Contents lists available at ScienceDirect

Journal of Hydrology

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

Highlights

Journal of Hydrology xxx (xxxxx) xxx

Quantifying the natural flood management potential of leaky dams in upland catchments, Part I: A data-based modelling approach

Z.R. van Leeuwen*, M.J. Klaar, M.W. Smith, L.E. Brown

- Limited evidence for natural flood management due to short, uncertain data
- Time series modelling simulates flood peaks (±2 cm) with 95% accuracy using BACI data
- Accuracy is adequate for assessing leaky dam impacts on flood peak magnitude
- Proof of concept valuable for applying to other environments and interventions

Graphical abstract and Research highlights will be displayed in online search result lists, the online contents list and the online article, but will not appear in the article PDF file or print unless it is mentioned in the journal specific style requirement. They are displayed in the proof pdf for review purpose only.

Contents lists available at ScienceDirect

Journal of Hydrology

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

Research papers

Quantifying the natural flood management potential of leaky dams in upland catchments, Part I: A data-based modelling approach

Z.R. van Leeuwen *, M.J. Klaar, M.W. Smith, L.E. Brown

School of Geography & water@leeds, University of Leeds, Woodhouse Lane, Leeds, LS2 9JT, United Kingdom

ARTICLE INFO

This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Szilágyi József, Associate Editor.

Keywords:
Empirical analysis
Transfer function model
ARIMA
Hydrograph analysis
BACI monitoring

ABSTRACT

A novel application of data-based time series methods is proposed in this study to help overcome barriers to quantifying the impacts of Natural Flood Management measures from hydrological timeseries data which hitherto have prevented accurate assessment of the effectiveness of interventions. To demonstrate the value of this method, a transfer function noise model was fitted to stage data from a three year before-after-control-impact style monitoring study of leaky dams in an upland catchment in North Yorkshire, England. Using the data-based time series method, uncertainties associated with stage data were overcome. The models were able to simulate the peaks of flood events on one stream to within ± 2 cm accuracy for 95% of events recorded during the baseline monitoring period. These simulations are used in this study's companion paper to quantify leaky dam impacts on flood peak magnitude. The level of accuracy achieved in this study provides proof of concept for application of the approach to data from other environments and natural flood management interventions, which is crucial if natural flood management is to be used as a mainstream flood risk management measure.

1. Introduction

Natural Flood Management (NFM) is an increasingly popular approach to flood risk management (Grabowski et al., 2019) which aims to slow and store flood water in the landscape by restoring natural hydrological and geomorphological processes (Forbes et al., 2015). NFM measures provide multiple benefits alongside flood risk management and include measures such as afforestation, soil management practices, and leaky dams (Lane, 2017). Due to difficulties in obtaining precise hydrological data both before and after NFM introduction, a lack of robust, empirical evidence describing the effectiveness of NFM measures at reducing downstream flood risk currently undermines confidence in its efficacy and limits its adoption (Bark et al., 2021; Waylen et al., 2018; Wingfield et al., 2019). Its implementation is, nevertheless, encouraged across Europe (Commission of the European Communities, 2009), forms part of flood risk management policy in the UK (Environment Agency, 2010; Forbes et al., 2015), is incentivised in agricultural policy (Defra et al., 2016), and is likely to form part of future approaches to environmental land management in England and Wales (Klaar et al., 2020). To increase confidence in the implementation of these approaches for flood risk management it is essential to build a robust evidence base of the impacts of NFM measures (Cook et al., 2016; Dadson et al., 2017; Ellis et al., 2021; Iacob et al., 2017; Lane, 2017; Beven et al., 2022).

Applying a Before After Control Impact (BACI) monitoring methodology to the collection of hydrological data has been identified as a way to gather the evidence required to 'mainstream' NFM approaches (Ellis et al., 2021). However, factors which contribute to uncertainty in quantifying the impacts of NFM, such as difficulties in isolating impacts of one type of NFM measure, capturing changes in effectiveness with event magnitude, insufficiently long monitoring timescales, and complexities of context and scale (Connelly et al., 2020) are not necessarily overcome by using a BACI approach. There are two main challenges with applying the BACI approach to the monitoring of NFM impacts on flood peak magnitude. Firstly, the opportunistic nature of NFM projects means there is a lack of lead time to collect long enough periods of baseline data to account for the stochastic nature of floods (Connelly et al., 2020; Ellis et al., 2021). For example, despite best efforts, a paucity of high flow events observed during the monitoring period hampered the empirical quantification of NFM impacts of two of the three government funded pilot projects initiated in 2009 (Nisbet et al., 2015; National Trust, 2015) and continues to affect the collection of evidence from the £15 million worth of NFM projects funded by the UK government in 2017 (Environment Agency, 2019). Furthermore, even when sufficient flood events are recorded, the signal of an intervention is often masked by the high levels of uncertainty typical of hydrological data (Black et al., 2021; Gebrehiwot et al., 2019; Lane, 2017; Ellis

* Corresponding author.
E-mail address: zoravanleeuwen@gmail.com (Z.R. van Leeuwen).

https://doi.org/10.1016/j.jhydrol.2023.130448
Received 12 May 2023; Received in revised form 25 September 2023; Accepted 17 October 2023

et al., 2021). For example, although impacts on flood peak timing were quantified, high levels of uncertainty in hydrological data prevented the quantification of leaky dam impacts on flood peak magnitude in a multi-scale experiment in Scotland, despite having a 9 year long, comprehensive monitoring strategy (Black et al., 2021).

As a result of the difficulties associated with collecting hydrological data there are few empirical studies which have successfully quantified the impacts of leaky dams in upland watercourses on downstream flood risk. Recent reviews of both the academic and 'grey' literature identified few studies in which the impact of instream wood on the flood hydrograph was successfully quantified (Addy and Wilkinson, 2019; Burgess-Gamble et al., 2017). In these studies the problems presented by the stochastic nature of flood events were avoided by generating artificial reservoir releases to emulate flood events (Keys et al., 2018; Wenzel et al., 2014). The feasibility of such an approach is, necessarily, limited to locations in which artificial flood releases can be generated.

The majority of attempts to overcome the difficulties associated with empirically quantifying leaky dam impacts have, therefore, been made using numerical fluvial hydraulic and hydrological models (Burgess-Gamble et al., 2017). These types of models rely on *a priori* assumptions about the physical processes governing the impacts of leaky dams. These processes are poorly understood (Dixon et al., 2016; Lane, 2017) and the lack of quantitative validation data means the representation of leaky dams in numerical hydraulic and hydrological models remains heuristic (Addy and Wilkinson, 2019).

An alternative is to use empirical BACI data supported by an inductive, 'top-down' data-based time series modelling approach, which minimises the need for *a priori* assumptions. This approach, described by Beven (2001) as 'doing hydrology backwards' uses statistical time series modelling methods to infer model structure and parameter values from empirical data (Young, 2003). Once the model structure and parameters have been determined from the data the model represents the dynamical properties of the system and can therefore be used to make predictions of the values of the output series and its uncertainty for unobserved periods (von Asmuth et al., 2002). This approach has been used to quantify the thermal response of streams to changes in both riparian vegetation (Gomi et al., 2006) and anthropogenic flow regulation (Dickson et al., 2012), as well as to assess the effects of forest treatments on streamflow (Watson et al., 2001).

Hydrological data typically possesses features such as seasonality, non-stationarity, and autocorrelation (Beven and Westerberg, 2011). Such features can be handled by the data-based time series approach by choosing an appropriate model based on the statistical properties of the data (Hipel and McLeod, 1994). Particularly, this type of modelling requires the underlying processes which generate the time series to be stationary, or in a state of 'statistical equilibrium' over time (Hipel and McLeod, 1994). If the statistical properties of the time series changed over time the inferences, forecasts or simulations generated using the fitted model would not be valid unless the underlying non-stationarity of the data was taken into account. To address this there is an array of time series modelling approaches and techniques which account for underlying non-stationarity. Artificial neural networks (Dorofki et al., 2012; Piotrowski and Napiorkowski, 2013; Thirumalaiah and Deo, 1998), support vector machines (Han et al., 2007; Lin et al., 2006), classification and regression trees (Noymanee and Theeramunkong, 2019; Yin et al., 2018) and transfer function models (Beven et al., 2008; Leedal et al., 2010; Romanowicz et al., 2008; Young, 2003) have all been applied to the modelling of hydrological data.

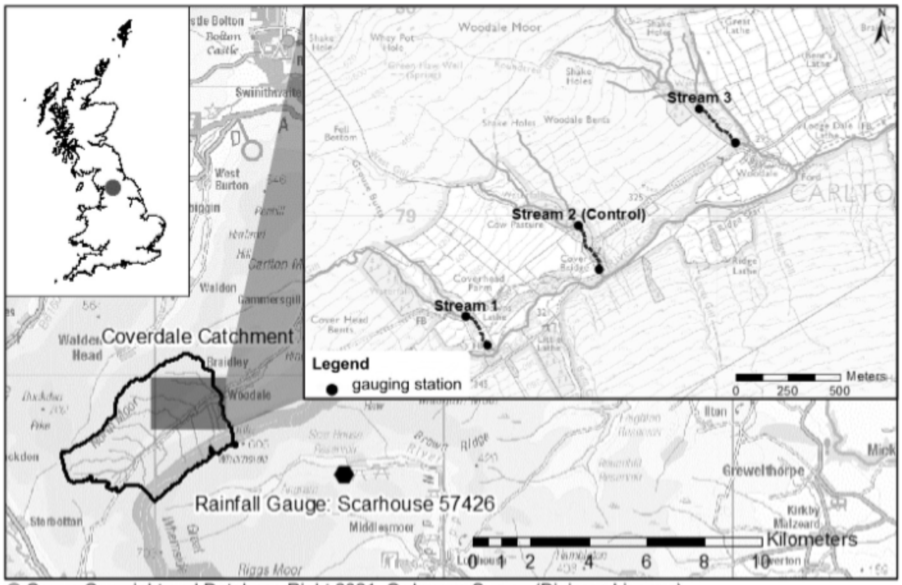
The transfer function noise (TFN) family of models are predominantly used when a time series can be modelled by linearly transforming one or more predictor time series and the resulting residuals of that transformation are autocorrelated. TFN models are therefore particularly well suited to modelling hydrological data (von Asmuth et al., 2002). Watson et al. (2001) demonstrated how the approach could be applied to paired catchment studies to address problems typical of hydrological data which invalidate the assumptions of traditional statistical methods. TFN models are more widely applied in the fields of systems engineering, econometrics and the social sciences (Okiy et al., 2015) but have been used to model hydrological data for several decades (e.g. Dooge, 1959; Jakeman et al., 1990; Young, 1986). In hydrology, TFN models are most commonly used to model rainfall-runoff relationships (e.g. Katimon et al., 2013; Ratto et al., 2007; Young, 2003), but they have also been used to fill gaps in hydrological records (Tencaliec et al., 2015), real-time level to level forecasting (Leedal et al., 2010; Young, 2002), modelling groundwater fluctuations (von Asmuth et al., 2002) and to detect impacts on hydrological processes (Dickson et al., 2012; Katimon et al., 2013; O'Driscoll et al., 2016). Transfer functions have been shown to produce simulations of peak event magnitude to a high level of accuracy: based on upstream stage series a transfer function model was able to simulate downstream stage on the River Severn to within 0.006 m to 0.139 m (0.1%–3.7%) for varying lead times (between 2 and 14 h) at the peak of an event (Romanowicz et al., 2008). They have been shown to produce accurate simulations of stage ($R^2 = 0.94$) even when fitted and validated using only a short period (20 days) of data (Young, 2003).

Transfer function noise modelling and other top-down, data-based time series modelling techniques, therefore, present an opportunity to extract information from typically short periods of baseline data collected before NFM interventions are installed. Successful application of the approach would allow for comparison between pre and post-intervention response of a stream even if directly comparable flood hydrographs were not captured in both monitoring periods. Given the difficulties in monitoring and processing hydrological data typically collected in NFM projects (Arnott et al., 2018), such a method would be a valuable tool to assess the efficacy of NFM interventions. This research evaluates the role which data-based time series modelling techniques could play in quantifying NFM impacts through application of the approach to data collected during a BACI monitoring campaign in three steep, upland streams. We assess whether linear TFN models are able to simulate the pre-intervention level to level response of upland streams during flood events to a sufficiently high degree of accuracy to inform the baseline conditions prior to the installation of in-stream NFM interventions. The models developed in this study were used in the companion paper, (van Leeuwen et al., 2023), to show that leaky dams in an upland stream significantly reduced the flood peak magnitude of events with a return period of up to 1 year, by 10% on average. The data based time series modelling approach allowed the effectiveness of leaky dams to be quantified for 50 events with a return period up to 6 years, giving novel insights about the variability of their impacts, particularly during single and multi peaked events. Together with the companion paper (van Leeuwen et al., 2023), this study demonstrates the potential that data-based time series modelling has for building the quantitative evidence base needed for NFM measures, such as leaky dams, to become more mainstream.

2. Methodology

2.1. Site description

The study streams were located in the headwaters of the River Cover (54.20045°N, −1.98617°E) on the Eastern flank of the Yorkshire Dales National Park, North Yorkshire, England (Fig. 1). The climate is cool and wet, with an average annual rainfall of 1270 mm (Environment Agency rain gauging station 57 426 data 1988–2018). The headwaters are formed of many small, parallel streams which flow into the River Cover at an altitude of ~ 400 m AOD. The study focuses on three of these streams, with a combined catchment area of 4.7 km^2. The watercourses are of type A in the Rosgen classification; steep, partially entrenched and cascading with step-pool streams (Rosgen, 1994). Land use in the catchment is pastoral agriculture on open, unimproved grass-



© Crown Copyright and Database Right 2021. Ordnance Survey (Digimap Licence)

Fig. 1. Location of the headwaters of the River Cover, Coverdale, North Yorkshire, UK, and the nearest Environment Agency operated rainfall gauge. The inset shows the location of the studied streams within the Coverdale catchment and the positions of the water level gauges are indicated using black points.

Table 1
Stream characteristics.

Stream	Gradient (m/m)	Catchment area (km^2)	Monitored length (m)	Mean width (m)
1	0.13	1.1	280	2.6
2 (control)	0.11	1.9	260	3.0
3	0.09	1.7	250	2.7

land whilst the moorland is managed for grouse shooting. The streams are representative of the type of sites in which the installation of leaky dams has been proposed and completed in North Yorkshire (Yorkshire Dales National Park Authority et al., 2017).

The hydrological characteristics of the streams are summarised in Table 1. Stream catchment areas were calculated using a 30 m resolution digital elevation model (NASA Shuttle Radar Topography Mission (SRTM), 2013) in the global information system ArcMap, version 10.6. The catchment areas were adjusted based on 5 m elevation contours from the OS Terrain 5 dataset (Ordnance Survey (GB), 2016). The monitored stream lengths were chosen to avoid including lateral inflows within the monitored reaches.

2.2. Monitoring network

Water stage data, defined as water level above the gauge datum, were collected following a Before After Control Impact (BACI) methodology as described by Smith (2002). Fig. 1 shows the control stream (Stream 2) and two impact streams (Stream 1 and 3) in which leaky dams were installed at the end of the baseline monitoring period. Baseline data were collected between March 2017 and September 2018 for a total period of 13 to 18 months before leaky dams were placed in the impact streams between September and October 2018. The length of the baseline monitoring period varied between the streams because of the timing at which the monitoring equipment and the leaky dams could be installed on the streams. The design and locations of the leaky dams are described in detail in van Leeuwen et al. (2023). Post-intervention stage data were collected for a further 16 months between October 2018 and February 2020.

Stage was monitored at one-minute intervals using In-Situ Inc. (Redditch, UK) Rugged TROLL 100 non-vented pressure transducers ($\pm$0.05% full scale accuracy) in stilling wells at the upstream and downstream extent of the study reaches on each stream (Fig. 1). The pressure readings from the transducers were corrected for atmospheric pressure using an In-Situ Inc. (Redditch, UK) Rugged BaroTROLL atmospheric pressure gauge ($\pm$0.05% full scale accuracy) which was installed near the bottom of stream 1, at a similar elevation to the water level pressure transducers. The control stream was monitored throughout the baseline and post-intervention monitoring periods without any interventions.

2.3. Data analysis

Stage data were quality assured, smoothed and aggregated from 1-minute to 15-minute time-steps using R version 4.0.2 (R Core Team, 2020). To meet the condition of stationarity, stage data were transformed by taking the first order difference after Box and Jenkins (1976). The KPSS stationarity test (Kwiatkowski et al., 1992) and ADF unit root test (Fuller, 1996) were performed on the differenced upstream and downstream stage series, U^* and D^*, to verify that the data were stationary as a result of the data transformation.

The analysis approach was guided by the characteristics of the data and was similar to that taken by Dickson et al. (2012), Watson et al. (2001) and Gomi et al. (2006) in analyses of water temperature response to catchment modifications. Based on exploratory regression and autocorrelation analysis of the differenced stage series, the decision was made to represent the relationship between upstream and downstream stage by a linear transfer function noise (TFN) model.

Linear transfer functions mathematically describe the dynamic linear relationship between a given input and output. The transfer function, or dynamic component of the model, is similar to a multiple linear regression, but one in which the predictive variables can include one or more lagged versions of a variable. Autocorrelation in the residual series, which would not meet the assumptions of independence required in linear regression, is taken into account by fitting a time series model to the residual series in the noise component of the model. Hence, the transfer function model consisted of a dynamic component and a noise component:

$$output = dynamic\ component + noise \tag{1}$$

The dynamic component of the TFN model was a linear function in which the forecast variable, D^*, the first order difference of downstream stage, was regressed against the predictive variable, U^*, the first order difference of the upstream stage, and k number of lagged values of U^*; thereby representing how the input, (U^*_t), dynamically affects the output, (D^*_t):

$$D_t^* = U_t^* + \upsilon_1 U_{t-1}^* + \upsilon_2 U_{t-2}^* + \cdots + \upsilon_k U_{t-k}^* \tag{2}$$

where υ_1 to υ_k are the impulse response weights, which were inferred from the data.

The residual series was represented using an autoregressive moving average (ARMA) model, which is common practice for hydrological data (Bell et al., 2001; Katimon et al., 2013; Yuan et al., 2009; Hipel and McLeod, 1994). Based on examination of the autocorrelation function and partial autocorrelation function of the residual series, the noise term was represented using both autoregressive and moving average terms. For the stationary noise series, N_t, the ARMA model was of the form:

$$N_t = \sum_{j=1}^{p} \phi_j N_{t-j} + \sum_{j=1}^{q} \theta_j a_{t-j} + a_t \tag{3}$$

where $a_t \sim N(0, \sigma^2)$ and where $\phi_j = (\phi_1, \phi_2, \ldots, \phi_p)$ and $\theta_j = (\theta_1, \theta_2, \ldots, \theta_q)$ are the vectors of model coefficients of order p and q, which denote the number of autoregressive and moving average parameters respectively, 'j' is the autoregressive or moving average term between one and p or q respectively, and a_t is the residual series. To obtain estimates for the model parameters, a_t was required to be independent and normally distributed with zero mean and fixed variance $N(0, \sigma^2)$.

The two parts of the transfer function noise model were fitted simultaneously. The dynamic linear regression component of the model, given by Eq. (2), and the ARMA noise model, given by Eq. (3), were fitted by maximum likelihood estimation (MLE) using the package forecast (v.8.12) (Hyndman and Khandakar, 2008) in R (v.4.0.2).

To fit a parsimonious model, the Minimum Akaike Information Criterion Estimation (MAICE) procedure introduced by Akaike (1973) was followed. The MAICE procedure is an adaptation of Box and Jenkin's model fitting procedure (Box and Jenkins, 1976) which uses the Akaike Information Criterion (AIC) as an indicator of predictive ability, which penalises for increasing the number of terms used, to discriminate between models. The number of dynamic regressors, k, was determined using a forward stepwise approach after Hyndman and Athanasopoulos (2018), and the number of AR and MA terms was determined by carrying out a search over the ARMA model order space to identify the optimum combination of terms by minimising the AIC score. The residuals of the model were checked for autocorrelation, heteroscedasticity, and non-normal distribution.

2.4. Assessing model performance

The accuracy of the model predictions was assessed using a blocked cross-validation approach. Based on the recommendations of Bergmeir and Benítez (2012) & Roberts et al. (2017) multiple out of sample time series were simulated by removing one ten-hour block of data from the series at a time and testing the model trained on the remaining data on the excluded block. The ten hour blocks of data had a 15 min time step and were centred on the peak of high flow events.

The Hydrological Model Assessment and Development (HydroMAD) v.0.9-26 R package (Andrews and Guillaume, 2018) was used to identify discrete storm events. A stage peak was considered a discrete high flow event if it was part of a defined flow event with duration >60 min and the upstream peak stage exceeded the mean stage recorded on the stream. Events were classed as independent if they were separated by at least 15 min of stage below or within 10% of baseflow stage.

The measures to describe the ability of the model to predict the event peak magnitude were absolute peak error (PE) and peak error percentage (PEP) given by Eqs. (4) and (5), where $\hat{D}_{peak}$ is the model simulated peak magnitude, and D_{peak} is the observed peak magnitude, in metres. The measures used to assess goodness of fit throughout the event, rather than just at the peak, were root mean square error (RMSE) and Nash–Sutcliffe Efficiency (NSE). Although the use of NSE assumes that the errors are independent, it was used here because of its wide application to assess the goodness of fit of hydrological models (McCuen et al., 2006).

$$PE = \hat{D}_{peak} - D_{peak} \tag{4}$$

$$PEP = \frac{(\hat{D}_{peak} - D_{peak})}{D_{peak}} \tag{5}$$

2.5. Model uncertainty

Theoretical prediction intervals are usually calculated based on the standard error of the innovation series and the residuals of the fitted model (Chatfield, 2001; Lee and Scholtes, 2014). However, although they are commonly used, it is widely accepted that theoretical prediction intervals are almost always too narrow in practice because they account only for the uncertainty due to random error (Hyndman et al., 2002; Makridakis and Winkler, 1989) and may not provide adequate cover if the assumptions of normal, independent and identically distributed residuals are not strictly met (Hyndman and Athanasopoulos, 2018). Therefore, a common alternative approach was used to calculate prediction intervals based on the empirical out of sample forecasts which account for random, parameter and model specification errors whilst only assuming that the error distribution of future simulations is similar to the error distribution of the out of sample simulations (Lee and Scholtes, 2014; Williams and Goodman, 1971). Empirical prediction intervals have been successfully applied in a wide range of fields (Isengildina-Massa et al., 2011; Lee and Scholtes, 2014; Rayer et al., 2009). After Williams and Goodman (1971), who introduced the approach, the prediction interval at each forecasting timestep was estimated using specified quantiles of the empirical error distribution at that timestep. As the error for multiple step ahead simulations was additive (Hyndman and Athanasopoulos, 2018) the simulation window was centred on the peak of the event so that the event peak estimation error was always calculated for the same timestep, $(\frac{N}{2})$, where N was the number of simulation timesteps. The 95% prediction interval, and the 80% prediction interval were calculated in this way. Although the 95% prediction interval is more stringent, the 80% prediction interval was included because it is recommended for error distributions with outliers, or 'tail problems' (Chatfield, 2001).

To assess the model's overall performance at predicting event peak magnitude for the range of events tested, the observed peak magnitudes were linearly regressed against the simulated peak magnitudes and the confidence interval of the relationship was calculated based on the standard errors of the linear relationship. The confidence interval describes the uncertainty associated with the regression coefficients, which are parameters of the population (Hahn and Meeker, 1991). It was therefore expected to be narrower than the prediction intervals, which quantify the uncertainty associated with the prediction of an individual data point. The NSE_p (Eq. (6)), which has been widely used to assess goodness of fit of hydrological models (McCuen et al., 2006), was used to test the closeness of the relationship to the one-to-one line. In Eq. (6), n is the total number of events and $\bar{D}_{peak}$ is the mean of the observed values of peak magnitude.

$$NSE_p = 1 - \frac{\sum^{N} (\hat{D}_{peak} - D_{peak})^2}{\sum (D_{peak} - \bar{D}_{peak})^2} \tag{6}$$

3. Results

The upstream and downstream baseline stage series collected on the two impact streams and the control stream are shown in Fig. 2. The downstream stage was highly cross-correlated to upstream stage

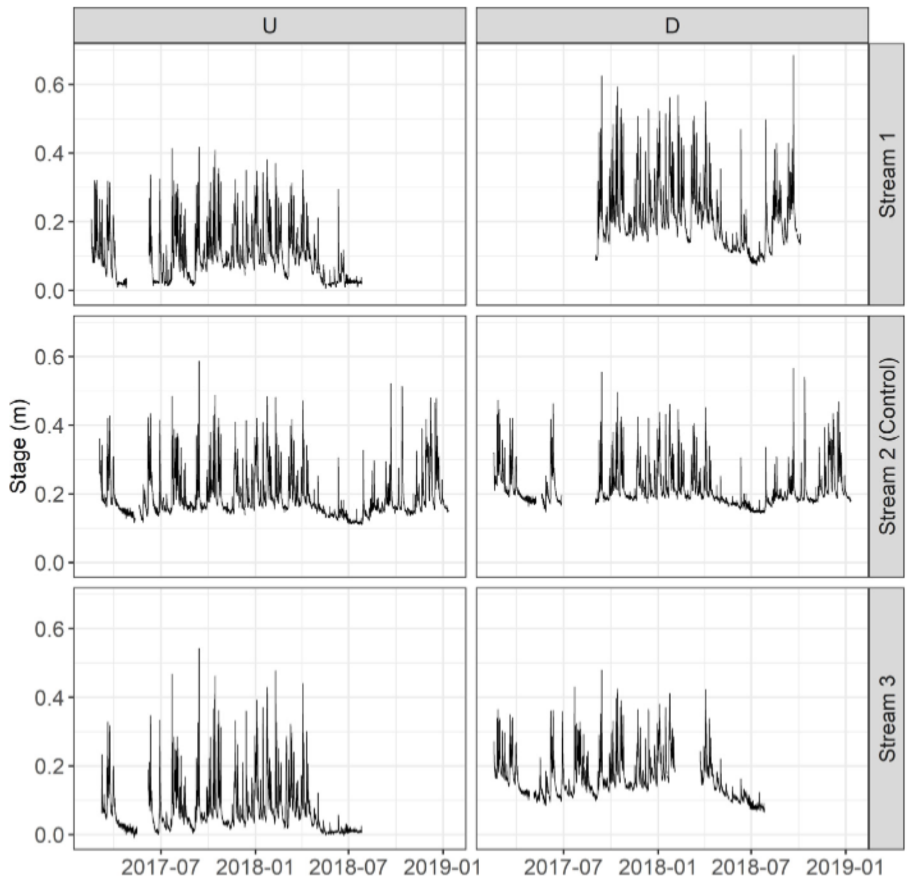


Fig. 2. Upstream (*U*) and Downstream (*D*) stage series collected in the three study streams.

at lag zero (i.e. instantaneously) (CCF 0.97–0.99) on all three streams reflecting the steep nature of the catchment. Seasonality was present in all three streams with periods of lower baseflows in the summer months, particularly in the summer of 2018, which was exceptionally dry across the UK (Met Office, 2018). Apart from the summer of 2018, high flow events were recorded regularly throughout the baseline monitoring period (Fig. 2). The highest stage peaks were recorded during Storm Aileen (13 September 2017), ex-hurricane Ophelia (14 October 2017) and Storm Bronagh (20 September 2018).

High levels of uncertainty were identified in the stage data during a thorough quality assurance process. Small, gradual changes in the relationship between upstream and downstream stage of magnitude up to ± 0.05 m occurred frequently. These changes were constant over the range of stage, for example, when there was a shift of 0.05 m at baseflow there was also a shift of 0.05 m at the peak. They were, therefore, assumed to be 'datum errors' brought about by a change in the reference datum or flow conditions at the gauging station. Based on field observations the most likely source of datum error was frequent blockage of the gauging stations with material on the outside, and sediment on the inside of the stilling well. By modelling the first order difference of the stage data, rather than the absolute stage, the models were made independent of these datum errors.

3.1. TFN model equations

The data for stream 1, stream 2 (control) and stream 3 were modelled with six, five and four dynamic regression terms, respectively. By testing all possible combinations of AR and MA terms for each noise series the model with minimum AIC score was found to be the ARMA (5, 3) model for stream 1, the ARMA (4, 2) model for stream 2 (control) and the ARMA (2, 3) model for stream 3, where ARMA (p, q) refers to the number of autoregressive terms, p, and moving average terms, q, of the ARMA model. The model for stream 1 included lag 0 to 3, lag 11 and lag 20, where each lag term refers to the number of 15 min timesteps by which the series was lagged. The model for stream 2 (control) included lag 0 to 4 and lag 11. The model for stream 3 included lag 0, 1, 3, and 4.

The terms and parameters of the TFN models describing baseflow conditions in stream 1, stream 2 (control) and stream 3 are provided in Table 2 along with their standard error. The close fit of the data to the one-to-one line on the plots of fitted against observed values in Fig. 3 show the good in-sample fit of the model to the data (RMSE 0.00097 m).

3.2. Simulations of downstream baseline stage during high flow events

Between 32 and 54 high flow events were identified in the data on each of the streams. Different numbers of high flow events were identified on the streams because of periods of missing data, and due to the interventions being installed a month earlier on stream 3 than stream 1. Each event was removed in turn and the model coefficients were re-estimated on the remaining data. The downstream stage of the 'hold out' event was then simulated by providing the model with the upstream stage series of the 'hold out' event. The simulated downstream stage was compared to the downstream stage observed during the event to assess the accuracy of the simulation. Overall, the model simulations fit the observed stage well (NSE 0.976–0.996, Fig. 4). For example, for an event in Autumn 2017 the models predicted peak stage to within 0.4 cm on the two impact streams, and under predicted peak stage by 1.0 cm on the control stream (Fig. 4) . These are fairly typical errors, with the majority of events predicted to within 0.4 cm on stream 1, and within 1 cm on streams 2 and 3 (Table 3).

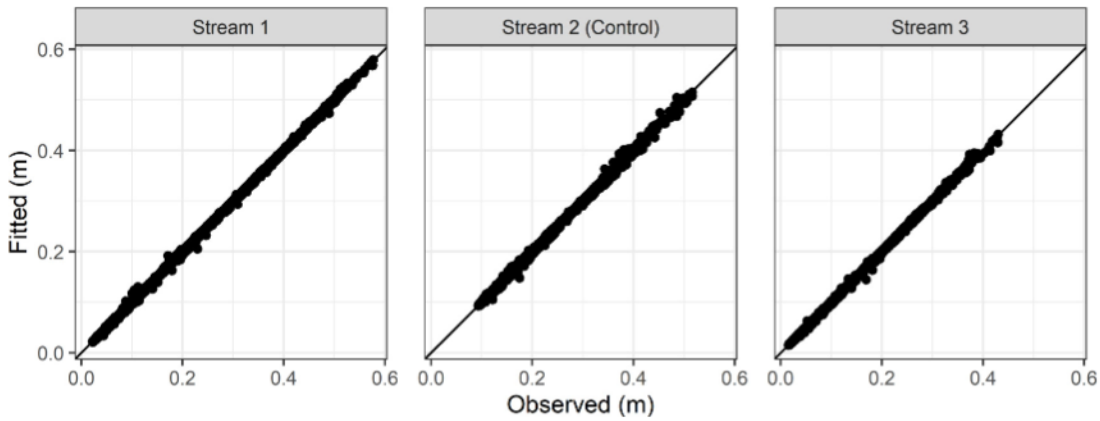


Fig. 3. Points of fitted downstream stage plotted against observed downstream stage with the one-to-one line for reference.

Table 2
Parsimonious TFN model parameter coefficients; coeff. = coefficient, s.e. = standard error in metres.

	Parameter	Stream 1		Stream 2 (control)		Stream 3	
		coeff.	s.e	coeff.	s.e.	coeff.	s.e.
Transfer Function Parameters	U^*_{t}	0.687	0.005	0.430	0.003	0.583	0.003
	U^*_{t-1}	0.332	0.005	0.198	0.003	0.039	0.004
	U^*_{t-2}	0.165	0.005	0.114	0.003	–	–
	U^*_{t-3}	0.102	0.005	0.083	0.003	0.040	0.004
	U^*_{t-4}	–	–	0.042	0.003	0.040	0.004
	U^*_{t-11}	−0.014	0.004	–	–	–	–
	U^*_{t-20}	−0.010	0.003	–	–	–	–
AR terms	N_{t-1}	0.893	0.083	1.035	0.039	1.147	0.040
	N_{t-2}	0.179	0.097	−0.196	0.050	−0.441	0.037
	N_{t-3}	−0.437	0.061	0.173	0.020	–	–
	N_{t-4}	0.288	0.032	−0.062	0.008	–	–
	N_{t-5}	−0.162	0.014	–	–	–	–
MA terms	a_{t-1}	−0.554	0.083	−0.739	0.039	−0.865	0.039
	a_{t-2}	−0.654	0.074	−0.253	0.039	0.021	0.030
	a_{t-3}	0.453	0.051	–	–	0.229	0.011

Table 3
Distribution of Peak error (PE) and Percentage error in peak (PEP) on the three streams.

	Stream 1		Stream 2		Stream 3	
	PE (m)	PEP (%)	PE (m)	PEP (%)	PE (m)	PEP (%)
Median	−0.001	−0.2	−0.002	−0.9	0.0003	0.1
Upper Quartile	0.003	0.9	0.002	0.5	0.011	3.9
Lower Quartile	−0.004	−1.3	−0.010	−3.0	−0.007	−2.3
IQR	0.006	2.2	0.011	3.5	0.018	6.2
Maximum	0.012	3.5	0.038	9.2	0.037	10.0
Minimum	−0.025	−4.4	−0.059	−11.4	−0.020	−8.5

The error in simulating event peak magnitude was < 0.03 m for all but four of the simulated events across all three streams (Fig. 5). On stream 1 the PE was smallest and was distributed evenly above and below zero, indicating that the simulations were not biased. On stream 2 (control) and stream 3 the range in PE and PEP was larger, and although the median value of PE was close to zero, the majority of event magnitudes were under predicted on stream 2 (control), and over predicted on stream 3. Whilst the PEP was < 5% for all simulations on stream 1, it was up to 11% on stream 2 (control) and 10% on stream 3 (Table 3). The interquartile range (IQR) of PEP was within 6%, for all three streams, with the IQR on stream 1 being as low as 2% (Table 3).

The observed event peak magnitude and the simulated event peak magnitude were strongly correlated for stream 1 (NSE 0.994, Fig. 6) with a residual standard error of 0.008 m. The simulated peak magnitudes were both over and under predicted and were not affected by the event peak magnitude (Fig. 6). On stream 2 (control), however, PE was relatively small (RMSE 0.006 m) for events < 0.35 m but increased to a RMSE of 0.027 m and 0.036 m for events with peak magnitude > 0.35 m and > 0.40 m, respectively. For the largest events the peak magnitude was under predicted more frequently than it was over predicted, resulting in a linear relationship which lay below the one-to-one line (NSE 0.97). There was a relationship between PE and event magnitude on stream 3; events with peak magnitude < 0.25 m were under predicted whilst events with peak magnitude > 0.25 m were over predicted. Although the NSE was relatively high (0.98) the coefficients of the linear regression reflect the bias in the model simulations.

The empirical 80% and 95% prediction intervals at the peak of the simulated events varied between streams (Table 4). The empirical prediction intervals indicated that the models were able to predict peak magnitude to within ±0.02 m at the 80% prediction interval on the control stream and stream 3, and at the 95% prediction interval on stream 1 (Table 4).

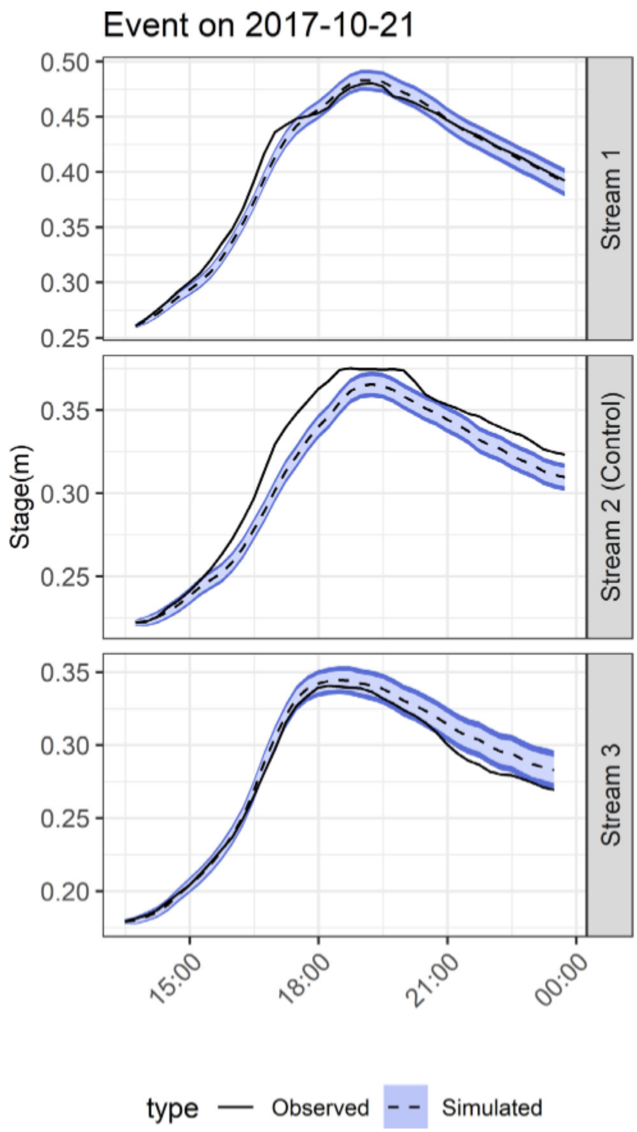


Fig. 4. Observed (solid line) and simulated (dashed line) downstream stage during an high flow event with 80% (dark blue shading) and 95% (light blue shading) empirical prediction intervals. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Empirical prediction interval width at event peak; PI = prediction interval.

	Lower 95% PI	Upper 95% PI	Lower 80% PI	Upper 80% PI
Stream 1	−0.017	0.012	−0.011	0.008
Stream 2 (control)	−0.051	0.027	−0.018	0.007
Stream 3	−0.016	0.032	−0.009	0.017

4. Discussion

Increasing the quantitative evidence base of NFM impacts on the flood hydrograph is crucial for measures such as leaky dams if they are to become more mainstream in flood risk and environmental land management (Ellis et al., 2021). Efforts to address the lack of quantitative evidence of leaky dam impacts on flood peak magnitude have been hampered by a lack of baseline data (Burgess-Gamble et al., 2017; Ellis et al., 2021), lack of comparable events monitored before and after the installation of leaky dams (National Trust, 2015), and high levels of uncertainty in stage and/or discharge series (Lane, 2017; Connelly et al., 2020). To grow the quantitative evidence base of NFM impacts new approaches which are able to overcome these barriers are needed. This study, together with van Leeuwen et al. (2023) signifies the value of the data-based time series modelling approach for quantifying leaky dam impacts on flood peak magnitude. This study demonstrates that data-based time series modelling can be used to overcome the barriers presented by poor quality, relatively short baseline data series.

4.1. Data quality

The finding that the collected stage data was highly uncertain is not unique to this study; it has long been acknowledged that hydrological data is ‘messy’ (Beven and Westerberg, 2011). Although uncertainty in the stage series can be one of the most significant sources of error (Di Baldassarre and Montanari, 2009), stage datum errors are often assumed to be negligible (Horner et al., 2018). Identifying such errors necessitates a thorough quality assurance process which requires sufficient resource allocation to post-processing of the data. Frequent inspection, maintenance and calibration of gauging stations, and the use of artificial gauging structures such as flumes or weirs, could mitigate these problems but has further implications for resource allocation to the monitoring of NFM schemes. This study has shown that, for BACI data with levels of uncertainty which mask the signal of NFM interventions, valuable information can be extracted by using a data-based time series modelling approach. The study emphasises that proper specialist resource is not only required to monitor NFM measures (Arnott et al., 2018), but also to post-process and analyse the data.

4.2. Model accuracy

The level of accuracy achieved by TFN models depends on many factors including the quality of the input data, choice of model timestep and model structure (Sene and Tilford, 2004), which makes comparison between studies difficult. The accuracy of the models was therefore assessed in terms of the intended purpose: detecting leaky dam impacts in the observed post-intervention stage by providing accurate simulations of baseline, pre-intervention stage for comparison. Identifying the level of accuracy required for this presents difficulties as the impacts of leaky dams in upland watercourses are not yet known (Burgess-Gamble et al., 2017). However, it is known that wood placed in upland streams for the purpose of river restoration reduced event peak magnitude by 8% and 2.2% in steep watercourses (Keys et al., 2018; Wenzel et al., 2014) and in lowland rivers peak magnitude reductions of 21% have been observed for combined planform and large wood restoration (Kitts, 2010). The results of these previous studies imply that for the majority of events on stream 1, the simulations are likely to be sufficiently accurate (± 1%) to detect leaky dam impacts on individual events. Furthermore, by combining the simulations of all events and calculating the confidence interval (0.02 m on average at the 95% confidence level) of the relationship between observed and simulated stage it becomes clear that there is a very high level of confidence in the ability of the models to replicate downstream event peak magnitude on average, particularly on stream 1.

4.3. Implications

Whilst there are many examples of high levels of predictive ability achieved using transfer functions in hydrology, the majority of applications are in rainfall runoff modelling (see for example, Katimon et al. (2013), Ratto et al. (2007) & Young (2003)). This study shows that it is possible to achieve sufficiently high levels of accuracy in predictions of the level-to-level response of small upland watercourses to detect leaky dam impacts. Unlike previous application of the approach to detect changes in long-term stream flow patterns (Watson et al., 2001), this was achieved for high temporal resolution data and can therefore be

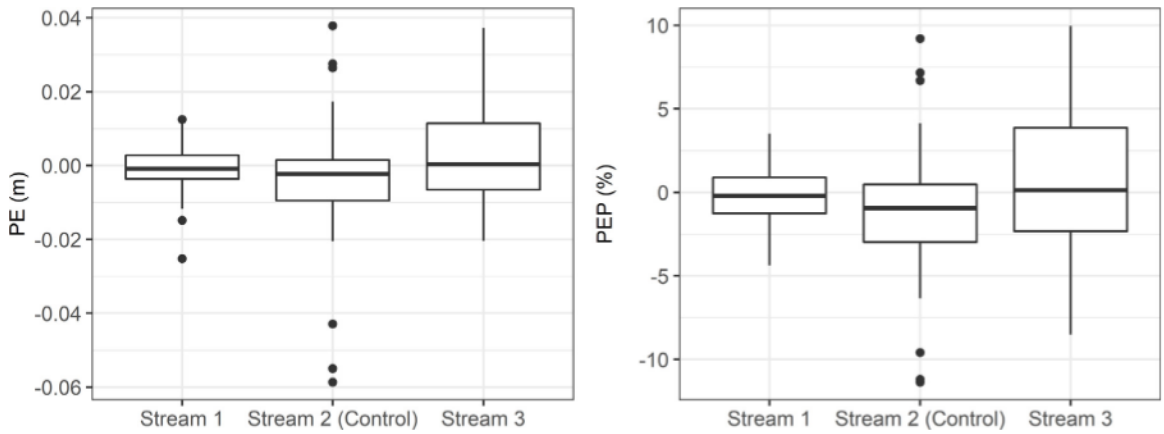


Fig. 5. Skill of TFN models at predicting out of sample event peak magnitude. Peak error (PE) and Percentage error in peak (PEP) is the difference between the observed and simulated downstream peak stage.

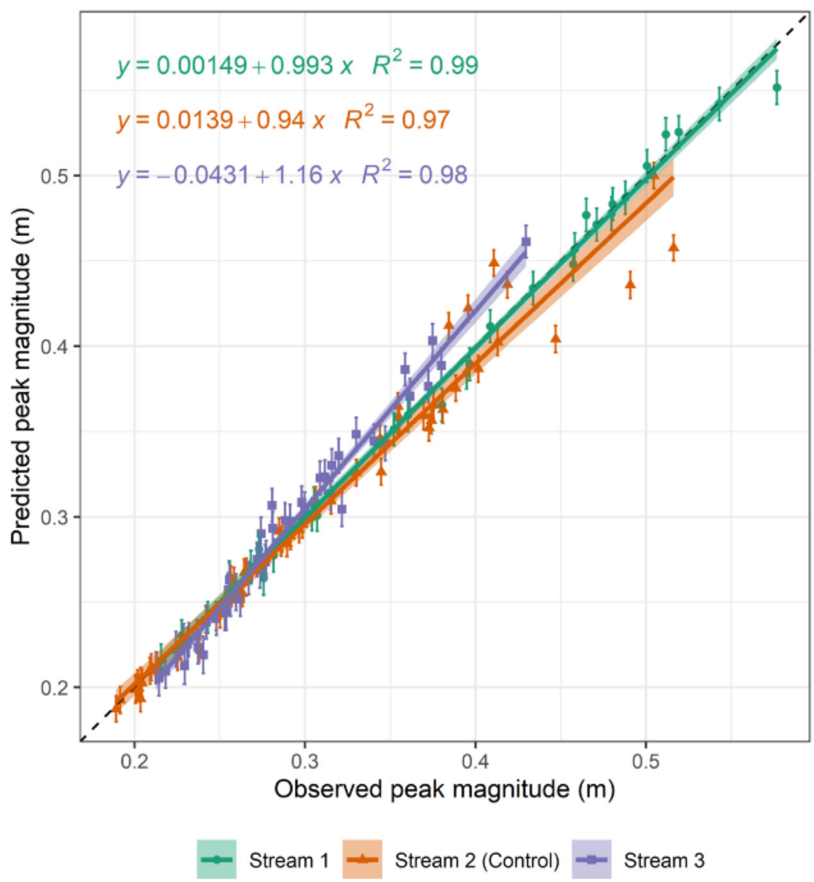


Fig. 6. Goodness of fit of out of sample simulations of event peak stage with theoretical prediction intervals. The shaded areas shows the confidence interval of the linear relationship between simulated and observed peak magnitude and the dashed line shows the 1-to-1 relationship.

used to detect impacts of small scale treatments, like leaky dams, on individual high flow events.

Assuming the basic properties of the (differenced) baseline stage data are constant over time and given adequate control data, any differences in the simulated baseline response and observed intervention response for an event can be attributed to the NFM interventions. Notably, baseline and post-intervention monitoring datasets are therefore not required to contain events which are so similar that their responses can be compared directly. This is a critical advancement because the stochastic nature of flood events has precluded the assessment of NFM impacts in several studies (Kitts, 2010; National Trust, 2015), or provided evidence of only one or two comparable events (Nisbet et al., 2015; Thomas and Nisbet, 2007; Wilkinson et al., 2010). By simulating downstream baseline stage, the impact of interventions can be assessed for every event in the post-intervention monitoring period, providing replicates and allowing impacts to be assessed at a range of event magnitudes. By using this approach the impacts of NFM measures may, therefore, be more conclusively evaluated from both new and existing datasets.

The quantitative evidence base of leaky dam effectiveness, which may be built by applying this approach, is needed to assess whether leaky dams are a viable flood risk management technique in upland catchments, during what types of flood events they are effective, and to inform the design of natural flood management schemes. Hydraulic and hydrological models are the tools which are commonly used to make such assessments, but confidence in their outputs is currently held back by a lack of empirical validation data (Addy and Wilkinson, 2019). This study demonstrates that the data-based time series modelling approach has the potential to provide the empirical evidence needed to validate the representation of leaky dams in such models. By providing the data to validate local effects, confidence in the assessment of catchment scale impacts is also increased, although problems of scaling need to be addressed (Hankin et al., 2021; Lavers et al., 2022). Alternatively, to avoid problems of scaling, a multi-local scale approach has been suggested which focuses on multiple smaller catchments with downstream communities at risk to quantify impacts within a larger catchment (Hankin et al., 2021). The data-based time series modelling approach can be readily applied to such catchments, as is demonstrated in van Leeuwen et al. (2023).

4.4. Limitations and further work

In this study several notable high flow events were captured, but baseline data which does not contain many high flow events cannot be used to train a model to simulate high flow events. The problem of capturing a range of events may, therefore, not always be overcome by using this approach. However, Young (2003) showed for the case of rainfall-runoff modelling that high levels of accuracy ($R^2 = 0.94$) can be achieved using a model fitted and validated using only 20 days of hourly observations. Hence, by using a data-based modelling approach assessing the impacts of NFM measures, even where only very short periods of baseline data are available, may become viable.

The results on stream 1 demonstrate the potential that linear TFN models have for extracting information about peak magnitude from highly uncertain baseline data (0.02 m accuracy at 95% prediction interval). Further work is required to identify models which provide a better fit for larger events on stream 2 (control) and stream 3 (0.02 m accuracy at 80% prediction interval). It is likely that the lack of fit at higher event magnitudes is due to non-linearity in the response (e.g. due to differences in the geometry of the gauging cross-sections (Romanowicz et al., 2008)), which can be incorporated in the TFN approach by including a non-linear transform to the data. The DBM approach, for example, has been developed to model typically non-linear relationships between rainfall and runoff (Young, 2003) and has been successfully applied to model non-linear level to level responses (Beven et al., 2008; Leedal et al., 2010; Romanowicz et al., 2008; Young, 2002). Alternatively, a class of models which accounts for non-constant variance, such as autoregressive conditional heteroscedasticity (ARCH) models could be explored.

This study has shown that the data-based time series modelling approach has the potential to extract information from new and existing BACI data, therefore, it has the capacity to provide the quantitative evidence base needed for NFM measures to become mainstream. To achieve this, further work is required to assess the potential of using data-based time series modelling technique to analyse BACI data from a range of environments, NFM interventions and scales. As outlined in van Leeuwen et al. (2023), application of the models to post-intervention or ‘after’ data collected from this study site has already provided a comprehensive dataset of leaky dam impacts on flood peak magnitude.

5. Conclusion

Whilst previous research has focused on using either an empirical approach or a deterministic modelling approach to detect NFM impacts on downstream flood risk, the novel application of the approach in this research demonstrates that where uncertainty masks the signal of the intervention, a top-down data-based time series modelling approach can provide the tools needed to make a meaningful comparison between empirical baseline and post-intervention data.

The linear TFN models were able to predict peak stage with 0.02 m accuracy at the 95% prediction interval on stream 1, and with 0.02 m accuracy at the 80% prediction interval on streams 2 and 3. Although different types of time series models may be necessary to demonstrate the full benefits of data-based time series modelling approaches, this study provides evidence that, where the underlying data generating processes are linear, linear TFN modelling can reproduce observed stage hydrographs to a high degree of accuracy. Given the upstream stage series, the baseline, pre-intervention response of a stream can therefore be accurately simulated for any chosen high flow event. Thus, for every flood peak observed after an NFM intervention is made the baseline flood peak magnitude can be simulated for comparison with a high degree of confidence. The impact of an NFM intervention can thereby be quantified for the full range of events observed after the interventions are installed. Hereby it is demonstrated that, by using a data-based time series modelling approach, BACI data can be used to assess the impact of NFM features such as leaky dams on downstream flood peak magnitude. The impact can be assessed even when lead times to collect baseline data are short, the data are highly uncertain and comparable high flow events are not observed before and after the interventions are installed. Data-based time series modelling techniques, therefore, provide a promising solution to the problems associated with quantifying the flood risk management benefits of NFM interventions.

CRediT authorship contribution statement

Z.R. van Leeuwen: Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **M.J. Klaar:** Conceptualization, Writing – review & editing, Supervision, Funding acquisition. **M.W. Smith:** Conceptualization, Writing – review & editing, Supervision, Funding acquisition. **L.E. Brown:** Conceptualization, Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

The authors would like to thank the Yorkshire Dales Rivers Trust for assistance with building leaky dams and the Mawle family for allowing this research to be completed on their land. This work was funded by Natural Environment Research Council with Industrial CASE support from JBA Trust and the Yorkshire Dales Rivers Trust, NE/N008065/1. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

References

Addy, S., Wilkinson, M.E., 2019. Representing natural and artificial in-channel large wood in numerical hydraulic and hydrological models. WIREs Water 6 (6), 1–20. http://dx.doi.org/10.1002/wat2.1389.

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B., Csaki, F. (Eds.), International Symposium on Information Theory. Budapest, pp. 267–281.

Andrews, F., Guillaume, J., 2018. hydromad: Hydrological model assessment and development. URL: http://hydromad.catchment.org/.

Arnott, S., Burgess-Gamble, L., Dunsford, D., Webb, L., Johnson, D., Andison, E., Slaney, A., Vaughan, M., Ngai, R., Rose, S., Maslen, S., 2018. Monitoring and Evaluating the DEFRA Funded Natural Flood Management Projects. Technical Report July, Environment Agency, pp. 1–18, URL: http://www.gov.uk/government/publications.

Bark, R.H., Martin-Ortega, J., Waylen, K.A., 2021. Stakeholders' views on natural flood management: Implications for the nature-based solutions paradigm shift? Environ. Sci. Policy 115, 91–98. http://dx.doi.org/10.1016/j.envsci.2020.10.018, URL: https://linkinghub.elsevier.com/retrieve/pii/S1462901120313678.

Bell, V.A., Carrington, D.S., Moore, R.J., 2001. Comparison of Rainfall-Runoff Models for Flood Forecasting Part 2: Calibration and Evaluation of Models. R& D Technical Review W242, Technical Report, Environment Agency, Bristol, p. 239. http://dx.doi.org/10.1097/PRS.0000000000002088.

Bergmeir, C., Benítez, J.M., 2012. On the use of cross-validation for time series predictor evaluation. Inform. Sci. 191, 192–213. http://dx.doi.org/10.1016/j.ins.2011.12.028.

Beven, K., 2001. Rainfall-Runoff Modelling: The Primer, second ed. Wiley-Blackwell, Chichester.

Beven, K., Page, T., Hankin, B., Smith, P., Kretzschmar, A., Mindham, D., Chappell, N., 2022. Deciding on fitness-for-purpose-of models and of natural flood management. Hydrol. Process. 36 (11), http://dx.doi.org/10.1002/hyp.14752, URL: https://onlinelibrary.wiley.com/doi/10.1002/hyp.14752.

Beven, K., Westerberg, I., 2011. On red herrings and real herrings: Disinformation and information in hydrological inference. Hydrol. Process. 25 (10), 1676–1680. http://dx.doi.org/10.1002/hyp.7963.

Beven, K., Young, P., Leedal, D., Romanowicz, R., 2008. Computationally efficient flood water level prediction (with uncertainty). Flood Risk Manag. Res. Pract. 281–289. http://dx.doi.org/10.1201/9780203883020.ch34.

Black, A., Peskett, L., MacDonald, A., Young, A., Spray, C., Ball, T., Thomas, H., Werritty, A., 2021. Natural flood management, lag time and catchment scale: Results from an empirical nested catchment study. J. Flood Risk Manag. 14 (3), 1–16. http://dx.doi.org/10.1111/jfr3.12717, URL: https://onlinelibrary.wiley.com/doi/10.1111/jfr3.12717.

Box, G.E.P., Jenkins, G.M., 1976. Time Series Analysis: Forecasting and Control, revised ed. Holden-Day, Oakland, California.

Burgess-Gamble, L., Ngai, R., Wilkinson, M., Nisbet, T., Pontee, N., Harvey, R., Kipling, K., Addy, S., Rose, S., Maslen, S., Jay, H., Nicholson, A., Page, T., Jonczyk, J., Quinn, P., 2017. Working with Natural Processes - Evidence Directory. Technical Report, Environment Agency, Bristol, URL: https://assets.publishing.service.gov.uk/.

Chatfield, C., 2001. Prediction intervals for time-series forecasting. In: Armstrong, J. (Ed.), Principles of Forecasting. In: International Series in Operations Research & Management Science, vol. 30, Springer, Boston, MA, pp. 475–494. http://dx.doi.org/10.1007/978-0-306-47630-3_21.

Commission of the European Communities, 2009. White Paper - Adapting to Climate Change - Towards a European Framework for Action. Technical Report, European Commission, Brussels, URL: https://eur-lex.europa.eu/legal-content/.

Connelly, A., Snow, A., Carter, J., Lauwerijssen, R., 2020. What approaches exist to evaluate the effectiveness of UK-relevant natural flood management measures? A systematic map protocol. Environ. Evid. 9 (1), 1–13. http://dx.doi.org/10.1186/s13750-020-00192-x.

Cook, B., Forrester, J., Bracken, L., Spray, C., Oughton, E., 2016. Competing paradigms of flood management in the Scottish/English borderlands. Disaster Prev. Manag. 25 (3), 314–328. http://dx.doi.org/10.1108/DPM-01-2016-0010, URL: https://www.emerald.com/insight/content/doi/10.1108/DPM-01-2016-0010/full/html.

Dadson, S.J., Hall, J.W., Murgatroyd, A., Acreman, M., Bates, P., Beven, K., Heathwaite, L., Holden, J., Holman, I.P., Lane, S.N., O'Connell, E., Penning-Rowsell, E., Reynard, N., Sear, D., Thorne, C., Wilby, R., 2017. A restatement of the natural science evidence concerning catchment-based 'natural' flood management in the UK. Proc. R. Soc. A 473 (2199), 20160706. http://dx.doi.org/10.1098/rspa.2016.0706, URL: https://royalsocietypublishing.org/doi/10.1098/rspa.2016.0706.

Defra, EAFRD, Natural England, 2016. Countryside Stewardship - How can it help reduce flooding? URL: https://www.gov.uk/government/uploads/.

Di Baldassarre, G., Montanari, A., 2009. Uncertainty in river discharge observations: A quantitative analysis. Hydrol. Earth Syst. Sci. 13 (6), 913–921. http://dx.doi.org/10.5194/hess-13-913-2009.

Dickson, N.E., Carrivick, J.L., Brown, L.E., 2012. Flow regulation alters alpine river thermal regimes. J. Hydrol. 464–465, 505–516. http://dx.doi.org/10.1016/j.jhydrol.2012.07.044.

Dixon, S.J., Sear, D.A., Odoni, N.A., Sykes, T., Lane, S.N., 2016. The effects of river restoration on catchment scale flood risk and flood hydrology. Earth Surf. Process. Landf. 41 (7), 997–1008. http://dx.doi.org/10.1002/esp.3919, URL: https://onlinelibrary.wiley.com/doi/10.1002/esp.3919.

Dooge, J., 1959. A general theory of the unit hydrograph. J. Geophys. Res. 64, 241–256.

Dorofki, M., Elshafie, A.H., Jaafar, O., Karim, O.A., Mastura, S., 2012. Comparison of artificial neural network transfer functions abilities to simulate extreme runoff data. In: International Conference on Environment, Energy and Biotechnology, Vol. 33. IACSIT Press, Singapore, pp. 39–44, URL: www.ipcbee.com/vol33/008-ICEEB2012-B021.pdf.

Ellis, N., Anderson, K., Brazier, R., 2021. Mainstreaming natural flood management: A proposed research framework derived from a critical evaluation of current knowledge. Prog. Phys. Geogr. Earth Environ. 1–23. http://dx.doi.org/10.1177/0309133321997299, URL: http://journals.sagepub.com/doi/10.1177/0309133321997299.

Environment Agency, 2010. Water framework directive (WFD) mitigation measures online manual - Management and use of large wood. URL: http://evidence.environment-agency.gov.uk/FCERM/.

Environment Agency, 2019. Natural Flood Management Programme : Interim Lessons Learnt. Technical Report January, Environment Agency, Bristol.

Forbes, H., Ball, K., McLay, F., 2015. Natural Flood Management Handbook. Technical Report, SEPA, Stirling, p. 142.

Fuller, W.A., 1996. Introduction to Statistical Time Series, second ed. John Wiley and Sons, New York.

Gebrehiwot, S.G., Di Baldassarre, G., Bishop, K., Halldin, S., Breuer, L., 2019. Is observation uncertainty masking the signal of land use change impacts on hydrology? J. Hydrol. 570, 393–400. http://dx.doi.org/10.1016/j.jhydrol.2018.12.058.

Gomi, T., Moore, R.D., Dhakal, A.S., 2006. Headwater stream temperature response to clear-cut harvesting with different riparian treatments, coastal British Columbia, Canada. Water Resour. Res. 42 (8), 1–11. http://dx.doi.org/10.1029/2005WR004162.

Grabowski, R.C., Gurnell, A.M., Burgess-Gamble, L., England, J., Holland, D., Klaar, M.J., Morrissey, I., Uttley, C., Wharton, G., 2019. The current state of the use of large wood in river restoration and management. Water Environ. J. 33, 366–377. http://dx.doi.org/10.1111/wej.12465, URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/wej.12465, https://onlinelibrary.wiley.com/doi/10.1111/wej.12465.

Hahn, G.J., Meeker, W.Q., 1991. Statistical Intervals: A Guide for Practitioners. In: Wiley Series in Probability and Statistics, Wiley, New York, http://dx.doi.org/10.1002/9780470316771.

Han, D., Chan, L., Zhu, N., 2007. Flood forecasting using support vector machines. J. Hydroinform. 9 (4), 267–276. http://dx.doi.org/10.2166/hydro.2007.027.

Hankin, B., Page, T.J.C., Chappell, N.A., Beven, K.J., Smith, P.J., Kretzschmar, A., Lamb, R., 2021. Using micro-catchment experiments for multi-local scale modelling of nature-based solutions. Hydrol. Process. 35 (11), http://dx.doi.org/10.1002/hyp.14418, URL: https://onlinelibrary.wiley.com/doi/10.1002/hyp.14418.

Hipel, K.W., McLeod, A.I., 1994. Time Series Modelling of Water Resouces and Environmental Systems. Elsevier, Amsterdam.

Horner, I., Renard, B., Le Coz, J., Branger, F., McMillan, H.K., Pierrefeu, G., 2018. Impact of stage measurement errors on streamflow uncertainty. Water Resour. Res. 54 (3), 1952–1976. http://dx.doi.org/10.1002/2017WR022039.

Hyndman, R.J., Athanasopoulos, G., 2018. Forecasting Principles and Practice, second ed. OTexts, p. 382, URL: https://otexts.com/fpp2/.

Hyndman, R.J., Khandakar, Y., 2008. Automatic Time Series Forecasting: The forecast Package for R. J. Stat. Softw. 27 (3), 22, URL: http://www.jstatsoft.org/v27/i03/paper.

Hyndman, R.J., Koehler, A.B., Snyder, R.D., Grose, S., 2002. A state space framework for automatic forecasting using exponential smoothing methods. Int. J. Forecast. 18 (3), 439–454. http://dx.doi.org/10.1016/S0169-2070(01)00110-8.

Iacob, O., Brown, I., Rowan, J., 2017. Natural flood management, land use and climate change trade-offs: the case of Tarland catchment, Scotland. Hydrol. Sci. J. 62 (12), 1931–1948. http://dx.doi.org/10.1080/02626667.2017.1366657.

Isengildina-Massa, O., Irwin, S., Good, D.L., Massa, L., 2011. Empirical confidence intervals for USDA commodity price forecasts. Appl. Econ. 43 (26), 3789–3803. http://dx.doi.org/10.1080/00036841003724429.

Jakeman, A.J., Littlewood, I.G., Whitehead, P.G., 1990. Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments. J. Hydrol. 117 (1–4), 275–300. http://dx.doi.org/10.1016/0022-1694(90)90097-H.

Katimon, A., Shahid, S., Abd Wahab, A.K., Ali, M.H., 2013. Hydrological behaviour of a drained agricultural peat catchment in the tropics. 1: Rainfall, runoff and water table relationships. Hydrol. Sci. J. 58 (6), 1297–1309. http://dx.doi.org/10.1080/02626667.2013.815759.

Keys, T.A., Govenor, H., Jones, C.N., Hession, W.C., Hester, E.T., Scott, D.T., 2018. Effects of large wood on floodplain connectivity in a headwater Mid-Atlantic stream. Ecol. Eng. 118 (March), 134–142. http://dx.doi.org/10.1016/j.ecoleng.2018.05.007.

Kitts, D.R., 2010. The Hydraulic and Hydrological Performance of Large Wood Accumulation in a Low-Order Forest Stream (Ph.D. thesis). University of Southampton, p. p. 368, URL: http://eprints.soton.ac.uk/185791/.

Klaar, M.J., Carver, S., Kay, P., 2020. Land management in a post-Brexit UK: An opportunity for integrated catchment management to deliver multiple benefits? Wiley Interdiscip. Rev. Water 7 (5), 1–6. http://dx.doi.org/10.1002/wat2.1479.

Kwiatkowski, D., Phillips, P.C.B., Schmidt, P., Shin, Y., 1992. Testing the null hypothesis of stationarity against the alternative of a unit root. J. Econometrics 54, 159–178.

Lane, S.N., 2017. Natural flood management. WIREs Water 4 (3), 1–14. http://dx.doi.org/10.1002/wat2.1211, URL: https://onlinelibrary.wiley.com/doi/10.1002/wat2.1211.

Lavers, T., Charlesworth, S.M., Lashford, C., Warwick, F., Fried, J., 2022. The performance of natural flood management at the large catchment-scale: A case study in the Warwickshire Stour Valley. Water 14 (23), http://dx.doi.org/10.3390/w14233836.

Lee, Y.S., Scholtes, S., 2014. Empirical prediction intervals revisited. Int. J. Forecast. 30 (2), 217–234. http://dx.doi.org/10.1016/j.ijforecast.2013.07.018.

Leedal, D., Neal, J., Beven, K., Young, P., Bates, P., 2010. Visualization approaches for communicating real-time flood forecasting level and inundation information. J. Flood Risk Manag. 3 (2), 140–150. http://dx.doi.org/10.1111/j.1753-318X.2010.01063.x.

Lin, J.Y., Cheng, C.T., Chau, K.W., 2006. Using support vector machines for long-term discharge prediction. Hydrol. Sci. J. 51 (4), 599–612. http://dx.doi.org/10.1623/hysj.51.4.599.

Makridakis, S., Winkler, R.L., 1989. Sampling distributions of post-sample forecasting errors. J. R. Stat. Soc. 38 (2), 331–342, URL: https://www.jstor.org/stable/2348063%0AJSTOR.

McCuen, R.H., Knight, Z., Cutter, A.G., 2006. Evaluation of the Nash- Sutcliffe efficiency index. J. Hydrol. Eng. 11 (6), 597–602. http://dx.doi.org/10.1061/(asce)1084-0699(2006)11:6(597).

Met Office, 2018. Summer 2018. Technical Report, Met Office, http://dx.doi.org/10.31390/cwbr.20.3.01.

NASA Shuttle Radar Topography Mission (SRTM), 2013. Shuttle Radar Topography Mission (SRTM) Global. Distributed by OpenTopography. http://dx.doi.org/10.5069/G9445JDF.

National Trust, 2015. From Source to Sea - Natural Flood Management - The Holnicote Experience. Technical Report, National Trust, p. 28.

Nisbet, T., Row, P., Marrington, S., Thomas, H., Broadmeadow, S., Valatin, G., 2015. Project RMP5455: Slowing the Flow at Pickering - Phase 2. Technical Report, Department for Environment, Food & Rural Affairs, p. 32.

Noymanee, J., Theeramunkong, T., 2019. Flood forecasting with machine learning technique on hydrological modeling. Procedia Comput. Sci. 156, 377–386. http://dx.doi.org/10.1016/j.procs.2019.08.214.

O'Driscoll, C., O'Connor, M., Asam, Z., de Eyto, E., Brown, L.E., Xiao, L., 2016. Forest clearfelling effects on dissolved oxygen and metabolism in peatland streams. J. Environ. Manag. 166, 250–259. http://dx.doi.org/10.1016/j.jenvman.2015.10.031, URL: https://linkinghub.elsevier.com/retrieve/pii/S0301479715303297.

Okiy, S., Chukwuemeka Nwobi-Okoye, C., Clement Igboanugo, A., 2015. Transfer function modelling: A literature survey. Res. J. Appl. Sci. Eng. Technol. 11 (11), 1265–1279. http://dx.doi.org/10.19026/rjaset.11.2234, URL: http://maxwellsci.com/jp/mspabstract.php?jid=RJASET{&}doi=rjaset.11.2234.

Ordnance Survey (GB), 2016. OS MasterMap Topography Layer, Scale 1:1000. EDINA Digimap Ordnance Survey Service, URL: http://digimap.edina.ac.uk.

Piotrowski, A.P., Napiorkowski, J.J., 2013. A comparison of methods to avoid overfitting in neural networks training in the case of catchment runoff modelling. J. Hydrol. 476, 97–111. http://dx.doi.org/10.1016/j.jhydrol.2012.10.019.

R Core Team, 2020. A Language and Environment for Statistical Computing, Vol. 2. R Foundation for Statistical Computing, Vienna, Austria, URL: http://www.r-project.org.

Ratto, M., Young, P.C., Romanowicz, R., Pappenberger, F., Saltelli, A., Pagano, A., 2007. Uncertainty, sensitivity analysis and the role of data based mechanistic modeling in hydrology. Hydrol. Earth Syst. Sci. 11 (4), 1249–1266. http://dx.doi.org/10.5194/hess-11-1249-2007.

Rayer, S., Smith, S.K., Tayman, J., 2009. Empirical prediction intervals for county population forecasts. Popul. Res. Policy Rev. 28 (6), 773–793. http://dx.doi.org/10.1007/s11113-009-9128-7.

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40 (8), 913–929. http://dx.doi.org/10.1111/ecog.02881.

Romanowicz, R.J., Young, P.C., Beven, K.J., Pappenberger, F., 2008. A data based mechanistic approach to nonlinear flood routing and adaptive flood level forecasting. Adv. Water Resour. 31 (8), 1048–1056. http://dx.doi.org/10.1016/j.advwatres.2008.04.015.

Rosgen, D.L., 1994. A classification of natural rivers. Catena 22 (3), 169–199. http://dx.doi.org/10.1016/0341-8162(94)90001-9.

Sene, K., Tilford, K., 2004. Review of Transfer Function Modelling for Fluvial Flood Forecasting. Technical Report, Defra/ Environment Agency, Bristol, pp. 1–135, URL: https://www.gov.uk/government/publications/.

Smith, E.P., 2002. BACI design. Encycl. Environ. 1, 141–148. http://dx.doi.org/10.1002/9780470057339.vab001.pub2.

Tencaliec, P., Favre, A.C., Prieur, C., Mathevet, T., 2015. Reconstruction of missing daily streamflow data using dynamic regression models. Water Resour. Res. 51, 9447–9463. http://dx.doi.org/10.1002/2015WR017399.

Thirumalaiah, K., Deo, M.C., 1998. River stage forecasting using artificial neural networks. J. Hydrol. Eng. 3 (1), 26–32. http://dx.doi.org/10.1061/(ASCE)1084-0699(1998)3:1(26), URL: http://ascelibrary.org/doi/10.1061/%28ASCE%291084-0699%281998%293%3A1%2826%29.

Thomas, H., Nisbet, T.R., 2007. An assessment of the impact of floodplain woodland on flood flows. Water Environ. J. 21 (2), 114–126. http://dx.doi.org/10.1111/j.1747-6593.2006.00056.x.

van Leeuwen, Z.R., Klaar, M.J., Smith, M.W., Brown, L.E., 2023. Quantifying the natural flood management potential of leaky dams in upland catchments, Part II: leaky dam impacts on flood peak magnitude. Journal of Hydrology http://dx.doi.org/10.1016/j.jhydrol.2023.130449.

von Asmuth, J.R., Bierkens, M.F., Maas, K., 2002. Transfer function-noise modeling in continuous time using predefined impulse response functions. Water Resour. Res. 38 (12), 23.1–23.12. http://dx.doi.org/10.1029/2001wr001136.

Watson, F., Vertessy, R., McMahon, T., Rhodes, B., Watson, I., 2001. Improved methods to assess water yield changes from paired-catchment studies: Application to the Maroondah catchments. Forest Ecol. Manag. 143 (1–3), 189–204. http://dx.doi.org/10.1016/S0378-1127(00)00517-X.

Waylen, K.A., Holstead, K.L., Colley, K., Hopkins, J., 2018. Challenges to enabling and implementing Natural Flood Management in Scotland. J. Flood Risk Manag. 11, 1078–1089. http://dx.doi.org/10.1111/jfr3.12301.

Wenzel, R., Reinhardt-Imjela, C., Schulte, A., Bölscher, J., 2014. The potential of in-channel large woody debris in transforming discharge hydrographs in headwater areas (Ore Mountains, Southeastern Germany). Ecol. Eng. 71, 1–9. http://dx.doi.org/10.1016/j.ecoleng.2014.07.004.

Wilkinson, M.E., Quinn, P.F., Welton, P., 2010. Runoff management during the September 2008 floods in the Belford catchment, Northumberland. J. Flood Risk Manag. 3 (4), 285–295. http://dx.doi.org/10.1111/j.1753-318X.2010.01078.x.

Williams, W.H., Goodman, M.L., 1971. A simple method for the construction of empirical confidence limits for economic forecasts. J. Amer. Statist. Assoc. 66 (336), 752–754. http://dx.doi.org/10.1080/01621459.1971.10482340.

Wingfield, T., Macdonald, N., Peters, K., Spees, J., Potter, K., 2019. Natural Flood Management: Beyond the evidence debate. Area 51 (4), 743–751. http://dx.doi.org/10.1111/area.12535, URL: http://doi.wiley.com/10.1111/area.12535, https://onlinelibrary.wiley.com/doi/10.1111/area.12535.

Yin, Z., Feng, Q., Wen, X., Deo, R.C., Yang, L., Si, J., He, Z., 2018. Design and evaluation of SVR, MARS and M5Tree models for 1, 2 and 3-day lead time forecasting of river flow data in a semiarid mountainous catchment. Stoch. Environ. Res. Risk Assess. 32 (9), 2457–2476. http://dx.doi.org/10.1007/s00477-018-1585-2.

Yorkshire Dales National Park Authority, Yorkshire Dales Rivers Trust, North Yorkshire City Council, 2017. Natural Flood Management Measures – A Practical Guide for Farmers. Technical Report, Natural England, Environment Agency, http://dx.doi.org/10.1002/wat2.1211, URL: https://thefloodhub.co.uk/wp-content/.

Young, P., 1986. Time series methods and recursive estimation in hydrological systems analysis. In: Kraijenhoff, D., Moll, J. (Eds.), River Flow Modelling and Forecasting. D. Reidel, Dordrecht, pp. 129–180.

Young, P.C., 2002. Advances in real–time flood forecasting. Phil. Trans. R. Soc. A 360, 1433–1450. http://dx.doi.org/10.1098/rsta.2002.1008, URL: https://royalsocietypublishing.org/doi/10.1098/rsta.2002.1008.

Young, P., 2003. Top-down and data-based mechanistic modelling of rainfall-flow dynamics at the catchment scale. Hydrol. Process. 17 (11), 2195–2217. http://dx.doi.org/10.1002/hyp.1328.

Yuan, X., Xie, Z., Liang, M., 2009. Sensitivity of regionalized transfer-function noise models to the input and parameter transfer method. Hydrol. Sci. J. 54 (3), 639–651. http://dx.doi.org/10.1623/hysj.54.3.639.