EPiC Series in Engineering
Volume 3, 2018, Pages 1030–1039
HIC 2018. 13th International Conference on Hydroinformatics

Predicting turbidity in water distribution trunk mains using nonlinear autoregressive exogenous artificial neural networks

Ehsan Kazemi[1], Stephen Mounce[1], Stewart Husband[1], Joby Boxall[1]

[1] Pennine Water Group, Department of Civil and Structural Engineering, University of Sheffield, Sheffield, S1 3JD, UK

e.kazemi@sheffield.ac.uk

Abstract

A nonlinear autoregressive exogenous artificial neural network model was developed to predict turbidity response in two different trunk mains with measured flow and turbidity data. Models were initially established to prepare the data and automatically select the appropriate events for model training. Then, an autoregressive exogenous network model was developed and applied to predict turbidity responses based on past events in the time series. A per site continual data driven calculation of turbidity event risk was included as an additional input to capture the effect of temporal distance between the selected events as well as increasing the accuracy of the predictions. The calculated normalised mean square error and mean absolute error showed that the developed model combined with the data preparation and pre-processing models provides good regressions on a future event with a period of 7 to 10 hours for a multi-step ahead prediction. Furthermore, the result of the autoregressive exogenous network was compared with the output of a feed-forward network where the former significantly outperformed the latter (R value of approximately 0.97 compared to 0.66).

Keywords: Machine learning, ANN, NARX, Trunk mains, Turbidity, Water distribution systems

1 Introduction

Although significant effort has been dedicated to solving the discolouration problem in Water Distribution Systems (WDS), it remains a challenging task for water service providers (WSP) to attain regulatory demands and maintain customer satisfaction. In 2016, WSP in the UK were contacted 12.2

times per 10,000 people due to discoloured water, the greatest source of customer dis-satisfaction (Drinking Water Inspectorate, 2016).

The primary cause of discolouration events in WDS has been shown to be changes in network hydraulic loading mobilising discolouration material that is constantly accumulating across the full pipe circumference and with a shear strength conditioned by prevailing hydraulics (Husband and Boxall, 2011). If flow can be associated with a discolouration response, and the discolouration therefore predicted in advance, smart alarms facilitating pro-active management of WDS would be possible. This could lead to a reduction in operational costs associated with the present re-active management strategies adopted by water companies (typically comprising of expensive trunk mains cleaning programmes).

Trunk mains are critical assets that are of particular concern to WSP with regards to discolouration risk as they lie upstream of large numbers of customers and may typically only experience low conditioning shear stresses, consequently there may be much weakly-bound material adhered to their large internal surfaces. Understanding discolouration processes in trunk mains is difficult due to the fact that these processes can only be observed directly through disturbing the system which is usually undesirable due to potential mobilisation risks (Husband et al., 2010).

Discolouration in WDS is a highly non-linear multivariate problem where various factors such as hydraulic condition, water quality, pipe material and age need to be considered (Husband and Boxall, 2011; Machell and Boxall, 2014). Therefore, empirical and hydraulic models developed for predicting discolouration are often unable to describe such a complex system and are not applicable for WDS with different network characteristics. Data-driven models such as Artificial Neural Networks (ANN), as an alternative to hydraulic models, may be able to deal with the complex nature of the problem and be applied to different system conditions where sufficient data is available to appropriately calibrate and validate the model.

In recent work, data-driven modelling has explored predicting discolouration in WDS (Meyers et al., 2017a and 2017b) in which three data-driven approaches (ANN, Random Forests and Support Vector Machines) were applied to develop a turbidity forecasting system. The models were developed based on two different scenarios: regression-based and classification-based turbidity forecasting. In the former, the actual value of turbidity was directly predicted from past values of flow and turbidity, while in the latter, the turbidity in the forecast horizon was classified as being above a pre-specified threshold. In the application of the models to their data following single filtering, it was shown regression-based models could predict turbidity up to 20 minutes ahead, while the classification-based models could forecast turbidity for up to 5 hours ahead. It was found that Random Forests based turbidity forecast models performed best for this approach but that ANNs also had good turbidity forecasting ability.

It this paper, we investigate the application of ANNs for turbidity prediction, as the main cause of discolouration problem in WDS, based on the regression-based models as these models provide more useful information compared to the classification-based ones. Moreover, in contrast to the regression-based model in Meyers et al., 2017a, where only the value of turbidity at n hours later is forecast, our aim is to predict the distribution of one future turbidity event (which may exceed several hours) from the past flow and turbidity events, i.e. predicting turbidity sequence from the current time to n hours ahead (multi-step ahead prediction). For this purpose, an ANN model is constructed based on the nonlinear autoregressive exogenous (NARX) network (Lin et al., 1996) which is suitable for time-series multi-step predictions. Before that, the data needs to be prepared for ANN analysis. Hence, the data preparation models, such as event detection models are introduced and applied first to extract the required events. In addition, a 'turbidity event risk' metric is calculated with two scenarios and included as input in the ANN model to capture the effect of temporal distance between the extracted events. Prediction performance is tested for two real world cases and compared with and without this additional input, and also with the output from a Feed-Forward network.

2 Datasets and data preparation models

In this section, the data preparation techniques will be discussed, and then applied to two different data-sets from the centre and north UK, which have been employed in the present study for the ANN analysis. These datasets are selected due to their different characteristics.

The *central* data is a flow and turbidity time series from a site with a 5.4 km, 500 mm ductile iron trunk main that supplied a service reservoir directly from the water treatment works with consistent 0.1 NTU output and flow control by variable speed pumps. With no take-offs along the length, controlled changes in flow were planned and implemented to create low-level discolouration events well within UK 4 NTU regulatory values (Husband and Boxall, 2017). The *northern* data set is from a gravity fed 6.4 km, 300 mm partially lined cast iron main with a ferric coagulated upland service reservoir source that was part of an operational hydraulic discolouration study (Sunny et al., 2017).

2.1 Filtering turbidity anomalies

There are several high frequency spikes in the turbidity time series, especially in the *central* data which are outliers and need to be treated. These are typically associated with reflective properties of particles used for measuring turbidity, possible air bubbles or resulting from instrument maintenance necessary to clean optics as discolouration material accumulates. These spikes occur in very short time periods and without any flow association. As the dataset is very large, manual removal of these anomalies is not practical. A searching algorithm based on the gradient of the turbidity time series is therefore applied to filter them. Changes in the gradient is computed from a data point to the next, and if it is greater than a threshold, the data point is removed, i.e. it will not be considered for next steps of analysis.

2.2 Averaging and smoothing the data

A simple averaging algorithm is applied to adjust the resolution to a desirable one for ANN training. The original flow and turbidity data was logged every 5 and 15 minutes, respectively, in the *central* and *northern* data. In the current analysis, the resolution of the *central* data is also resampled to 15 minutes by arithmetic averaging. One may use a weighted arithmetic average for this purpose.

The data still requires smoothing for undesirable noise to be removed. A cubic spline function is used here to smooth the time series (Equation 1). The function works based on a weighting kernel and a smoothing length which defines the influence domain of the smoothing (the influence domain is twice the smoothing length).

$$W(t-t',h_\Upsilon)=\lambda_\Upsilon\begin{cases}2/3-q_\Upsilon^2+q_\Upsilon^3/2 & 0\le q_\Upsilon<1\\ \left(2-q_\Upsilon^3\right)/6 & 1\le q_\Upsilon<2\\ 0 & q_\Upsilon>2\end{cases} \quad (1)$$

where Υ refers to the influence domain associated to a certain data point over which the data is smoothed; t and t' refer to the data point at which the smoothing is carried out and its neighbouring data points within the influence domain, respectively; $q_\Upsilon=|t-t'|/h_\Upsilon$; and $\lambda_\Upsilon=1/h_\Upsilon$.

By adjusting the smoothing length, the roughness and trend preservation of the smoothed time series can be controlled. Here, it is set to $1.2\Delta t$, i.e. $h_\Upsilon=18$ minutes. Figure 1 shows a part of the *central* data during an operationally controlled flow event and the resulting turbidity response occurring over two days pre- and post-smoothing.

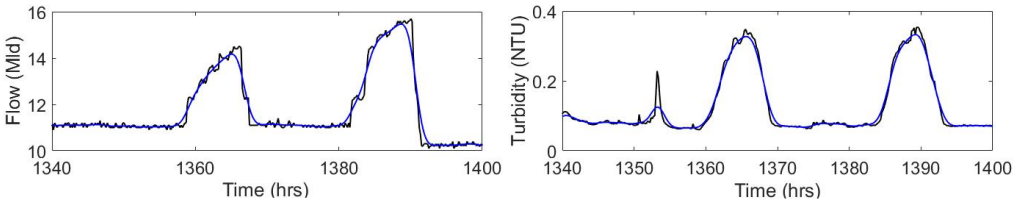

Figure 1: The original data (black) and the smoothed data (blue).

2.3 Turbidity event detection model

The turbidity events which are sufficiently high for analysis need to be extracted from the data. Therefore, an automated system was designed, as described in the following, to find those events in the time series.

Firstly, turbidity peaks higher than a threshold are detected. They are the maximum values of the events. Then, the start and end points of the events need to be determined. This is performed based on the change in the gradient of the time series in a specified period around the peak of the event. Then the 'base value' of the event which is 'the pre-event value of turbidity which starts increasing at the beginning of the event and then decreasing again to that value at the end of the event' is computed using the smoothing function with a large smoothing length without taking the detected events into account. This value can be used later to calculate the rise in the turbidity as the difference between the maximum and base values, i.e. removing the effect of seasonality in the data. Finally, the detected turbidity events, i.e. their maximum, start and end points are updated according to the calculated base line. Figure 2 shows two examples of turbidity event detection in the data of *Central* (a) and *Northern* (b).

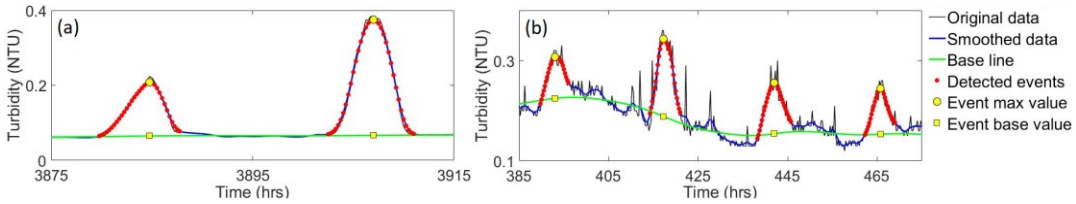

Figure 2: Turbidity event detection examples: (a) *Central*, (b) *Northern*.

2.4 Flow event detection model

Once turbidity events are detected, flow events associated with them are also detected by searching over a certain period of time (several hours here) previous to each detected turbidity event. Firstly, the flow peak in that period is located. Then, from there, the change in the flow in both directions is investigated within a time window based on the gradient of the time series. The flow variations need to follow certain conditions to be considered as a flow event. If not, no flow event is detected and therefore the associated turbidity event is also removed based on the assumption that there must have been no flow association at that particular turbidity event so that it will not be used in the ANN analysis.

Then, the 'base value' for flow events are calculated by averaging past flow values over a short period of time. The rise in the flow can then be calculated as the difference between the peak and base values. Figure 3 shows an example of detected flow events and associated turbidity events.

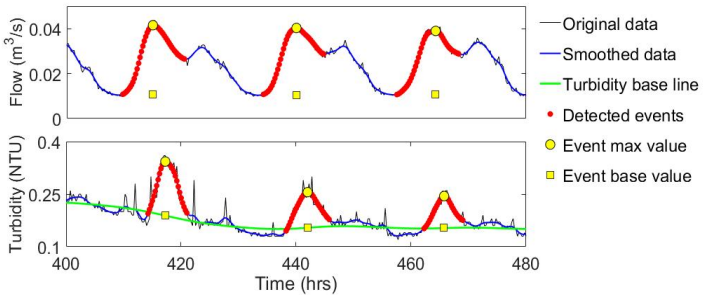

Figure 3: An example of detected flow events with their associated turbidity events. (Data: *Northern*).

The turbidity and flow event detection systems have several coefficients and constants that define which and what type of events are to be detected; for example, high or low events (based on a threshold), or the events with steeper or milder profile gradients.

3 ANN analysis

The aim of this section is to predict a future turbidity event using the past events of flow and turbidity. In this analysis, the target is turbidity and the input is flow, as the turbidity events are regarded to be flow associated.

Detected events are extracted and used for constructing input and target matrices. However, the effect of temporal distance between the events is missing while it can have an important effect on a future event. For instance, if there has been a long period with no turbidity event, the risk of the occurrence of a new turbidity event is higher compared with when there has just been a recent significant event. Hence, an extra input, namely, the turbidity event risk parameter (E_p) is introduced to resolve this deficiency. Two scenarios are explored for the determination of this parameter, as now described in the following.

3.1 Determination of the turbidity event risk

Scenario #1

It is assumed that 1) material deposits continually on the pipe wall, and releases from it suddenly during an event; and that 2) the total build up and release volumes (over a certain time period) are equal. Based on the first assumption, the risk of an event increases gradually with a certain slope, which is defined here as 'deposition per time step (m_d)'. This drops suddenly at the occurrence of a turbidity event. According to the second assumption, m_d is equal to total area below turbidity events in the time series (red lines in Figure 4, see Figure 2 for better visibility) divided by the total time period. Therefore, the following equation is used to estimate the turbidity event risk E_p.

$$E_p\big|_t = E_p\big|_{t-\Delta t} + m_d \Delta t - \delta_t R_t \Delta t \tag{2}$$

where t and t-Δt denote the time step (data point) where E_p is calculated, and its previous time step, respectively; Δt is the time interval size; R_t is the increase in the turbidity volume from the base value at time t (during an event); and δ_t is equal to one if the data point at time t is within a turbidity event, and is equal to zero, if it is not. The second term on the right-hand side of the equation describes the gradual increase in the E_p profile with the slope of m_d, while the third term introduces the sudden reduction during an event. Figure 4 (b) shows the distribution of E_p calculated by Equation (2) for the

central data, shifted up as to have no negative values and then normalised between 0 and 1. An initial value is required for E_p. However, it is not easy to set that value due to the lack of knowledge on the condition of the main before the data was measured. Here, it was initially set to zero, which led to a value greater than zero after applying the shift in the E_p profile.

A limitation with this technique is the difficulty in the estimation of m_d, which is a site-specific parameter. Based on the current calculation, m_d is estimated as the total area below the detected turbidity events in the time series divided by the total time period. Therefore, the number of events included in the calculation or the time period over which the calculation is carried out, influences the magnitude of m_d. In other words, using different numbers of events, or different time periods in the analysis, gives different values for m_d. Hence, scenario #2 is also introduced as an alternative for the estimation of E_p.

Scenario #2

Here, the turbidity event risk parameter E_p is estimated locally, based on the average of a few past events in the time series, so that there will be no need for an extra parameter like m_d. To estimate E_p at time t, an imaginary event which is a weighted average of a few past events is set at a certain temporal distance to the current time t, and then E_p is estimated as the temporal distance divided by the magnitude of the imaginary event, as follows in Equation (3).

$$E_p\big|_t = \frac{L_{EM}}{H_{EM}}, \text{ where } H_{EM} = \frac{1}{N}\sum_{i=1}^{N} H_{Ei} \text{ and } L_{EM} = \sum_{i=1}^{N}\left(L_{Ei}H_{Ei}\right)\Big/\sum_{i=1}^{N} H_{Ei} \tag{3}$$

where H_{EM} and L_{EM} are the magnitude of the imaginary event and the temporal distance between t and that event, respectively. H_{Ei} and L_{Ei} are the magnitude of the past events (i = 1,2,…,N) that are used to estimate the imaginary event, and the temporal distance between t and those event, respectively. i denotes the past events used for this calculation and N is the total number of them.

According to this equation, when the temporal distance to past events is longer or the magnitudes of the past events are lower, the risk of a new event is higher, and vice versa. In fact, the effect of several past events is captured with a higher impact from the larger events. N = 1 means that the risk of a new event at time t is affected by only the past event, right before t. However, more than one event usually affects the risk of a new event. An example is when there is a small event in a certain temporal distance to time t and there is also a large event just before the small event. In this situation, considering only the small event in the estimation of E_p leads to a high risk, while the risk is not actually high due to the larger event. The choice of N depends on the data. Figure 4 (c) shows the estimated E_p for the *central data*, by using N = 4 and the initial value of zero, normalised between 0 and 1. It can be seen that both scenarios provide relatively similar profiles of E_p, while scenario #2 is simpler and also more robust with regards to the independence to m_d.

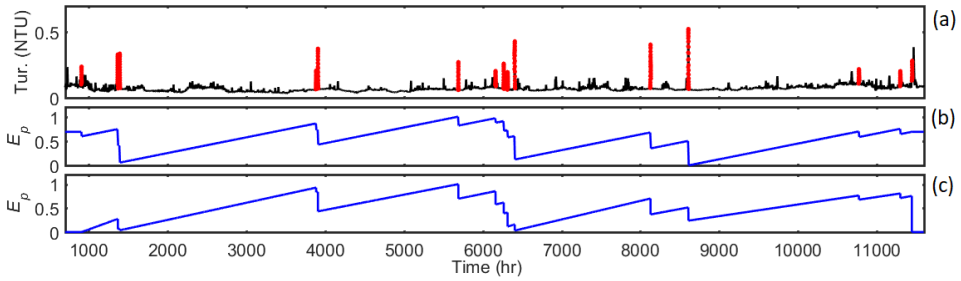

Figure 4: Turbidity event risk E_p. a) Turbidity time series (black) and detected events (red); b) E_p (hr.NTU) estimated by scenario #1; and c) E_p (hr/NTU) estimated by scenario #2.

3.2 NARX: results and discussion

A multi-step prediction ANN model is constructed based on the NARX network in MATLAB. At a certain time, the turbidity is predicted several time-steps ahead (forecast horizon) using the values of flow, E_p and turbidity in the past events. First, an open-loop version of the NARX network is used to train the model, then the network is closed and applied for turbidity prediction. One hidden layer with the size of 10 is employed and the input and feedback delays are set to 3 (derived empirically).

The inputs are the detected flow events as well as E_p at the beginning of the turbidity events; and the target is the turbidity events. note that that the difference between the magnitude of flow/turbidity during an event and the base value of that event, i.e. flow/turbidity rise during an event, is considered rather than the absolute values of flow and turbidity, in order to remove the effect of data seasonality.

Figure 5 represents the results of the model in comparison with the measured data for the *central* and *northern* cases (Figures 5 (a) and (b), respectively) for the prediction of a future event (forecast horizon of about 7-10 hrs) using 10 past events, with using E_p as input based on scenarios #1 and #2. The result without using E_p is also presented for comparison. According to the figure, including E_p enhances the predictions. In the results related to the *central* data, the Normalised Mean Square Error (NMSE) of the closed-loop training for the cases without using E_p, scenario #1, and scenario #2 are 0.15, 0.115 and 0.056, respectively; and the Mean Absolute Error (MAE) of the predicted event is 0.107, 0.057, and 0.042, respectively. For the *northern* case, the NMSE of the NARX training for the cases without using E_p, and scenarios #1 and #2 are 0.035, 0.061 and 0.058, respectively; and the MAE of the predicted event is 0.028, 0.026, and 0.016, respectively. The performance of the model in the analysis of the *northern* data seems more accurate with lower errors. This can be explained by this data being more uniform with less variation in the size of the events compared to the *central* data.

It is noted that the inclusion of E_p in the inputs does not necessarily lead to more accurate prediction for all events in the data. For instance, in Figure 6 (a), the result of the model for another part of the *northern* data (to predict a different event) without using E_p is presented and compared with the results of the model when scenarios #1 and #2 are used to estimate E_p. As can be seen, the inclusion of E_p leads to an overestimation of this particular event. The reason may be related to the probable overestimation of the calculated E_p at that particular event as shown in Figure 6 (b). According to the figure, the estimated E_p at the beginning of the predicted event with both scenarios #1 and #2 is about 0.86 (marked with circles on the profiles) which is a relatively high risk, while the event is relatively small. This means that it is not always possible to have a precise estimation of the risk of all turbidity events in the time series.

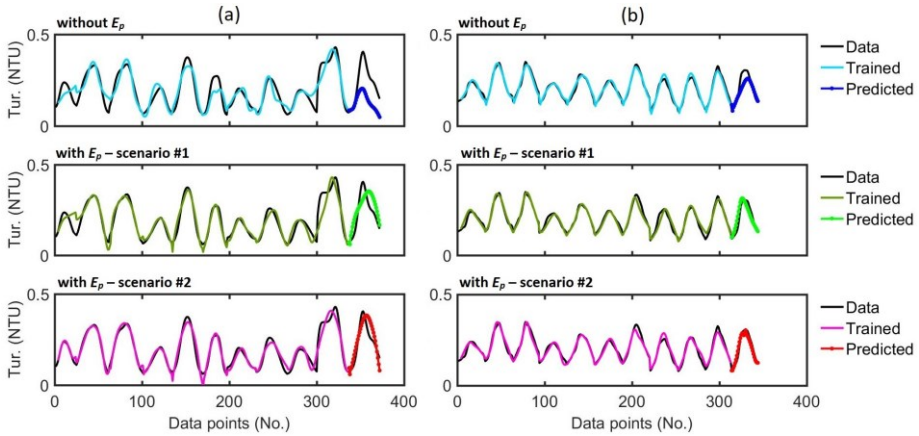

Figure 5: Prediction of a future event using 10 past events, without using E_p and with scenarios #1 and #2 for the a) *central* and b) *northern* cases.

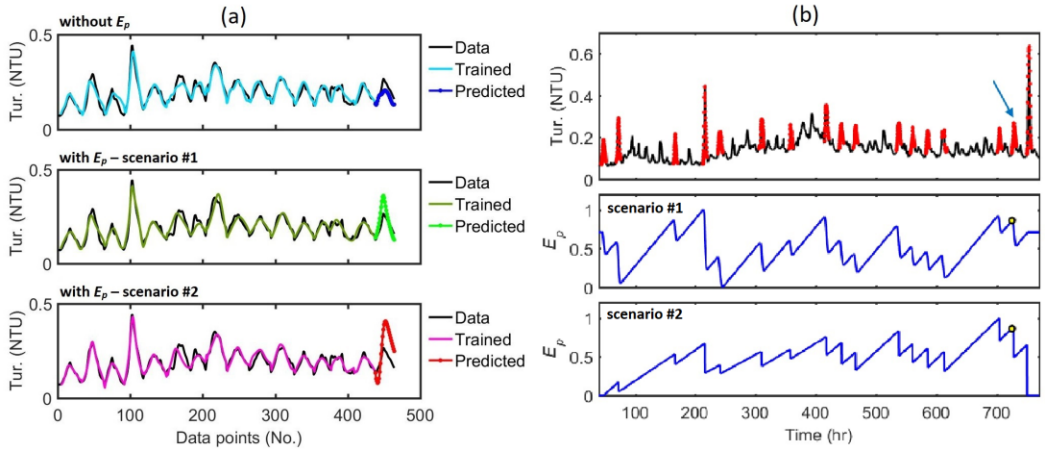

Figure 6: A part of the *northern* data: a) prediction of a future event without using E_p, and with using it based on scenarios #1and #2; and b) the predicted event shown in the turbidity time series and E_p calculated using scenarios #1 and #2.

3.3 NARX network vs. Feed-Forward network

In this section, the analysis of the *northern* data (presented in Figure 5 (b), scenario #2) is repeated with a Feed-Forward ANN network in order to assess the performance of the NARX model in comparison with the Feed-Forward network for the present case study. The result is shown in Figure 7. The NMSE of the trained data (past events) is 0.559 and 0.058 for the Feed-Forward and NARX analysis, respectively; and the MAE of the future (predicted) event is estimated as 0.04 and 0.016 NTU, respectively. Figure 8 presents train and test regressions of these two ANN models with an R value of about 0.66 for the Feed-Forward analysis and 0.97 for the NARX model. This information reveals that the NARX network significantly outperforms the Feed-Forward network for this type of multi-step ahead prediction.

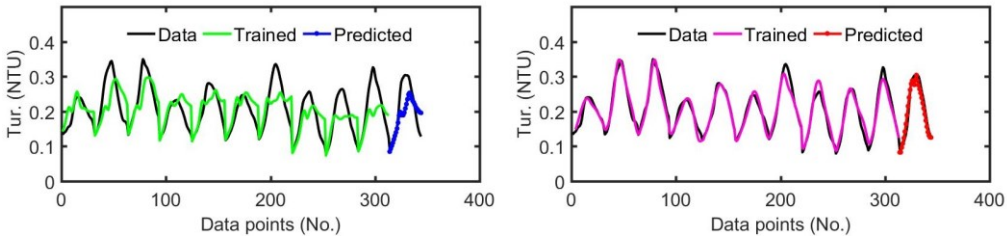

Figure 7 Result of a Feed-Forward network (left) and NARX network (right) analysis for a part of the *northern* data.

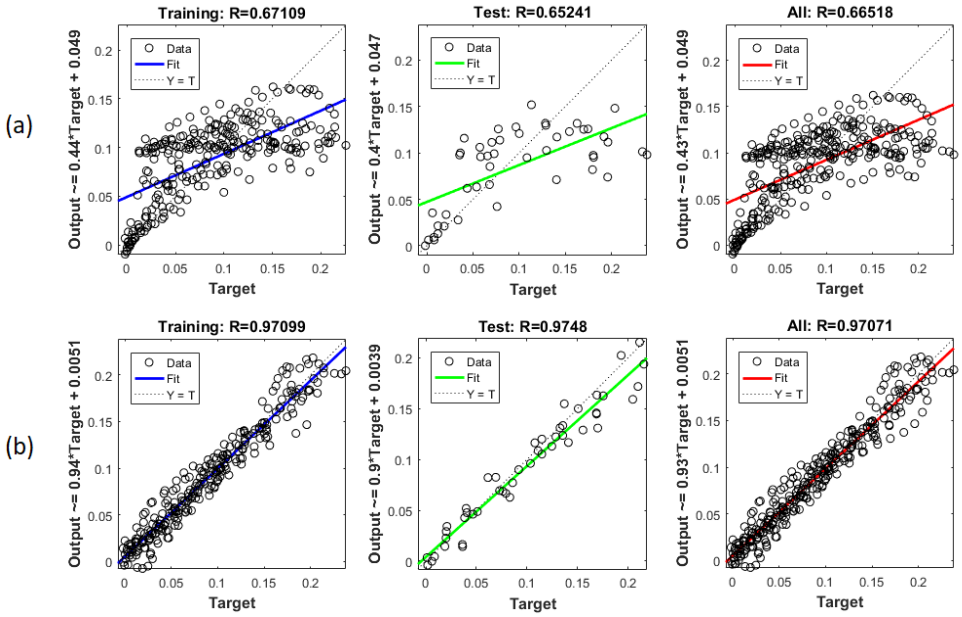

Figure 8 Regressions of train and test of the data with a) Feed-Forward network and b) NARX network.

4 Further work

In future work, the developed models will be trained, calibrated and validated using a larger number of datasets (incorporating many more events with different characteristics), to examine the trade-offs of data availability, variability, accuracy and generalisation. The model includes a flow warning system that could be useful for practical applications. The warning system detects if a considerable change takes place in the flow. At that specific time, the trained model can be applied to forecast a potential future turbidity event and whether it exceeds a particular threshold. In addition, the model is not computational demanding hence ideal as a real time smart alarm system, i.e. can run updated series at regular intervals to estimate a new alert for exceedance thresholds.

5 Conclusions

The data preparation models were introduced to tailor the data for ANN analysis. The presented automated turbidity and flow event detection models extracted the required events for ANN training. Then, the NARX network was employed to develop a time series prediction model for predicting one future event from the past detected events. Due to the discontinuity in the extracted events, an extra input, E_p, was introduced to capture the effect of temporal distance between the events. Key conclusions are summarised as follows;

- The estimated errors (NMSE and MAE) show that the NARX model combined with the models developed for data preparation provides a good fitting to the detected events.
- Including E_p as an input leads to not only avoidance of incorrect mathematical representation of the physical problem, but also an enhancement in the predictions.

- Both scenarios introduced for E_p estimation perform very similarly. However, scenario #2 is more robust and easily implemented, though estimation of the risk of a turbidity event is still challenging due to the complexity of the problem, resulting in inaccurate predictions of some events.
- The NARX network performs significantly better than the Feed-Forward network for the present problem.
- The present model provides the (multi-step ahead) distribution of a future turbidity event rather than a single value (e.g. maximum or mean of an event) or a classification of the event. An advantage of the model is that it is capable of predicting the distribution of one future event with a period of several hours (7 to 10 hours in the present analysis) and could be used to predict several future events if there is a sufficient number of various events available for training.

Acknowledgment

The authors would like to thank the PODDS and the UK water companies for supporting the ongoing programme of discolouration research at the University of Sheffield and for data provision and permission to publish the details included herein.

Reference

Drinking Water Inspectorate; England and Wales. Jan 2016 - Dec 2016, http://www.dwi.gov.uk/about/annual-report/2016/

Husband, P.S., Boxall, J.B. (2011). Asset deterioration and discolouration in water distribution systems. Water Res. 45: 113–124.

Husband, P.S., Whitehead, J., Boxall, J.B. (2010). The role of trunk mains in discolouration. Proceedings of the ICE - Water Management, 163: 397–406.

Machell, J., Boxall, J.B. (2014). Modeling and field work to investigate the relationship between age and quality of tapwater. J. Water Res. Plan. Manag., 140: 431–439.

Meyers, G., Kapelan, Z., Keedwell, E. (2017a). Short-term forecasting of turbidity in trunk main networks. Water Res., 124: 67–76, ISSN 0043-1354.

Meyers, G., Kapelan, Z., Keedwell, E. (2017b). Data-driven approach to short-term forecasting of turbidity in a trunk main network. CCWI 2017, Sheffield, Sep. 2017.

Lin, T., Horne, B.G., Tino, P., Giles, C.L. (1996). Learning long-term dependencies in NARX recurrent neural networks. IEEE Trans. Neural Netw., 7: 1329-1351.

Husband, P.S., Boxall, J.B. (2017). Real time modelling of pipe wall material and managing the discolouration risk in distribution systems. Water Quality Technology Conference, USA, Nov. 2017.

Sunny, I., Husband, P.S., Moore, G., Drake, N., Mckenzie, K., Boxall, J.B. (2017). Discolouration Risk Management and Chlorine Wall Decay. CCWI 2017, Sheffield, Sep. 2017.