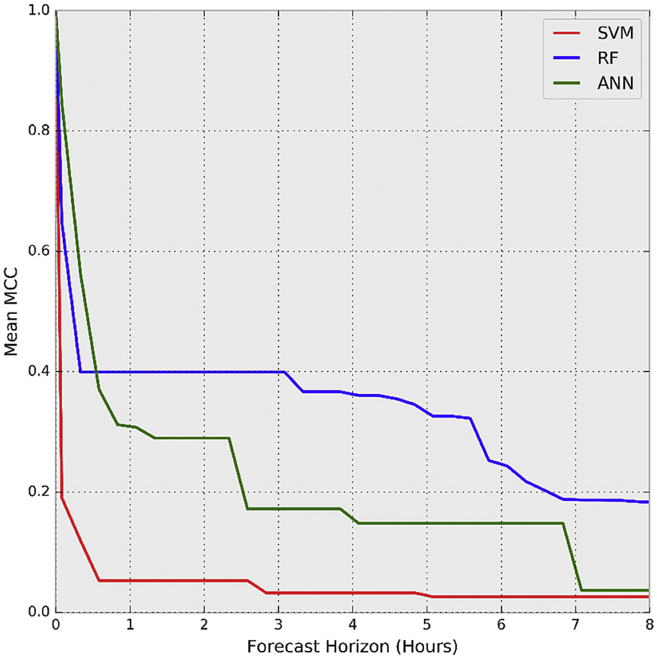


Fig. 6. The mean Matthews correlation coefficient (MCC) is shown for each of the classification based turbidity prediction models over an 8 h forecast horizon. A lower value indicates worse performance.

The sharp drop in the mean MMC value seen from forecast horizons less than 30 min across all models is primarily due to an increasing number of false positives which have a significantly stronger influence on the MMC because of the large class skew (significantly fewer actual event measurements limits the number of possible true positives but not false positives). The particularly poor performance of the SVM based model indicates that the SVM was unable to generalise well and produced a greater number of false predictions as a result. From the results in Fig. 6 it is determined that of the models presented here, the RF based model has the best performance overall.

Based on these results a closer examination of the RF based turbidity model is shown in Fig. 7. This figure shows the values of TPRs and FDRs for the RF model at the three different threshold values of 1 NTU, 2 NTU and 4 NTU and for different forecasting horizons ranging from 5 min to 8 h.

From Fig. 7 it can be observed that it is possible to obtain high TPR values (e.g. values above 0.9) even for longer forecast horizons of over 5 h. The TPR values obtained for all models are high initially for all models however the 2 and 4 NTU threshold models quickly drop to a mid-region values (0.3–0.45) for lead times around 30 min. TPRs for the 2 and 4 NTU threshold models then again increase to high values of above 0.9 for lead times of over 5 h before the 2 NTU threshold model once again drops. The increases in each model's TPR are explained by the models' sensitivity and trade-off between detecting real turbidity events (i.e. the TPR) and raising false alarms (i.e. the FPR). This is because each forecast horizon of the classification based model was calibrated with more weight placed on keeping the FPR low, and thus also keeping the FDR low, than was placed on keeping the TPR high. This extra weight was removed if the TPR drops to 0.3 or lower during calibration.

The TPR and FDR of the 1 NTU threshold model increases at approximately 3 h and 30 min because during calibration the TPR would have dropped below 0.3, thus the model made the trade-off of being more sensitive and accepting more false positives in order to keep the TPR higher. This is also why the FDR for all models in Fig. 7 is seen to only monotonically increase.

It can also be observed from Fig. 7 that unlike the 2 and 4 NTU threshold prediction models, the 1 NTU threshold prediction model never reaches a TPR of 1. A closer examination of the model showed that this is primarily because the 1 NTU threshold prediction model struggled to forecast the many sudden transient turbidity spikes that slightly exceed 1 NTU for a single time step. Some of these are accurately forecasted at very short lead times (less than 20 min) by autocorrelation with previous turbidity but are quickly lost when forecasting further into the future. However, events of this magnitude and duration are highly unlikely to be detected by customers and their short duration means that it is not clear what remedial action an operator could take to mitigate this.

Note that the FDRs shown in Fig. 7 can be misleadingly high as the majority of FPs that make up the FDRs are not randomly spread out, but instead clustered around actual turbidity events. This is shown in Fig. 8 where the 2 and 4 NTU threshold RF turbidity prediction models are shown forecasting 5 h and 20 min ahead. As it can be seen from this figure, both models correctly predict that turbidity will go above the corresponding thresholds hours before any rise in turbidity is observed. This clearly shows that the model is able to detect the mobilisation of discolouration material via flow meter measurements alone. However, the high number of FPs that

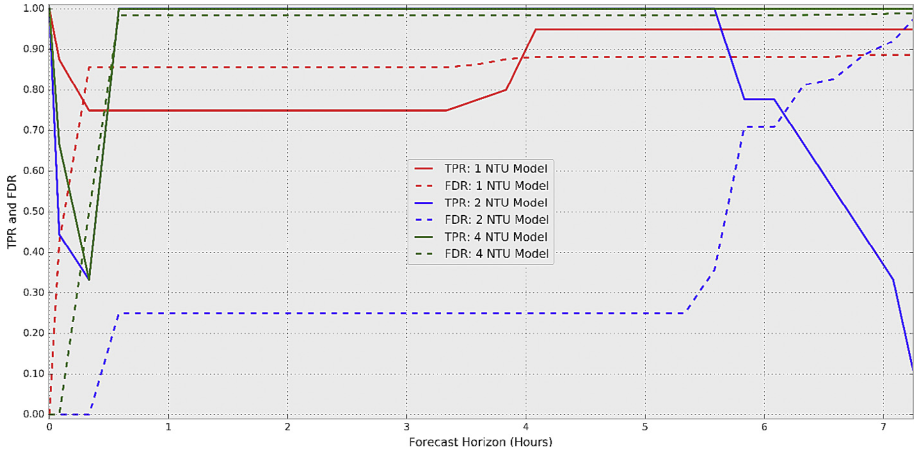


Fig. 7. The validation performances of the 1, 2, 4 NTU threshold classification based models across a range of forecasting horizons. The True Positive Rate (TPR) shows how well a model can detect real events and the False Detection Rate (FDR) gives the likelihood of a real event that has been detected turning out to be a false positive. Perfect results would show a TPR = 1 and FDR = 0.

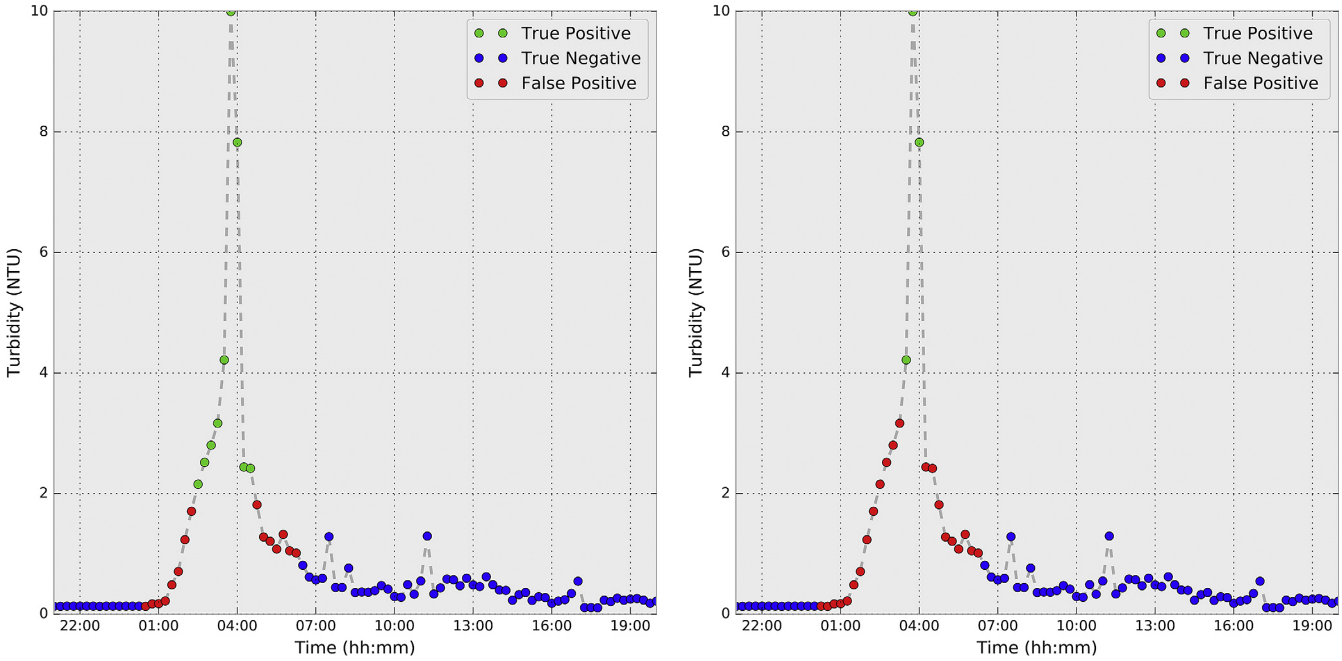


Fig. 8. The 2 and 4 NTU threshold RF turbidity models forecasting 5 h and 20 min ahead over a typical turbidity event in the validation set. (a) 2 NTU Threshold Prediction Model. (b) 4 NTU Threshold Prediction Model.

surround the event show that the models struggle to predict exactly when the discolouration material would reach the downstream turbidity meter. Because the models detected the mobilisation of discolouration material but struggled to estimate its travel time to the downstream turbidity meter, the models learned to predict positive for all values close to when the discolouration material was estimated to arrive.

While this predicting positive for all values method ensures that TPs are not missed and enables the models to have high TPRs for long lead times, it also means many FPs are generated in the process which raises the FDRs. For the 2 and 4 NTU threshold RF turbidity models forecasting 5 h and 20 min ahead, the FPs that occur 6 h before and after all actual events account for 79% and 88% of FPs respectfully.

It is worth noting that as the FDR is normalised by the number of true and false positives, the FDR for the 4 NTU threshold prediction model was expected to be higher than the other models because there are fewer 4 NTU or greater observations and thus fewer positives for the model can accurately predict. This is shown to some extent in Fig. 8 where the unequal number of TPs per threshold prediction model results in this individual event having a FDR of 63% for the 2 NTU threshold prediction model but a FDR of 88% for the 4 NTU threshold prediction model.

While the 2 and 4 NTU threshold prediction models may be more appealing to detecting larger impact discolouration events, the high sensitivity required by the 1 NTU threshold prediction model should not be overlooked. Fig. 9 shows the 1 and 2 NTU threshold RF turbidity prediction model forecasting over a low NTU event 5 h and 20 min ahead. The event seen in Fig. 9 is a non-typical turbidity event caused by trunk mains cleaning in an upstream branch of the network.

It can be observed here that that the 1 NTU threshold prediction

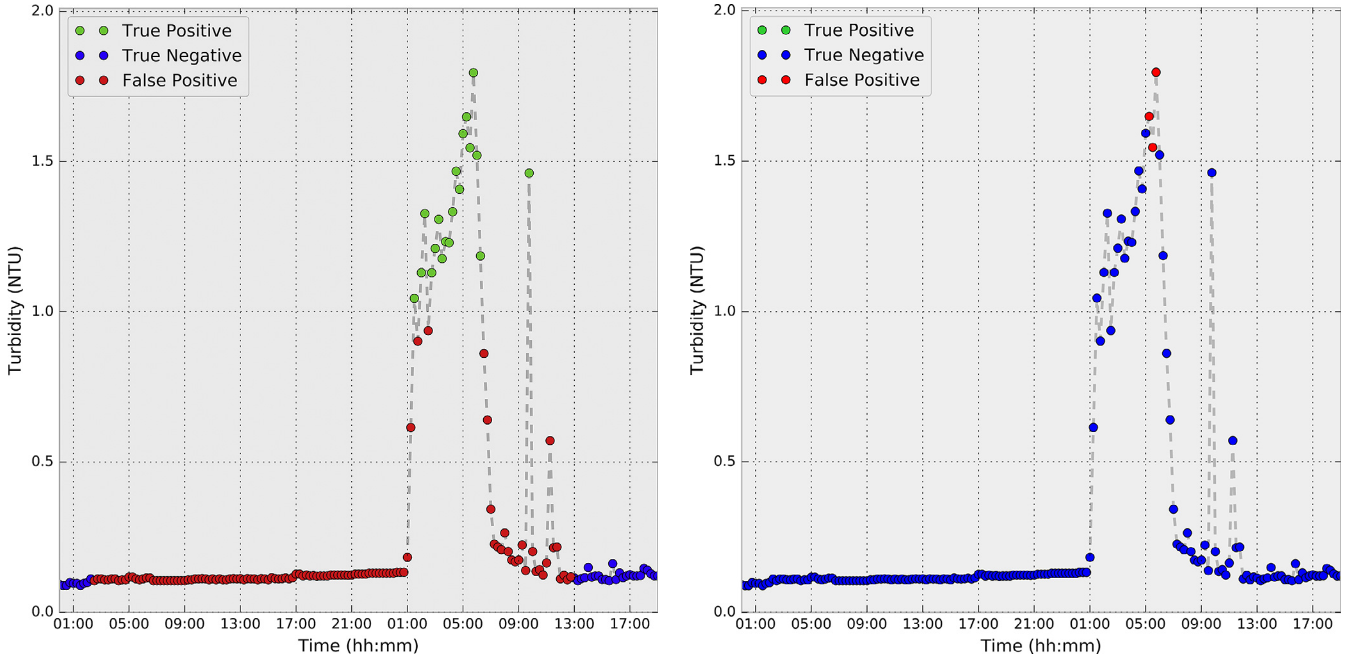


Fig. 9. The 2 and 4 NTU threshold RF turbidity models forecasting 5 h and 20 min ahead over a non-typical low concentration turbidity event in the validation set. (a) 1 NTU Threshold Prediction Model. (b) 2 NTU Threshold Prediction Model.

model picked up a change in the WDS over a day before the resulting turbidity event. The change that was detected was when operational staff started filling up the upstream and downstream service reservoirs in preparation for the cleaning works. Fig. 9 (b) shows that because the 2 NTU threshold prediction model was calibrated to not be as sensitive to smaller changes in measurement data, it did not raise the same alarm.

5. Discussion

Three data-driven turbidity forecasting models based on Artificial Neural Networks, Random Forests and Support Vector Machines were developed and evaluated in this paper. Each of these models takes current and past flow and turbidity measurements at a number of locations in the system to either directly predict turbidity (regression based model) or classify turbidity as being above a pre-specified threshold (classification based model). Three classification threshold values of 1, 2 and 4 NTU were used to define the occurrence of different discolouration events and to cover a variety of possible model applications. Forecasting horizons were systematically increased from 5 min to 8 h ahead and the methodology was tested and verified on a real water system in the UK with flow and turbidity measurement data available for 11 months.

The ANN and RF regression based turbidity prediction models were able to produce relatively accurate forecasts up to approximately 20 min ahead. This is almost certainly an insufficient amount of lead time for a water company's operational staff to act upon.

The RF classification based turbidity prediction models performed very well at predicting turbidity events with a lead time of several hours, as shown by the high TPR values obtained. However, as lead time increases the model struggles to predict when the increased turbidity will arrive at the downstream turbidity meter thus there is a significant increase in the number of false positives around the forecasted turbidity events. Ultimately, at the expense of the FDR, the largest reliable, increased turbidity forecasting time horizon for all classification models in this trunk main system is just over 5 h. While lead times of the order of several hours does not allow the system operators to do anything more substantial in nature (e.g. flush the relevant pipe(s)), this should provide water company operational staff with a sufficient lead time to act on limiting or mitigating the discolouration event (e.g. to reduce flow rates to prevent further discolouration mobilisation). This can be considered more proactive compared to the commonly adopted reactive approach of waiting for customers to report occurrences of discoloured water and then cleaning the trunk main only after a sufficient number of these complaints have been received. Thus of the two modelling approaches, the classification based approach is deemed to be more practical than making direct predictions of turbidity values (regression based approach).

In the classification based approach a maximum turbidity threshold of 4 NTU was assessed which is the UK regulatory limit for discoloured water at customers' taps. However the visibility threshold where water can be deemed to be different from normal by a customer it is estimated to be approximately 10 FTU (Slaats et al., 2003). While this does mean a 4 NTU prediction may not actually result in a customer contact, immediate reduction of the WDS flow rates may prevent higher concentrations of discolouration material from being mobilised that would be visible. While this is limited to WDS where the flow rates can be automatically controlled, the immediate reduction of WDS flow rates after a high turbidity forecast could also significantly increase the travel time of already mobilised discolouration material and thus significantly increase the lead time for operational staff to act therein.

The discovery of the optimal trade-off between TPR, FDR and lead time will heavily depend on the purpose and requirements of the resulting proactive management strategy. For example, a relatively high FDR may not be an issue if the flow rate can be immediately reduced to prevent further discolouration material from mobilising. Similarly, a long forecasting lead time may not be required if a downstream actuated valve can automatically divert discoloured water (e.g. into a separate reservoir cell or elsewhere) and a very low FDR would be important to not needlessly spam customers who are warned of potential discolouration in advance via text message.

Most of these proactive management strategies should be possible to implement providing a modern SCADA system is already in use. These strategies could not only directly reduce the number of discolouration contacts but could also aid indirectly by reducing discolouration material build up in downstream distribution pipes. This further emphasises the cost benefit in comparison to the expensive alternative of regular invasive mains cleaning and their associated risks.

While the trunk main network used in this paper is arguably simple, constructing and maintaining a hydraulic model for this network would still require some time and capital expense. Furthermore, building and calibrating a hydraulic model alone is not sufficient to accurately predict turbidity, even in a relatively simple configuration system like the trunk main system shown here. The reason for this is that discolouration mobilisation is not limited to any one point in the WDS which then simply travels downstream. The mobilisation of material leading to discolouration can occur anywhere along the flow path and the actual quantity mobilised depends on a number of factors. These factors include the quantity of material available locally (sediments deposited or material available for ‘peeling off’ the pipe wall) and the local hydraulic forces, both of which are changing spatially and temporally throughout the pipe system. As a consequence, continually predicting turbidity at the downstream end of a WDS, even in the case of pipes in series, is more complex than just accounting for travel time at an upstream point in the system. This explains why physically based models such as the PODDS model (Husband and Boxall, 2016) require multiple additional discolouration related parameters, in addition to a calibrated hydraulic model. Furthermore, these discolouration related parameters have to be calibrated with onsite turbidity observations for each trunk main being modelled.

The turbidity forecasting methodology presented here is data-driven and hence does not require a hydraulic model. Only readily available flow and turbidity observation data is needed from the trunk main network to forecast turbidity. Therefore, the methodology has the potential to be transferable to WDS that have suitable meters installed and sufficient historical data already captured. However, the collected historical data will need to include a range of observed turbidity events within the WDS for the data driven models to forecast accurately. Thus while there is conceptually no issue with transferring the methodology to other WDS, the accuracy and leads times that could be achieved in other WDS is currently unknown and may vary significantly between different WDS.

Once a model is calibrated for a specific water system/site, the turbidity forecasting model would require recalibration if that water system's configuration is altered in some significant way (e.g. by doing some rehabilitation type work or otherwise). Depending on the nature of the alteration, it is possible that no human interaction would be required as the model's parameters can be automatically recalibrated. This will however require new observed data to be collected for a period of time which would be dependent on the requisite range of discolouration events taking place in the new system. This, however, is not seen as a major limitation of the methodology proposed here as most water supply and distribution systems are rather well developed and with network configurations that are not subject to regular, significant change.

Finally, due to the difficulty in interpreting the internal workings of data-driven models, little new knowledge about the accumulation and mobilisation processes of discolouration has been generated by using these models, although this remains an avenue for further investigation.

6. Conclusions

This paper presents a turbidity forecasting methodology that, according to our best knowledge, is the first verified discolouration model capable of predicting turbidity continually in a pipe network. The methodology can forecast turbidity above pre-specified thresholds and hence detect the corresponding discoloration event in a real system by using only a data-driven (i.e. non-physically based) approach. However, despite being data-driven, the methodology is generic and therefore should be readily transferable to other water distribution systems.

In a real water system analysed here, relatively accurate forecasts of turbidity events can be made up to approximately 5 h ahead using a classification based approach. This methodology was capable of detecting the mobilisation of discolouration material, estimating if sufficient turbidity will be generated to exceed the pre-specified threshold and approximating the travel time required for that material to reach the downstream turbidity meter.

Future work will focus on further testing and validation of the methodology presented here on additional real systems.

Acknowledgements

The authors are grateful to the Engineering and Physical Sciences Research Council (EPSRC) for providing the financial support as part of the STREAM EngD project and to Julian Collingbourne of South West Water for supplying the data used in this paper.

References

Abarbanel, H.D.I., Brown, R., Sidorowich, J.J., Tsimring, L.S., 1993. The analysis of observed chaotic data in physical systems. Rev. Mod. Phys. 65, 1331–1392. http://dx.doi.org/10.1103/RevModPhys.65.1331.

Abe, Y., Skali-Lami, S., Block, J.-C., Francius, G., 2012. Cohesiveness and hydrodynamic properties of young drinking water biofilms. Water Res. 46, 1155–1166. http://dx.doi.org/10.1016/j.watres.2011.12.013.

Arad, J., Housh, M., Perelman, L., Ostfeld, A., 2013. A dynamic thresholds scheme for contaminant event detection in water distribution systems. Water Res. 47, 1899–1908. http://dx.doi.org/10.1016/j.watres.2013.01.017.

Armand, H., Stoianov, I., Graham, N., 2015. Investigating the impact of sectorized networks on discoloration. Procedia Eng. 119, 407–415.

Armstrong, J.S., 2001. Evaluating forecasting methods. In: Scott Armstrong, J. (Ed.), Principles of Forecasting: A Handbook for Researchers and Practitioners. Kluwer.

ASCE, 1993. Criteria for evaluation of watershed models. J. Irrig. Drain. Eng. 119, 429–442. http://dx.doi.org/10.1061/(ASCE)0733-9437(1993)119:3(429).

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A.F., Nielsen, H., 2000. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16, 412–424. http://dx.doi.org/10.1093/bioinformatics/16.5.412.

Blokker, E.J.M., 2010. Stochastic Water Demand Modelling for a Better Understanding of Hydraulics in Water Distribution Networks. Delft University of Technology, TU Delft.

Blokker, E.J.M., Schaap, P.G., 2015. Particle accumulation rate of drinking water distribution systems determined by incoming turbidity. In: Procedia Eng., Computing and Control for the Water Industry (CCWI2015) Sharing the Best Practice in Water Management, vol. 119, pp. 290–298. http://dx.doi.org/10.1016/j.proeng.2015.08.888.

Blokker, E.J.M., Vreeburg, J.H.G., Van Dijk, J.C., 2009. Simulating residential water demand with a stochastic end-use model. J. Water Resour. Plan. Manag. 136, 19–26.

Blokker, E.J.M., Schaap, P.G., Vreeburg, J.H.G., 2011. Comparing the fouling rate of a drinking water distribution system in two different configurations. In: Savic, D., Kapelan, Z., Butler, D. (Eds.), CCWI 2011 Urban Water Management: Challenges and Opportunities. Centre for Water Systems, University of Exeter, Exeter, pp. 583–588.

Boxall, J.B., Saul, A.J., Gunstead, J.D., Dewis, N., 2003. Regeneration of discolouration in distribution systems. In: ASCE/EWRI/world Water and Environmental Resources Conference, Philadelphia.

Breiman, L., 2001. Random Forests. Mach. Learn 45, 5–32. http://dx.doi.org/10.1023/A:1010933404324.

Cook, D.M., Boxall, J.B., 2011. Discoloration material accumulation in water distribution systems. J. Pipeline Syst. Eng. Pract. 2, 113–122. http://dx.doi.org/10.1061/(ASCE)PS.1949-1204.0000083.

Cook, D.M., Husband, P.S., Boxall, J.B., 2015. Operational management of trunk main discolouration risk. Urban Water J. 13, 382–395. http://dx.doi.org/10.1080/

1573062X.2014.993994.

Dietterich, T.G., 2002. Machine learning for Sequential data: a review. In: Caelli, T., Amin, A., Duin, R.P.W., Ridder, D. de, Kamel, M. (Eds.), Structural, Syntactic, and Statistical Pattern Recognition, Lecture Notes in Computer Science. Presented at the Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR). Springer Berlin Heidelberg, pp. 15–30. http://dx.doi.org/10.1007/3-540-70659-3.

DWI, 2014. Drinking Water Quality Events in 2013 [WWW Document]. URL. http://dwi.defra.gov.uk/about/annual-report/2013/dw-events.pdf.

Frank, R.J., Davey, N., Hunt, S.P., 2001. Time series prediction and neural networks. J. Intell. Robot. Syst. 31, 91–103. http://dx.doi.org/10.1023/A:1012074215150.

Furnass, W.R., Collins, R.P., Husband, P.S., Sharpe, R.L., Mounce, S.R., Boxall, J.B., 2014. Modelling both the continual erosion and regeneration of discolouration material in drinking water distribution systems. Water Sci. Technol. Water Supply 14, 81. http://dx.doi.org/10.2166/ws.2013.176.

Gershenfeld, N.A., Weigend, A.S., et al., 1993. The Future of Time Series: Learning and Understanding.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Mach. Learn 63, 3–42. http://dx.doi.org/10.1007/s10994-006-6226-1.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. In: In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS'10). Society for Artificial Intelligence and Statistics.

Green, I.R.A., Stephenson, D., 1986. Criteria for comparison of single event models. Hydrol. Sci. J. 31, 395–411. http://dx.doi.org/10.1080/02626668609491056.

Husband, P.S., Boxall, J.B., 2011. Asset deterioration and discolouration in water distribution systems. Water Res. 45, 113–124. http://dx.doi.org/10.1016/j.watres.2010.08.021.

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, 4309–4318. http://dx.doi.org/10.1016/j.watres.2008.07.026.

Husband, P.S., Whitehead, J., Boxall, J.B., 2010a. The role of trunk mains in discolouration. Proc. Inst. Civ. Eng. - Water Manag. 163, 397–406. http://dx.doi.org/10.1680/wama.900063.

Husband, S., Boxall, J., 2016. Understanding and managing discolouration risk in trunk mains. Water Res. 107, 127–140.

Husband, S., Williams, R., Boxall, J.B., 2010b. Discolouration risk management for trunk mains. In: Proceedings of the Water Distribution System Analysis Conference, Tucson, Arizona, ASCE Conf. Proc., p. 425.

Hyndman, R.J., Koehler, A.B., 2006. Another look at measures of forecast accuracy. Int. J. Forecast 22, 679–688. http://dx.doi.org/10.1016/j.ijforecast.2006.03.001.

Kingma, D., Ba, J., 2014. Adam: a method for stochastic optimization. In: ArXiv14126980 Cs, 3rd International Conference for Learning Representations.

Leeder, A., Mounce, S.R., Boxall, J.B., 2012. Analysis of multi-parameter water quality data using event detection software on laboratory simulated events. In: WDSA 2012: 14th Water Distribution Systems Analysis Conference, 24-27 September 2012 in Adelaide, South Australia. Engineers Australia, p. 1018.

Machell, J., Boxall, J., Saul, A., Bramley, D., 2009. Improved representation of water age in distribution networks to inform water quality. J. Water Resour. Plan. Manag. 135, 382–391.

Machell, J., Mounce, S.R., Boxall, J.B., 2010. Online modelling of water distribution systems: a UK case study. Drink. Water Eng. Sci. 3, 21–27.

Machell, J., Mounce, S.R., Farley, B., Boxall, J.B., 2014. Online data processing for proactive UK water distribution network operation. Drink. Water Eng. Sci. 7, 23–33. http://dx.doi.org/10.5194/dwes-7-23-2014.

Martinec, J., Rango, A., 1989. Merits of Statistical Criteria for the Performance of Hydrological Models.

Meyers, G., Kapelan, Z., Keedwell, E., Randall-Smith, M., 2016. Short-term forecasting of turbidity in a UK water distribution system. In: Procedia Eng., 12th International Conference on Hydroinformatics (HIC 2016) - Smart Water for the Future, vol. 154, pp. 1140–1147. http://dx.doi.org/10.1016/j.proeng.2016.07.534.

OFWAT, 2009. Serviceability Outputs for PR09 Final Determinations.

Prince, R., Goulter, I., Ryan, G., 2001. Relationship between velocity profiles and turbidity problems in distribution systems. In: World Water and Environmental Resources Congress, Orlando, Florida.

Rhodes, C., Morari, M., 1997. The false nearest neighbors algorithm: an overview. Comput. Chem. Eng. 21, S1149–S1154.

Skipworth, P.J., Machell, J., Saul, A.J., 2002. Empirical travel time estimation in a distribution network. In: Proceedings of the Institution of Civil Engineers-Water and Maritime Engineering. Thomas Telford Ltd., pp. 41–49

Slaats, P.G.G., Rosenthal, L.P.M., Siegers, W.G., van den Boomen, M., Beuken, R.H.S., Vreeburg, J.H.G., 2003. Processes Involved in the Generation of Discolored Water. AWWA Research Foundation and Kiwa Water Research.

Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Stat. Comput. 14, 199–222. http://dx.doi.org/10.1023/B: STCO.0000035301.49549.88.

Tao, T., Huang, H., Li, F., Xin, K., 2014. Burst detection using an artificial immune network in water-distribution systems. J. Water Resour. Plan. Manag. 140, 04014027. http://dx.doi.org/10.1061/(ASCE)WR.1943-5452.0000405.

Van Thienen, P., Vreeburg, J.H.G., Blokker, E.J.M., 2011. Radial transport processes as a precursor to particle deposition in drinking water distribution systems. Water Res. 45, 1807–1817.

Vreeburg, I.J.H.G., Boxall, D.J.B., 2007. Discolouration in potable water distribution systems: a review. Water Res. 41, 519–529. http://dx.doi.org/10.1016/j.watres.2006.09.028.

Vreeburg, J.H.G., 2007. Discolouration in Drinking Water Systems: a Particular Approach. Delft University of Technology, TU Delft.

Vreeburg, J.H.G., Schaap, P.G., Van Dijk, J.C., 2005. Particles in the drinking water system: from source to discolouration. Water Sci. Technol. Water Supply 4, 431–438.

Vreeburg, J.H.G., Schippers, D., Verberk, J.Q.J.C., van Dijk, J.C., 2008. Impact of particles on sediment accumulation in a drinking water distribution system. Water Res. 42, 4233–4242. http://dx.doi.org/10.1016/j.watres.2008.05.024.