Hydrol. Earth Syst. Sci. Discuss., 10, 145–187, 2013
www.hydrol-earth-syst-sci-discuss.net/10/145/2013/
doi:10.5194/hessd-10-145-2013

Hydrology and Earth System Sciences Discussions

Legitimising neural network river forecasting models: a new data-driven mechanistic modelling framework

N. J. Mount[1], **C. W. Dawson**[2], **and R. J. Abrahart**[1]

[1]School of Geography, University of Nottingham, Nottingham, NG7 2RD, UK
[2]Department of Computer Science, Loughborough University, Loughborough, LE11 3TU, UK

Received: 11 December 2012 – Accepted: 17 December 2012 – Published: 9 January 2013

Correspondence to: N. J. Mount (nick.mount@nottingham.ac.uk)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

In this paper we address the difficult problem of gaining an internal, mechanistic understanding of a neural network river forecasting (NNRF) model. Neural network models in hydrology have long been criticised for their black-box character, which prohibits adequate understanding of their modelling mechanisms and has limited their broad acceptance by hydrologists. In response, we here present a new, data-driven mechanistic modelling (DDMM) framework that incorporates an evaluation of the legitimacy of a neural network's internal modelling mechanism as a core element in the model development process. The framework is exemplified for two NNRF modelling scenarios, and uses a novel adaptation of first order, partial derivate, relative sensitivity analysis methods as the means by which each model's mechanistic legitimacy is explored. The results demonstrate the limitations of standard, goodness-of-fit validation procedures applied by NNRF modellers, by highlighting how the internal mechanisms of complex models that produce the best fit scores can have much lower legitimacy than simpler counterparts whose scores are only slightly inferior. The study emphasises the urgent need for better mechanistic understanding of neural network-based hydrological models and the further development of methods for elucidating their mechanisms.

1 Introduction

In this paper we consider the complex question of how to determine the mechanistic legitimacy of a black-box, data-driven hydrological model. This question is of fundamental importance for the validation of data-driven models in hydrology and their wider acceptance by hydrologists, yet has to date received very little attention by researchers. In addressing it, we take the relatively simple case of neural network river forecasting models as a starting point, as these models have become one of the most popular application areas for data-driven modelling in hydrology over recent years (Abrahart et al., 2012a). In common with established, statistical river forecasting approaches (e.g. Hipel

HESSD
10, 145–187, 2013

et al., 1977), each neural network river forecaster (NNRF) is a simple, short-step-ahead hydrological forecasting model whose predictions are derived from a core set of lagged, autoregressive model inputs recorded for the point at which the prediction is required (e.g. Firat, 2008), and/or gauged locations upstream. These inputs are, to a varying extent, augmented by a range of relevant, lagged hydrometeorologic variables that act to further refine the model output (e.g. Anctil et al., 2004). Unlike the alternative approaches, that utilise explicit and widely-accepted statistical methods to define and constrain the model structure, the definition of an NNRF model is a largely implicit process in which a neural network (NN) is used to discover the modelling mechanisms directly from the calibration data that are used to train it. The result is a black box model that lacks an explicit documentation of its internal mechanisms.

The main benefit of NNRFs over statistical models is that they have been found to deliver enhanced levels of model fit when assessed against calibration and validation data sets (e.g. Abrahart and See, 2000). Consequently, it has been suggested that NNRFs can deliver forecasts with reduced error, and can be used to extend the horizon over which forecasts can reliably be made (de Vos, 2013). However, these claims should not be taken to mean that NNRF models are more valid than their established, empirical counterparts. Indeed, determining the validity of a model should include a requirement for the developer to establish its legitimacy (Oreskes et al., 1994; Sargent, 2011). In hydrology, one way that this has been addressed is by demonstrating that the behaviour of the internal numerical mechanisms by which the model fit is achieved is sufficiently real in the context of a given modelling task (e.g. Young and Beven, 1994; Beven, 2002). This has always been a problem for data-driven models in general, and NNRFs in particular, because it is not easy to gain an explicit representation of their internal numerical mechanisms from which the reality can be assessed. Indeed, even when the explicit governing equations of an NN are examined their complexity prohibits straightforward interpretation (Aytek et al., 2008; Abrahart et al., 2009). It is, therefore, difficult to provide a convincing justification that NNRFs behave in a manner that conforms to hydrological domain knowledge (Cunge, 2003; de Vos and Rientjes, 2005). It

HESSD
10, 145–187, 2013

also means that NNRFs have little heuristic value because the extent and legitimacy of hydrological process representation within the model's response function has not been sufficiently established. Similarly, generalising NNRFs is difficult as there is no means of demonstrating whether the modelling mechanisms employed adequately encapsulate the hydrological processes that influence catchment responses. Consequently, NNRFs have been criticised as offering hydrological researchers little more than an advanced method of non-linear curve-fitting (Abrahart et al., 2011), whose high degree of performance in replicating calibration and validation data is an inevitable consequence of the low numerical complexity of the underlying curve fitting operation required in river forecasting models (Mount and Abrahart, 2011a).

The inability to legitimise the internal behaviour of an NNRF's numerical mechanisms also has important methodological implications for the model development process. Developing an NNRF is complicated by the wide range of structural and training parameters that must be configured (e.g. learning function, constraints, number of hidden layers, number of hidden units, stopping point, etcetera), and a general lack of consensus about what the most appropriate configurations are for a given modelling task (Jayawardena and Fernando, 2001). Consequently, it is necessary for the modeller to determine the best configuration for their model via empirical means (e.g. Kisi, 2004). Standard methods apply a trial-and-error approach in which the parameter configuration is systematically altered to deliver a set of candidate models; each model containing slightly different internal numerical mechanisms and delivering slightly different outputs (e.g. Kisi, 2008). The "best" model is then selected by quantifying its fit to calibration and validation data in a process that seeks the optimum combination of NN components and parameters irrespective of any mechanistic understanding of the candidate models (Mount and Abrahart, 2011b). The result is that the NNRF model selection process arguably fails to deliver an adequate assessment of whether mechanisms employed by candidate solutions with lower levels of fit might actually be more valid or legitimate from the perspective of relevant domain knowledge.

HESSD
10, 145–187, 2013

In response to this issue, a limited number of attempts to provide a hydrologic explanation of NNRF numerical mechanisms can be identified in the literature. The majority have focussed on isolating and interpreting the outputs associated with different structural elements of the NN; specifically weights (e.g. Prada-Sarmiento and Obregon-Neira, 2009) and hidden units. Evaluation of hidden units has arguably delivered the greatest insights, with studies by Wilby et al. (1993), Jain et al. (2004), See et al. (2008), Fernando and Shamseldin (2009) and Jain and Kumar (2009) all indicating that different hidden units are responsible for modelling different components of the flood hydrograph (i.e. rapid overland flow, interflow and baseflow). Similar results were reported by Sudheer and Jain (2004), who found that their NNRF mapped a function that approximated a flow duration curve, with the different hidden units capturing low, medium and high magnitude flows. These studies offer important insights into the ways in which NNRFs partition their overall response function into defined components within the NN structure. However, the primary focus in these studies is the delivery of a structural assessment and the extent to which they focus on the numerical mechanisms that the structures deliver is limited.

If the criticisms surrounding NNRF model legitimacy are to be fully addressed it is necessary to move beyond investigations of NN structure, and towards the development of techniques that can elucidate the behaviour of the NN response function in a more mechanistic sense. In this context, previous papers on the development of data-based mechanistic (DBM) approaches to hydrological model development are of particular relevance (Young and Beven, 1994) as they offer a recognised means by which mechanistic legitimacy can be delivered to black-box hydrological models. The term mechanistic simply refers to the structure and interactions of the internal mechanisms that control a model's behaviour. In the DBM approach, these are examined using a formal process of statistical inference through which the modelling mechanisms are identified prior to building the model, and interpreted according to the extent to which they conform to the nature of the system under study (Young et al., 2004) (Fig. 1, A1–A4). The model is then accepted, or rejected, on the basis of its conformance.

HESSD
10, 145–187, 2013

The application of the DBM approach in data-driven modelling in general, and NNRF modelling in particular, is prevented due to the means by which the model learns from the data; limiting the a priori application of statistical inference from which a mechanistic interpretation can be made. The DBM process can, however, be reordered to address this issue by examining modelling mechanisms and assessing their legitimacy immediately after model development (Fig. 1, B1–B4). This results in a new, generic, data-driven mechanistic modelling (DDMM) framework that includes a specific requirement for mechanistic analysis and assessment as a post model building activity. In this adaptation, the DDMM framework is more loosely defined and need not necessarily be constrained to a demonstration of adequate representation of a natural system by a model, which is a key feature of the DBM approaches. Indeed, it may also be used as a tool to direct broader mechanistic investigations, including the complexity and functionality of the internal workings of a model, and the extent to which these can be justified by the modelling task.

In this study we show how our DDMM framework can be applied to NNRF models so that the evaluation of different candidate models includes an appraisal of both their performance and mechanistic legitimacy. In contrast to previous studies (e.g. Coulibaly et al., 2000; Huang et al., 2004; Kisi and Cigizoglu, 2007; Kisi, 2008), the primary objective of this paper not to demonstrate the best possible river forecast that can be achieved using NNRF techniques. Indeed, we restrict our modelling to only simple examples that we accept may not be optimal. Instead, its objective is to exemplify how the application of input sensitivity analysis, delivered within the DDMM framework, provides an important new means by which NNRF modellers can identify the most legitimate model mechanisms in a set of candidate models. Thus, we here apply two of the most commonly used, and simple, NNRF input scenarios for short step-ahead discharge forecasting on the River Ouse, Yorkshire, UK: antecedent, at-gauge inputs and antecedent upstream inputs. The candidate NNRF models that are developed are then evaluated within the DDMM framework according to their external fit metrics and internal mechanisms.

HESSD
10, 145–187, 2013

2 Elucidating NNRF model mechanisms

The DDMM framework potentially offers both NNRF modellers, and data-driven modellers more generally, a more rigorous and defensible approach to model development. However, its successful application is reliant on the availability of techniques by which data-driven model mechanisms can be elucidated and interpreted. These are not generally well developed or widely applied at present. Nonetheless, recent developments in techniques for delivering sensitivity analyses for NN models (Yeung et al., 2010) have provided new opportunities for exploring their internal mechanisms. In this context, parameter sensitivity analysis (Hamby, 1994) is of particular interest as it offers an important and established approach for elucidating numerical modelling mechanisms, as well as delivering improved model verification (Howes and Anderson, 1988) and validation (e.g. Kleijnen, 1995; Kleijnen and Sargent, 2000; Fraedrich and Goldberg, 2000; Smith et al., 2008; Mishra, 2009). Indeed, it is recognised as an important means by which model validation can be extended beyond fit, to include deeper insights into a model's mechanistic behaviours (e.g. Sun et al., 2009).

Existing attempts to explore the internal numerical mechanisms of NNRFs via sensitivity analysis are restricted to a single study by Sudheer (2005). This study applied an indirect, perturbation analysis in which each input series to the NNRF was systematically varied by a pre-determined proportion, with all other input series remaining unaltered. The impact of the input variation on the output magnitude was used to identify those inputs that had the greatest influence on the model. However, temporal dependencies in the data that drive river forecasting models should act to limit the perturbation ranges that can realistically be applied (Abrahart et al., 2012b), resulting in very small perturbations that limit the appropriateness of the perturbation method. Input sensitivity analysis approaches that are based on the quantification of a model's partial derivatives offer an alternative and more direct means of mechanistic examination. These express the ratio of change between a model's inputs and outputs directly

HESSD
10, 145–187, 2013

from the equations that define the model, and thus offer an important means by which internal modelling mechanisms can be inferred.

Computational techniques for determining first order partial derivatives of certain NNs have been developed. One such technique, outlined by Hashem (1992), involves the application of a simple backward chaining partial differentiation rule. His general rule is adapted in Eq. (1) for NNs with sigmoid activation functions, a single hidden layer, i input units, n hidden units and one output unit (O), so that the partial derivative of the network's output can be calculated with respect to each of its inputs:

$$\frac{\partial O}{\partial I_i} = \sum_{j=1}^{n} w_{ij} w_{jO} h_j \left(1 - h_j\right) O(1 - O) \quad (1)$$

where, w_{ij} is the weight from input unit i to hidden unit j; w_{jO} is the weight from hidden unit j to the output unit O; h_j is the output of hidden unit j; and O is the output from the network.

The relative sensitivity of each input, i, is thus calculated as:

$$\frac{\partial O}{\partial I_i} \cdot \frac{I_i}{O} = \sum_{j=1}^{n} w_{ij} w_{jO} h_j \left(1 - h_j\right) O(1 - O) \cdot \frac{I_i}{O} = (1 - O) I_i \sum_{j=1}^{n} w_{ij} w_{jO} h_j \left(1 - h_j\right). \quad (2)$$

However, simply using first order partial derivative approaches does not overcome the challenges presented by temporal-dependence in the input data that drive NNRFs. Standard, local scale sensitivity analysis techniques (e.g. Turanayi and Rabitz, 2000; Spruill et al., 2000; Holvoet et al., 2005; Hill and Tiedeman, 2007) require the establishment of a representative base case (Krieger et al., 1997) for all inputs. This is usually defined according to their mean or median values on the assumption that all inputs are independent of one another. However, in NNRF modelling this assumption is not valid and the identification of a representative base case is very difficult. Moreover, local scale analyses can only provide mechanistic insights for the specific location in the

HESSD
10, 145–187, 2013

input hyperspace to which the base case corresponds, and it should not be assumed that mechanistic insights can be generalised beyond it (Helton, 1993).

The application of a global (Muleta and Nicklow, 2005; Salteli et al., 2008) or regional (e.g. Spear and Hornberger, 1980; Beven and Binley, 1992) sensitivity analysis can overcome this issue by delivering a generalised sensitivity index, which incorporates input probability distributions that describe all of the input hyperspace, or specific regions within it. However, these methods are very dependent on the particular method used to sample and compute the distributions (Pappenberger et al., 2008), and strong temporal dependence in NNRF inputs makes the determination of an appropriate sampling strategy problematic. In addition, the summary, lumped indices output by global and regional techniques mask the detailed, local patterns of input-output sensitivity that must be understood in order to fully characterise a model's mechanistic behaviour.

One solution for overcoming these difficulties is to adopt a brute-force approach in which relative first order, partial derivatives for all model inputs are computed separately for every data point in a given time series, using the specific input values recorded at each point as a datum-specific base case. In this way, a "global-local" parameter sensitivity analysis is developed in which local scale input sensitivity analysis is performed across the global set of available data points. Issues associated with temporal dependence in river forecasting data are overcome because every datum in the analysis effectively becomes its own, specific base case. NNRF modelling mechanisms can then be characterised and interpreted across the full forecast range by plotting the relative sensitivity of each input (y-axis) against the forecast values delivered by the model (x-axis), and interpreting the patterns that can be observed in the plots.

These patterns can be interpreted mechanistically, by considering the magnitude, stability, continuity and coherency of the model's sensitivity to each input, the relative differences in these patterns for different inputs, and how they vary across the model's forecast range (Fig. 2). The magnitude of a model's sensitivity to its inputs characterises the relative extent to which each model forecast is sensitive to variation in each of its inputs. It can therefore reveal the relative importance of each input as a driver of

HESSD
10, 145–187, 2013

the model output at any given point in the forecast range. The stability of the input sensitivity characterises the consistency with which each input influences the model output across different forecast ranges. Invariance in an input's relative sensitivity across the entire range (the most stable case) indicates that it is being used as a constant multiplier by the model's internal mechanism. Lower levels of stability will indicate increasingly non-linear influences. The existence of local discontinuities in the model's sensitivity to an input indicates the existence of thresholds in the model's mechanisms that may result in distinctly different internal mechanistic behaviour at neighbouring locations in the forecast range. Coherency reflects the extent to which the model's sensitivity to its inputs varies from point to point. Low coherence is indicative of a model that applies a distinctly different modelling mechanism to each local data point and is a means by which data overfitting may be detected.

3 Exemplifying the DDMM framework: study area, datasets and modelling scenarios

In the following sections we present a simple example study to show how global-local parameter sensitivity analysis can be used, in the context of the DDMM framework, to assess the legitimacy of two simple NNRF models, developed for the River Ouse at Skelton, Yorkshire, UK. The complexity of the modelling tasks is purposefully minimised to ensure that the parameter sensitivity patterns delivered are relatively straightforward, and the characterisation and interpretation of modelling mechanisms can be clearly understood. Therefore, we restrict our model input parameters to temporally-lagged instances of the model output; accepting that alternative input configurations may be able to deliver models with a higher degree of fit. The first NNRF model (Scenario A) represents the most simplistic, autoregressive river forecasting case, in which at-a-gauge discharge is forecast from lagged discharge inputs recorded at the same location. The second, more complex, NNRF model (Scenario B) predicts at-a-gauge

HESSD
10, 145–187, 2013

discharge from a set of three lagged discharge inputs recorded at gauges located in tributary rivers immediately upstream.

The catchment upstream of the Skelton gauge (Fig. 3) covers a catchment of 3315 km^2 with a maximum drainage path length of 149.96 km, and an annual rainfall of 900 mm. The catchment contains mainly rural land uses with < 2 % urban land cover. It exhibits significant areas of steep, mountainous uplands that extend over 12 % of the catchment, and includes the three sub-catchments of the rivers Swale, Ure and Nidd. Each of these tributaries is gauged in its lowland reaches, upstream of its confluence with the Ouse. Details of these gauges and contributing catchments are provided in Table 1.

All NNRF models were developed using daily mean discharge records, downloaded from the Centre for Ecology and Hydrology National River Flow Archive (www.ceh.ac.uk/data/nrfa). The data extend over a period of 30 yr, from 1 January 1980 to 31 December 2010 (Fig. 4). Several short gaps exist in the observed records at irregular periods across the different stations; necessitating approximately 8 % of the 30-yr record to be omitted due to missing records at one or more gauges.

The data were partitioned so that the first 75 % of the available record (7762 data points) was used for model calibration, leaving 25 % (2588 data points) for use as an independent validation data set. The split places the three unusually high-magnitude flood peaks observed at Skelton (identified by the arrows in Fig. 4) in the calibration data. This is important in the context of this study, as it ensures that the internal mechanisms of the calibrated models have been developed to accommodate the largest observed floods in the record. Therefore, any mechanistic interpretation is informative across the full forecast range for each model. Nonetheless, we also recognise that the simplicity of this splitting procedure contrasts with more complex approaches that have been used by NN modellers (e.g. Snee, 1977; Baxter et al., 2000; Wu et al., 2012) to deliver improved validation consistency (LeBaron and Weigend, 1998) by ensuring representative subsetting procedures. Therefore, exceedence curves for the calibration

HESSD
10, 145–187, 2013

and validation data (Fig. 5), were checked to ensure high conformance in the discharge probability distributions for calibration and validation data subsets at all gauges.

Input selection and model development

Scenario A is a straightforward, autoregressive NNRF for Skelton that predicts instantaneous discharge (S_t) from the three most recently gauged discharges (S_{t-1}; S_{t-2}; S_{t-3}). The modelling is developed directly from the daily mean discharge record for Skelton, with no pre-processing having been applied. Three antecedent predictors were used, such lags having the strongest correlation with observed flow at Skelton at time t (Fig. 6) over the entire 30-yr record. Scenario B predicts S_t on the basis of antecedent discharges recorded for the three tributary gauges at Crakehill (C), Skip Bridge (SB) and Westwick (W). The strength of the correlation between each tributary gauge and Skelton over a range of lags was used to determine the lag time at each tributary that represented the strongest predictor for S_t. The three inputs to Scenario B are thus C_{t-1}; SB_{t-1}; and W_{t-1}.

The proportion of the discharge at S_t that is accounted for by discharge at C_{t-1}, SB_{t-1} and W_{t-1} is summarised as a box plot in Fig. 7. The plot shows that, summarised over the whole record, lagged discharge at Crakehill and Westwick accounts for a similar proportion of the instantaneous discharge at Skelton, with comparable median values (~ 40 %) and interquartile ranges. Skip Bridge is proportionally less important with a median value of 18 %. This highlights its relative weakness as a physical driver of S_t, which is in contrast to its relative strength as a statistical driver (i.e. it has the second highest correlation coefficient at $t - 1$). It should be noted that, due to timing effects and the use of summary, daily mean data, the maximum proportional contributions values in Fig. 7 exceed 100 %.

In order to reflect the lack of consensus surrounding NNRF parameterisation, and the empirical process that underpins model selection in the majority of previous studies (see introduction), four candidate single-hidden-unit NNs were developed for Scenarios A and B. Each candidate was structurally-distinct, incorporating either 2, 3, 4 or

HESSD
10, 145–187, 2013

5 hidden units. In this way, a range of alternative candidate models of varying complexity were developed in each NNRF scenario for subsequent mechanistic comparison. All candidate models used the back-propagation of error learning algorithm (Rummelhart et al., 1986) and root mean squared error (RMSE) as the objective function. The training rate was fixed at 0.1 and a momentum term of 0.9 was used. Each of the candidate models was trained for epochs ranging from 100 to 20 000 iterations, and tested at 100 epoch intervals, with the optimum number of epochs identified by means of cross-validation with the independent dataset, according to the lowest RMSE value obtained. The preferred number of epochs for each hidden unit configuration for the different scenarios is shown in Table 2, with the relative strength of the autoregressive relationship in Scenario A reflected in its lower number of training epochs. Similarly, the relative simplicity of the NN configurations comprising fewer hidden units is reflected in their generally lower number of training epochs. For a simple, linear benchmark (Abrahart and See, 2007; Mount and Abrahart, 2011a), multiple linear regressions models (MLR) were also developed on the calibration data for both scenarios. Their equations are:

$$\text{Scenario A}: S_t = 6.014 + 1.12 \cdot S_{t-1} + 0.455 \cdot S_{t-2} + 0.216 \cdot S_{t-3} \quad (3)$$

$$\text{Scenario B}: S_t = 5.715 + 0.424 \cdot C_{t-1} + 1.556 \cdot SB_{t-1} + 1.055 \cdot W_{t-1}. \quad (4)$$

4 Scenario A: performance, mechanistic interpretation and model choice

4.1 Candidate model fit

The calibration and validation performance of each candidate NNRF model, driven by autoregressive inputs, are presented in Tables 3 and 4. A wide range of metrics have been proposed for assessing hydrological model performance (Dawson et al., 2007, 2010), along with a range of mechanisms for their integration (e.g. Dawson et al., 2012). Nonetheless, consensus has still to be achieved on the metrics that should be used in assessing NNRF model performance. Here we restrict our metrics to three simple and

HESSD
10, 145–187, 2013

widely used examples that cover key aspects of model fit. This restriction is justified on the basis that the mechanistic exploration delivered by the DDMM framework reduces the overall reliance on metric-based assessment and the importance of arguments that surround the subtleties of metric choice in model assessment. Pearson's product moment correlation coefficient, squared (R-squared) is included as a general, dimensionless measure of model fit that indicates the proportion of overall variance in the data that is explained by each candidate model. RMSE is included because it is a metric that is disproportionately influenced by the extent to which each candidate model forecasts high-magnitude discharges. In contrast, the relative metric mean squared relative error (MSRE) is included because its scores emphasises the extent to which low-magnitude discharges are correctly forecast by the candidates. The formula for each metric can be found in Dawson et al. (2007).

The metric scores highlight almost identical levels of performance across the candidates, irrespective of the metric against which fit is assessed, or whether the fit is assessed relative to the calibration or validation data. Metric scores for the validation data are slightly better than those for the calibration data in all metrics, with the greatest differences observed in RMSE scores. This reflects the fact that the three highest magnitude floods are within the calibration data and, in common with most other autoregressive river forecasting models, there is a general underestimation of flood peaks. These two aspects combine to produce the observed improvement in RMSE in the validation data. Importantly, the MLR benchmark performs well, with RMSE and R-squared scores that are comparable with the NNRF model candidates for the calibration data and better for the validation data. This serves to highlight the near-linear nature of the modelling problem. Despite there being no clear winner on the basis of metrics alone, the 5-hidden-unit model does achieve the best NNRF candidate metric scores in three out of six cases.

HESSD
10, 145–187, 2013

4.2 Candidate model mechanisms

For each of the four candidate solutions, relative first order partial derivatives were computed according to the global-local approach outlined in Sect. 2. Equation (2) was used to compute local first order partial derivatives for all 7762 data points, on which the candidate models were calibrated. Values of w_{ij}, $w_{j\mathrm{O}}$, h_j were determined for each forecast, according to its specific input value set at each point. This resulted in a set of 7762 separate relative parameter sensitivity values for each of the three inputs in each candidate model. These values are plotted against their respective forecasted discharge values in Fig. 8.

Figure 8 highlights the fact that, mechanistically, all four candidate models behave in very similar ways. In all cases, the relative sensitivity of the model forecast to variation in S_{t-1} is substantially greater than to either S_{t-2} or S_{t-3}; indicating its primary importance as the driver of the model forecasts. This result is entirely in line with expectations of a simple autoregressive model. Indeed, the overriding importance of S_{t-1} is further highlighted by the opposing directionality in the generally low-magnitude, relative sensitivities associated with S_{t-2} and S_{t-3}. This pattern indicates the existence of internal NN mechanisms that largely cancel out the influence of these variables; resulting in a modelling mechanism with redundant complexity. This mechanism can be observed, to varying extents, in all candidate models, suggesting a mis-match between the scope of the modelling problem and the complexity of technique by which is has been solved. The MLR equations and performance metrics further support this view, with the coefficients for S_{t-2} and S_{t-3} being substantially smaller than for S_{t-1} and the good metric scores for the calibration and validation data (Table 4) highlighting the near-linear nature of the modelling problem. Nonetheless, moderate instability in the relative sensitivity of all candidate models to S_{t-1} is evident, with a consistent pattern that approximates a third order polynomial. This indicates some non-linearity in the modelling mechanism associated with S_{t-1}, although this non-linearity results in little if any performance gain over the linear, MLR benchmark.

HESSD
10, 145–187, 2013

One characteristic by which the candidate modelling mechanisms can be more clearly discerned from one another is their coherency, with different candidates displaying varying degrees of scatter in their relative sensitivity plots. Of particular note is a moderate reduction in the coherency of the relative sensitivity plots for S_{t-1} and S_{t-2} as the number of hidden units in the candidate models increases; with lower coherency indicating an internal modelling mechanism that is increasingly data point specific (i.e. is tending towards overfitting the data). As S_{t-1} is the main driver of the forecast discharge across all candidates, high coherency in the relative sensitivity of the model to this input is desirable; suggesting that the highest level of mechanistic legitimacy can be argued for the 2-hidden-unit candidate model.

4.3 Model selection

The simplistic, near-linear forecasting challenge presented by this scenario has, unsurprisingly, resulted in similarity across the candidate models, in terms of both their performance and internal mechanisms. Indeed, the lack of clear differentiation between each candidate model's metric score performance would suggest that any of the candidates might be reasonably chosen. However, the selection of the most parsimonious model is usually preferable (Dawson et al., 2006), especially for simple modelling problems. Therefore, in the absence of conclusive metrics-based evidence, selection of the 2-hidden-unit NNRF model could be argued as the most appropriate. Examination of the internal mechanisms adds additional evidence to support this choice. Although there is little evidence by which the candidates can be distinguished with respect to mechanistic stability or consistency, the 2-hidden-unit model displays a greater degree of coherency in its key driver (S_{t-1}) than its counterparts. This delivers additional, mechanistic support for its preferential selection. However, the high degree of redundancy observed in all candidate model mechanisms raises important questions about the appropriateness of using a NNRF model for such a simple modelling task at all and the number of inputs included. Indeed, the mechanistic evidence corresponds with previous criticisms (e.g. Mount and Abrahart, 2011a), which argue that, in most cases,

HESSD
10, 145–187, 2013

standard MLR-based methods can offer a more appropriate means for simple step-ahead river forecasting tasks.

5 Scenario B: performance, mechanistic interpretation and model choice

5.1 Candidate model fit

Calibration and validation performance for the four candidate NNRF models, driven by upstream inputs, are presented in Tables 5 and 6. The metric scores for Scenario B provide limited evidence by which to discern the relative validity of the candidate models, with all candidates again delivering similar metric scores. However, in contrast to Scenario A, one candidate consistently achieves the best metric scores. The 5-hidden-unit candidate achieves the best metric scores for two of the three calibration metrics, and all validation metrics. On this basis, its preferential selection could be argued, and this selection would be in line with previously-published data-driven modelling studies in which candidate model preference has been determined on the basis of consistent, best metric scores that represent relatively small overall performance gains (Kisi and Cigizoglu, 2007). It should also be noted that, in this scenario, the performance of all NNRF candidates exceed that of the MLR benchmark; highlighting the importance of non-linearity associated with river forecasting based on upstream inputs.

5.2 Candidate model mechanisms

Global-local relative sensitivity plots for each upstream input used in each candidate model are presented in Fig. 9. W_{t-1} is the strongest driver of S_t, particularly at low forecast ranges, with moderate sensitivity to SB_{t-1} also being evident. A clear mechanistic distinction between the 2 and 3-hidden-unit candidates and their 4 and 5-hidden-unit counterparts can be observed based on the coherency of their mechanisms. The 4 and 5-hidden-unit candidates display low coherency, particularly at moderate to high forecast ranges, and this is particularly evident for inputs C_{t-1} and W_{t-1}. This suggests

HESSD
10, 145–187, 2013

that modelling mechanisms in the more complex candidates may be overfitting the upper-range data; a tendency that is well known when NN-based hydrological models are used to fit heteroscedastic data (Mount and Abrahart, 2011b). The importance of avoiding overfitting in NN models is well known (Guistolisi and Lauocelli, 2005), and the lack of coherency in the 4 and 5-hidden-unit candidates thus raises concerns over their mechanistic legitimacy.

Low sensitivity to variation in the discharge at C_{t-1} is a particular feature of the 2 and 3-hidden-unit candidates. This pattern contrasts with the MLR coefficients (Equation 4) that highlight SB_{t-1} as the strongest model driver in the regression model. It also contrasts with the proportional contribution that each lagged, upstream discharge makes to overall discharge at S_t (Fig. 7). Indeed, the significant proportional contribution made by C_{t-1} is minimised by the candidates – a factor that highlights the signal-based, rather than physically-based nature of their modelling mechanisms. Reduction in the relative sensitivity to SB_{t-1} and W_{t-1} as the forecast range increases is evident in both the 2 and 3-hidden-unit candidates, and highlights the presence of non-linearity in the modelling mechanism. The high degree of stability in these plots is indicative of relative low-complexity in the non-linearity mechanism.

In differentiating the mechanistic legitimacy of these two candidates, however, the relative sensitivity plots for C_{t-1} and SB_{t-1} are of particular interest. The increase from 2 to 3-hidden-units is accompanied by a moderate reduction in the coherency of the relative sensitivity to SB_{t-1} at medium forecast ranges, and the existence of some negative values. To some extent, these negative sensitivity values are counteracted by slightly higher positive sensitivity to C_{t-1} at similar forecast ranges. Nonetheless, in the context of an upstream river forecasting model, it is difficult to justify a modelling mechanism that acts to reduce downstream discharge forecasts as discharge increases upstream. Consequently, the legitimacy of the 3-hidden-unit candidate is difficult to argue. Indeed, the 2-hidden-unit candidate appears to have the greatest mechanistic legitimacy of the candidates; combining high coherency and appropriate stability in its

HESSD
10, 145–187, 2013

relative sensitivity to its inputs; albeit with the predictive power of C_{t-1} minimised to near-zero.

5.3 Model selection

Scenario B represents a situation in which the fit metrics associated with different candidate models provide only limited evidence to inform model selection. On the basis of fit metrics alone, the 5-hidden-unit model appears to offer the best modelling solution as it consistently has the best scores. However, the actual performance gains are small, questioning whether a simpler model with only marginally lower performance might actually be preferable. Indeed, examination of the 5-hidden-unit candidate's internal mechanism reveals low coherency that is very difficult to legitimise over its more coherent counterparts. Taking into account both fit metric scores and the legitimacy of internal mechanisms, the 2-hidden-unit candidate offers the best overall modelling solution. It combines high coherency and an appropriate degree of stability in its modelling mechanisms, with fit metric scores that are only fractionally lower than the best performing 5-hidden-unit candidate.

6 Summary

The analysis presented demonstrates that fit metric scores alone are an insufficient basis by which to assess and discriminate between different NNRF models. The high degree of equifinality in the metric scores for the candidate models masks important differences in their complexity, mechanistic behaviour and legitimacy, which is only exposed when internal modelling mechanisms are explored. The importance of a mechanistic evaluation is particularly evident for Scenario B, where small improvements in metrics are associated with a substantial reduction in mechanistic legitimacy. Thus, the study responds to the issue of whether the end point of a model (i.e. its fit) is a sufficient basis by which to justify its means (i.e. the numerical basis by which the fit is

HESSD
10, 145–187, 2013

achieved). This question remains a vital one for all hydrological modellers. To a large extent, the scope and objectives of a hydrological model will determine the relative emphasis that should be placed on its mechanistic and performance validation (Jakeman et al., 2006). However, if these are to exceed basic data-specific curve-fitting tasks, some assessment of the mechanistic-legitimacy of the model is required. In the last two decades of NN-based hydrological modelling, relatively little research effort has been directed towards the development of methods for their mechanistic interpretation, explanation and/or evaluation. This is despite recognition that the lack of availability of such methods has been a fundamental constraint to progress in the field over the last 20 yr (Abrahart el al., 2012).

This lack of progress is in stark contrast to the advances made by physical and conceptual modellers that have centred on the development of new model evaluation methods that incorporate mechanistic insights into model behaviour and uncertainty (e.g. Beven and Binley, 1992). As a result, NN modelling in particular, and data-driven modelling more generally, has remained a niche area of hydrological research that has had only limited success in convincing the wider hydrological research community of its potential value beyond optimised curve fitting tasks. The DDMM framework and methodological approach that has been developed in this study represents an important early step in redressing this problem. The inclusion of a requirement for the elucidation and assessment of modelling mechanisms within the NN model development process ensures that the validation of any NN model makes explicit both the performance of the model, and the legitimacy of the means by which it is achieved. This opens up the possibility of developing NN hydrological models with greater heuristic value and transferability. By adapting a partial derivative sensitivity analysis method as the means by which this is done, we parallel existing approaches for mechanistic model exploration that are long-standing and well-established within hydrology (cf. McCuen, 1973). In so doing we increase the alignment between NN model development methodologies and those applied during the development of their conceptual and physical counterparts: an outcome that should lead to their wider acceptance.

HESSD
10, 145–187, 2013

The input scenarios that we have used to exemplify the DDMM approach in this paper are more simplistic than those used in many NNRF models that include an additional array of hydro-meteorological inputs with varying degrees of temporal dependence (cf. Zealand et al., 1999; Dibike and Solomatine, 2001). Similarly, the application of a standard, back-propagation algorithm is not fully representative of the wide range of NN variants that have been explored in NNRF modelling studies (cf. Hu et al., 2001; Shamseldin and O'Connor, 2001). Consequently, the relative ease with which we have been able to quantify and interpret input relative sensitivity in this study may not be mirrored in more complex studies that use an increased number and diversity of inputs and/or variants of the standard back-propagation algorithm. Thus, developing techniques that can deliver clear mechanistic interpretation of input relative sensitivity patterns in more complex NN modelling scenarios repesents an important consideration for future research. Nonetheless, the results we present serve as a clear demonstration of the dangers associated with evaluating NN models on the basis of performance validation approaches alone. Indeed, in our examples we are able to show that, in order to achieve moderate performance gains, the mechanistic legitimacy of the candidate NNRF models may be substantially reduced. This finding is particularly clear in Scenario B. This finding has important implications for previous river forecasting studies that have concluded that NNRF models offer benefits over other, established techniques based on limited performance gains. Indeed, an argument could be made for revisiting both previous NNRF studies and NN-based hydrological models more generally, to determine the extent to which their enhanced levels of performance validation are matched by their levels of mechanistic legitimacy.

7 Conclusions

This paper has argued that gaining an understanding of the internal mechanisms by which a hydrological model generates its forecasts is an important element of the model development process. It has also argued that the development of methods for delivering

HESSD
10, 145–187, 2013

mechanistic insights into NNRF models in particular, and data-driven hydrological models more generally, have not been afforded sufficient attention by researchers. As a result, "black-box" criticisms associated with NNRF models persist, and they remain a dataset-specific, non-transferrable curve-fitting optimisation technique that is delivering only limited heuristic knowledge to the hydrological community. This limitation is one of several problems that must be overcome if wider acceptance of NNRF models by hydrologists is to be achieved (for a discussion see Tsai et al., 2013).

This study represents an important step in addressing these limitations by shifting the focus of an NNRF model from its performance to its internal mechanisms. We have presented a generalised framework that explicitly includes a mechanistic evaluation of NNRF models as a part of the model development process. The framework comprises a set of high-level model development and evaluation procedures into which specific NN model development methodologies can be positioned. Through the development and application of a brute-force, global-local relative sensitivity analysis, we have overcome difficulties associated with quantifying relative sensitivity across a model's full forecast range, when the model inputs are temporally-dependent. Our adaptation of partial derivative input sensitivity analyses as a means of examining the mechanistic behaviour of a model during its construction, is reflective of long-established uses of sensitivity analyses for the mechanistic examination of hydrological models during their development (e.g. McCuen, 1973). To an extent, this contrasts with current advances in hydrological modelling that use sensitivity analyses as a means of examining the causes and impacts of uncertainty in the outputs of existing models (e.g. Pappenberger et al., 2008). Nonetheless, it serves as a useful reminder of its importance as an established and proven tool during hydrological model construction.

HESSD
10, 145–187, 2013

References

Abrahart, R. J. and See, L. M.: Neural network modelling of non-linear hydrological relationships, Hydrol. Earth Syst. Sci., 11, 1563–1579, doi:10.5194/hess-11-1563-2007, 2007.

Abrahart, R. J., Ab Ghani, N., and Swan, J.: Discussion of "An explicit neural network formulation for evapotranspiration", Hydrolog. Sci. J., 54, 382–388, 2009.

Abrahart, R. J., Mount, N. J., Ab Ghani, N., Clifford, N. J., and Dawson, C. W.: DAMP: a protocol for contextualising goodness-of-fit statistics in sediment-discharge data-driven modelling, J. Hydrol., 409, 596–611, 2011.

Abrahart, R. J., Anctil, F., Coulibaly, P., Dawson, C. W., Mount, N. J., See, L. M., Shamseldin, A. Y., Solomatine, D. P., Toth, E., and Wilby, R. L.: Two decades of anarchy? Emerging themes and outstanding challenges for neural network modelling of surface hydrology, Prog. Phys. Geogr., 36, 480–513, 2012a.

Abrahart, R. J., Dawson, C. W., and Mount, N. J.: Partial derivative sensitivity analysis applied to autoregressive neural network river forecasting, in: Proceedings of the 10th International Conference on Hydroinformatics, 14–18 July 2012, Hamburg, Germany, 8 pp., 2012b.

Anctil, F., Michel, C., Perrin, C., and Andreassian, V.: A soil moisture index as an auxiliary ANN input for stream flow forecasting, J. Hydrol., 286, 155–167, 2004.

Aytek, A., Guven, A., Yuce, M. I., and Aksoy, H.: An explicit neural network formulation for evapotranspiration, Hydrolog. Sci. J., 53, 893–904, 2008.

Baxter, C. W., Stanley, S. J., Zhang, Q., and Smith, D. W.: Developing artificial neural network process models: A guide for drinking water utilities, in: Proceedings of the 6th Environmental Engineering Society Specialty Conference of the CSCE, London, 376–383, 2000.

Beven, K. J.: Towards a coherent philosophy for modelling the environment, Philos. T. Roy. Soc. Lond. A, 458, 2465–2484, 2002.

Beven, K. J. and Binley, A.: The future of distributed models: model calibration and uncertainty prediction, Hydrol. Process., 6, 279–298, 1992.

Coulibaly, P., Anctil, F., and Bobe, B.: Daily reservoir inflow forecasting using artificial neural networks with stopped training approach, J. Hydrol., 230, 244–257, 2000.

Cunge, J. A.: Of data and models, J. Hydroinform., 5, 75–98, 2003.

Dawson, C. W., Abrahart, R. J., Shamseldin, A. Y., and Wilby, R. L.: Flood estimation at ungauged sites using artificial neural networks, J. Hydrol., 319, 391–409, 2006.

HESSD
10, 145–187, 2013

Dawson, C. W., Abrahart, R. J., and See, L. M.: HydroTest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts, Environ. Model. Softw., 22, 1034–1052, 2007.
Dawson, C. W., Abrahart, R. J., and See, L. M.: HydroTest: further development of a web resource for the standardised assessment of hydrological models, Environ. Model. Softw., 25, 1481–1482, 2010.
Dawson, C. W., Mount, N. J., Abrahart, R. J., and Shamseldin, A. Y.: Ideal point error for model assessment in data-driven river flow forecasting, Hydrol. Earth Syst. Sci., 16, 3049–3060, doi:10.5194/hess-16-3049-2012, 2012.
de Vos, N. J.: Echo state networks as an alternative to traditional artificial neural networks in rainfall-runoff modelling, Hydrol. Earth Syst. Sci., in press, 2013.
de Vos, N. J. and Rientjes, T. H. M.: Constraints of artificial neural networks for rainfall-runoff modelling: trade-offs in hydrological state representation and model evaluation, Hydrol. Earth Syst. Sci., 9, 111–126, doi:10.5194/hess-9-111-2005, 2005.
Dibike, B. Y. and Solomatine, D. P.: River flow forecasting using artificial neural networks, Phys. Chem. Earth, 26, 1–7, 2001.
Fernando, D. A. K. and Shamseldin, A. Y.: Investigation of internal functioning of the radial-basis-function neural network river flow forecasting models, J. Hydrol. Eng., 14, 286–292, 2009.
Firat, M.: Comparison of Artificial Intelligence Techniques for river flow forecasting, Hydrol. Earth Syst. Sci., 12, 123–139, doi:10.5194/hess-12-123-2008, 2008.
Fraedrich, D. and Goldberg, A.: A Methodological framework for the validation of predictive simulations, Eur. J. Oper. Res., 124, 55–62, 2000.
Gass, S. I.: Model accreditation: a rationale and process for determining a numerical rating, Eur. J. Oper. Res., 66, 250–258, 1993.
Giustolisi, O. and Laucelli, D.: Improving generalization of artificial neural networks in rainfall–runoff modelling, Hydrolog. Sci. J., 50, 439–457, 2005.
Hamby, D. M.: A review of techniques for parameter sensitivity analysis of environmental models, Environ. Monit. Assess., 32, 135–154, 1994.
Hashem, S.: Sensitivity analysis for feedforward artificial networks with differentiable activation functions, in: Proceedings of the International Joint Conference on Neural Networks, Baltimore, USA, 7-11 June, 1992, 1, 419-424, 1992.

HESSD
10, 145–187, 2013

Helton, J. C.: Uncertainty and sensitivity analysis techniques for use in performance assessment for radioactive waste disposal, Reliab. Eng. Syst. Safe., 42, 327–367, 1993.
Hill, M. C. and Tiedeman, C. R.: Effective Groundwater Model Calibration with Analysis of Sensitivities, Predictions, and Uncertainty, Wiley, New York, 2007.
Hipel, K. W., McLeod, A. I., and Lennox, W. C.: Advances in Box-Jenkins modeling 1. model construction, Water Resour. Res., 13, 567–575, 1977.
Holvoet, K., van Griensven, A., Seuntjens, P., and Vanrolleghem, P. A.: Sensitivity analysis for hydrology and pesticide supply towards the river in SWAT, Phys. Chem. Earth, 30, 518–526, 2005.
Howes, S. and Anderson, M. G.: Computer simulation in geomorphology, in: Modeling Geomorphological Systems, John Wiley and Sons Ltd, Chichester, 1988.
Hu, T. S., Lam, K. C., and Ng, S. T.: River flow time series prediction with a range-dependent neural network, Hydrolog. Sci. J., 46, 729–745, 2001.
Huang, W., Xu, B., and Chan-Hilton, A.: Forecasting flows in Apalachicola River using neural networks, Hydrol. Process., 18, 2545–2564, 2004.
Jain, A. and Kumar, S.: Dissection of trained neural network hydrologic models for knowledge extraction, Water Resour. Res., 45, W07420, doi:10.1029/2008WR007194, 2009.
Jain, A., Sudheer, K. P., and Srinivasulu, S.: Identification of physical processes inherent in artificial neural network rainfall runoff models, Hydrol. Process., 18, 571–581, 2004.
Jakeman, A. J., Letcher, R. A., and Norton, J. P.: Ten iterative steps in development and evaluation of environmental models, Environ. Model. Softw., 21, 602–614, 2006.
Jayawardena, A. W. and Fernando, T. M. K. G.: River flow prediction: an artificial neural network approach, in: Regional Management of Water Resources, 6th IAHS Scientific Assembly, July 2001, IAHS Publ. no. 268, Maastricht, The Netherlands, 2001.
Kişi, Ö.: River flow modelling using artificial neural networks, J. Hydrolog. Eng., 9, 60–63, 2004.
Kişi, Ö.: River flow forecasting and estimation using different artificial neural network techniques, Hydrol. Res., 27–40, 2008.
Kişi, Ö. and Cigizoglu, H. K.: Comparison of different ANN techniques in river flow prediction, Civ. Eng. Environ. Syst., 24, 211–231, 2007.
Kleijnen, J. P. C.: Verification and validation of simulation-models, Eur. J. Oper. Res., 82, 145–162, 1995.
Kleijnen, J. P. C. and Sargent, R. G.: A methodology for fitting and validating metamodels in simulation, Eur. J. Oper. Res., 120, 14–29, 2000.

HESSD
10, 145–187, 2013

LeBaron, B. and Weigend, A. S.: A bootstrap evaluation of the effect of data splitting on financial time series, IEEE T. Neural Netw., 9, 213–220, 1998.
McCuen, R. H.: The role of sensitivity analysis in hydrologic modelling, J. Hydrol., 18, 37–53, 1973.
Mishra, S.: Uncertainty and sensitivity analysis techniques for hydrologic modelling, J. Hydroinform., 11, 282–296, 2009.
Mount, N. J. and Abrahart, R. J.: Discussion of "River flow estimation from upstream flow records by artificial intelligence methods" by M. E. Turan, M. A. Yurdusev [J. Hydrol. 369 (2009) 71–77], J. Hydrol., 396, 193–196, 2011a.
Mount, N. J. and Abrahart, R. J.: Load or concentration, logged or unlogged? Addressing ten years of uncertainty in neural network suspended sediment prediction, Hydrol. Process., 25, 3144–3157, 2011b.
Muleta, M. K. and Nicklow, J. W.: Sensitivity and uncertainty analysis coupled with automatic calibration for a distributed watershed model, J. Hydrol., 306, 127–145, 2005.
Oreskes, N., Shrader-Frechette, K., and Belitz, K.: Verification, validation and confirmation of numerical models in the Earth Sciences, Science, 263, 641–646, 1994.
Pappenberger, F., Beven, K. J., Ratto, M., and Matgen, P.: Multi-method global sensitivity analysis of flood inundation models, Adv. Water Resour., 31, 1–14, 2008.
Prada-Sarmiento, F. and Obregon-Neira, N.: Forecasting monthly streamflows based on artificial neural networks, J. Hydrol. Eng., 14, 1390–1395, 2009.
Rumelhart, D. E., Hinton, G. E., and Williams, R. J.: Learning internal representations by error propagation, in: Parallel Distributed Processing: Explorations in the Microstructures of Cognition, Volume 1, The MIT Press, Cambridge, Massachusetts, USA, 318–362, 1986.
Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S.: Global Sensitivity Analysis, The primer, Wiley, Chichester, 304 pp., 2008.
Sargent, R. G.: Verification and validation of simulation models, in: Proceedings of the 2011 Winter Simulation Conference, Informs Simulation Society, Phoenix, Arizona, USA, 183–197, 2011.
See, L. M., Jain, A., Dawson, C. W., and Abrahart, R. J.: Visualisation of hidden neuron behaviour in a neural network rainfall-runoff model, in: Practical Hydroinformatics: Computational Intelligence and Technological Developments in Water Applications, Springer, Berlin, 87–99, 2008.

HESSD
10, 145–187, 2013

Shamseldin, A. Y. and O'Connor, K. M.: A non-linear neural network technique for updating of river flow forecasts, Hydrol. Earth Syst. Sci., 5, 577–598, doi:10.5194/hess-5-577-2001, 2001.
Smith, E. D., Szidarovszky, F., Karnavas, W. J., and Bahill, A. T.: Sensitivity analysis, a powerful system validation technique, Open Cybernet. System. J., 2, 39–56, 2008.
Snee, R. D.: Validation of regression models: methods and examples, Technometrics, 19, 415–428, 1977.
Spear, R. C. and Hornberger, G. M.: Eutrophication in Peel Inlet, II Identification of critical uncertainties via generalized sensitivity analysis, Water Resour. Res., 14, 43–49, 1980.
Spruill, C. A., Workman, S. R., and Taraba, J. L.: Simulation of daily and monthly stream discharge from small watersheds using the SWAT model, T. Am. Soc. Civil. Eng., 43, 1431–1439, 2000.
Sudheer, K. P.: Knowledge extraction from trained neural network river flow models, J. Hydrol. Eng., 10, 264–269, 2005.
Sudheer, K. P. and Jain, A.: Explaining the internal behaviour of artificial neural network river flow models, Hydrol. Process., 18, 833–844, 2004.
Sun, F., Chen, J., Tong, Q., and Zeng, S.: Structure validation of an integrated waterworks model for trihalomethanes simulation by applying regional sensitivity analysis, Sci. Total Environ., 408, 1992–2001, 2009.
Tsai, M.-J., Abrahart, R. J., Mount, N. J., and Chang, F.-J.: Including spatial distribution in a data-driven, rainfall-runoff model to improve reservoir inflow forecasting in Taiwan, Hydrol. Process., in press, 2013.
Turanayi, T. and Rabitz, H.: Local methods, in: Sensitivity Analysis, Wiley Series in Probability and Statistics, Wiley, Chichester, 2000.
Wilby, R. L., Abrahart, R. J., and Dawson, C. W.: Detection of conceptual model rainfall-runoff processes inside an artificial neural network, Hydrolog. Sci. J., 48, 163–181, 2003.
Wu, W., May, R., Dandy, G. C., and Maier, H. R.: A method for comparing data splitting approaches for developing hydrological ANN models, in: Proceedings of the 6th Biennial Meeting of the International Environmental Modelling and Software Society, 2012 International Congress on Environmental Modelling and Software Managing Resources of a Limited Planet, Leipzig, Germany, 2012.
Yeung, D. S., Cloete, I., Shi, D., and Ng, W. W. Y.: Sensitivity Analysis for Neural Networks, Springer, Berlin, 86 pp., 2010.

HESSD
10, 145–187, 2013

Young, P. C. and Beven, K. J.: Databased mechanistic modelling and the rainfall flow nonlinearity, Environmetrics, 5, 335–363, 1994.
Young, P. C., Chotai, A., and Beven, K. J.: Data-based mechanistic modelling and the simplification of environmental systems, in: Environmental Modelling: Finding Simplicity in Complexity, Wiley, Chichester, 371–388, 2004.
Zealand, C. M., Burn, D. H., and Simonovic, S. P.: Short term streamflow forecasting using artificial neural networks, J. Hydrol., 214, 32–48, 1999.

HESSD
10, 145–187, 2013

Table 1. Description of the River Ouse catchment and its primary sub-catchments.

Gauge	ID	Catchment physiography	Land cover
Ouse at Skelton	27009	Area 3315 km^2 Max elevation 714 m AOD Min elevation 4.6 m AOD Majority high to moderate permeability bedrock	Woodland 7 % Arable/horticultural 31 % Grassland 44 % Mountain/heath/bog 12 % Urban 2 % Other 4 %
Swale at Crakehill	27071	Area 1363 km^2 Max Elevation 714.3 m AOD Min elevation 12 m AOD Majority high to moderate permeability bedrock	Woodland 6 % Arable/horticultural 35 % Grassland 41 % Mountain/heath/bog 12 % Urban 1 % Other 5 %
Nidd at Skip Bridge	27062	Area 516 km^2 Max elevation 702.6 m AOD Min elevation 8.2 m AOD Majority high to moderate permeability bedrock	Woodland 8 % Arable/horticultural 22 % Grassland 49 % Mountain/heath/bog 13 % Urban 3 % Other 5 %
Ure at Westwick	27007	Area 915 km^2 Max elevation 710.0 m AOD Min elevation 14.2 m AOD Majority moderate permeability bedrock	Woodland 8 % Arable/horticultural 14 % Grassland 56 % Mountain/heath/bog 19 % Urban 1 % Other 2 %

HESSD
10, 145–187, 2013

Table 2. Epochs for preferred NNRF models based on validation data.

Model scenario	Hidden units			
	2	3	4	5
A	700	1100	3000	800
B	1000	7000	20 000	20 000

HESSD
10, 145–187, 2013

Table 3. Calibration performance of candidate models for Scenario A.

Hidden units	RMSE $m^3 s^{-1}$	MSRE	R-squared
2	27.19	0.0934	0.7977
3	27.10	0.0900	0.7992
4	**27.07**	0.0875	**0.7998**
5	27.21	**0.0833**	0.7987
MLR benchmark	27.61	0.1969	0.7909

Best performing NN models for each metric are in bold.

HESSD
10, 145–187, 2013

Table 4. Validation performance of candidate models for Scenario A.

Hidden units	RMSE $m^3 s^{-1}$	MSRE	R-squared
2	**26.25**	0.0825	0.8034
3	26.26	0.0809	0.8035
4	26.28	0.0794	0.8034
5	26.32	**0.0752**	**0.8042**
MLR benchmark	21.69	0.1151	0.8657

Best performing NN models for each metric are in bold.

HESSD
10, 145–187, 2013

Table 5. Calibration performance of candidate models for Scenario B.

Hidden units	RMSE $m^3 s^{-1}$	MSRE	R-squared
2	22.32	**0.0694**	0.8665
3	22.04	0.0841	0.8674
4	21.85	0.0718	**0.8710**
5	**21.83**	0.0732	**0.8710**
MLR benchmark	23.10	0.2151	0.8537

Best performing NN models for each metric are in bold.

HESSD
10, 145–187, 2013

Table 6. Validation performance of candidate models for Scenario B.

Hidden units	RMSE $m^3 s^{-1}$	MSRE	R-squared
2	21.94	0.0653	0.8697
3	21.63	0.0599	0.8708
4	21.62	0.0567	0.8712
5	**21.58**	**0.0564**	**0.8714**
MLR benchmark	23.6	0.1043	0.8513

Best performing NN models for each metric are in bold.

HESSD
10, 145–187, 2013

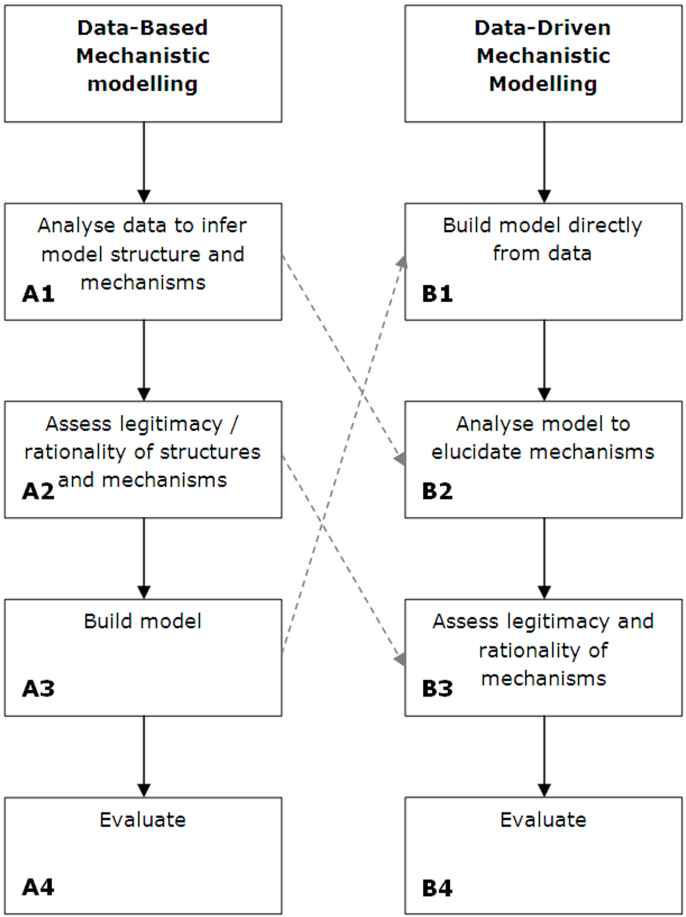

Fig. 1. Reordering of the DBM framework to generate the DDMM framework.

HESSD
10, 145–187, 2013

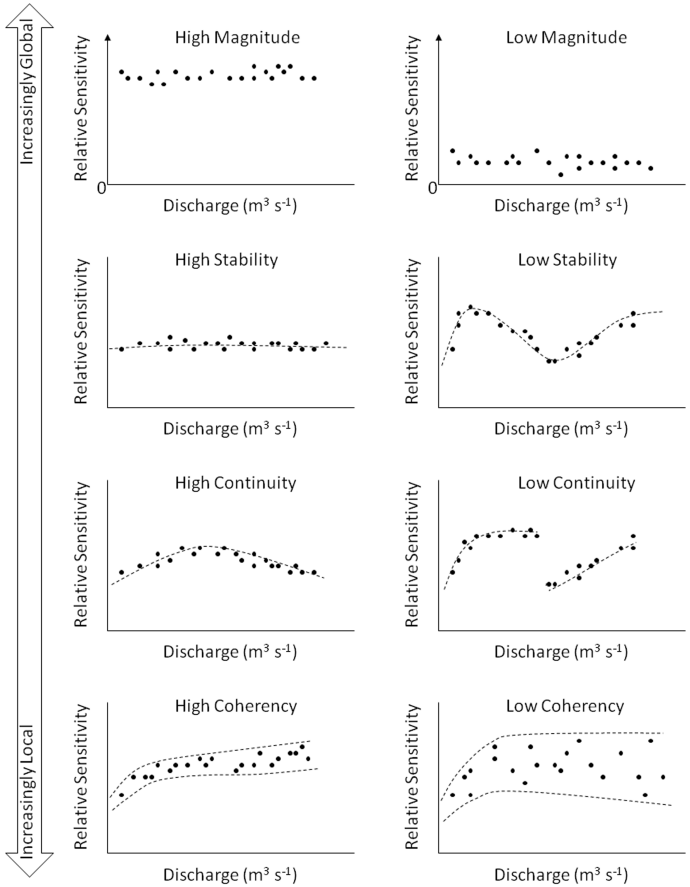

Fig. 2. Characteristic patterns of global-local model sensitivity. The continuum represented by the arrow on the left indicates the relative focus of each sensitivity characteristics on a range between global and local.

HESSD
10, 145–187, 2013

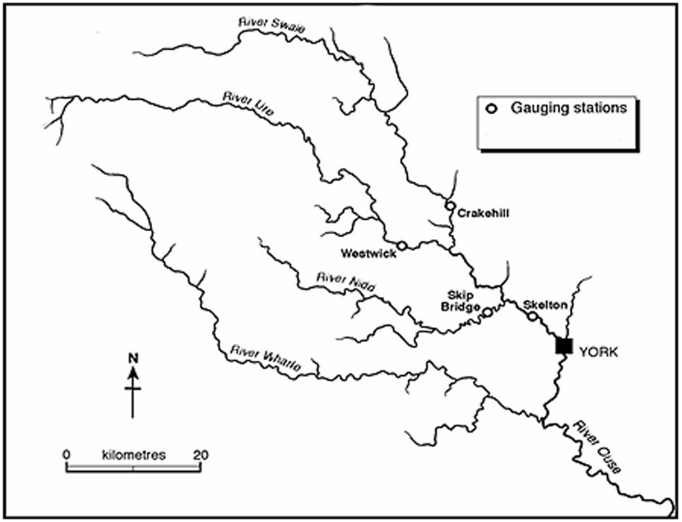

Fig. 3. River Ouse catchment in North Yorkshire, UK.

HESSD
10, 145–187, 2013

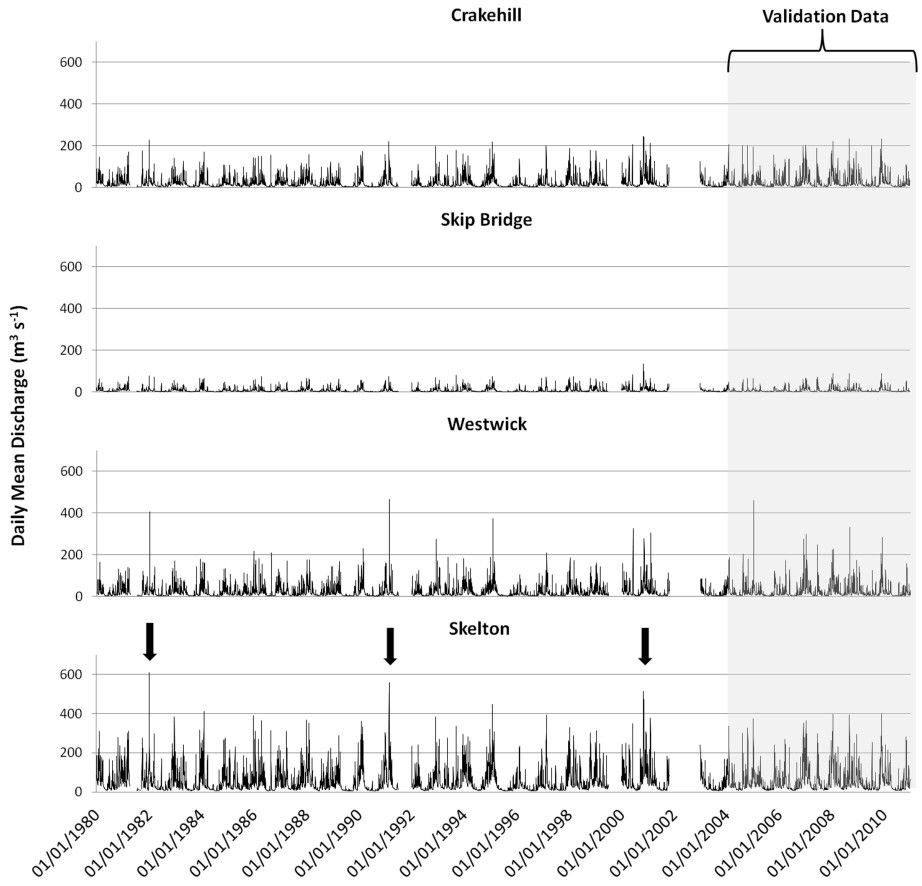

Fig. 4. Hydrographs for the four gauging stations showing data partitioning.

HESSD
10, 145–187, 2013

Legitimising neural network river forecasting models

N. J. Mount et al.

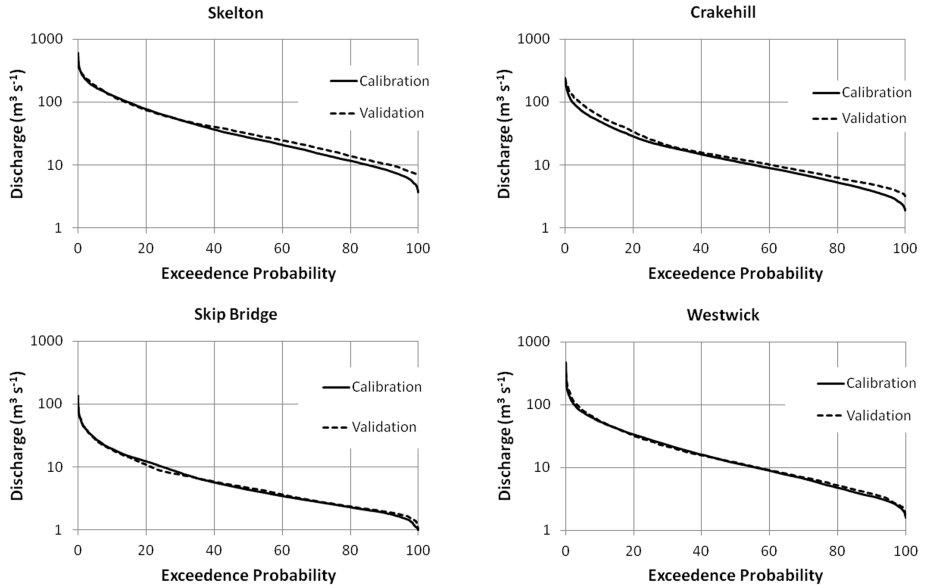

Fig. 5. Exceedence probability plots for the four gauging stations.

HESSD
10, 145–187, 2013

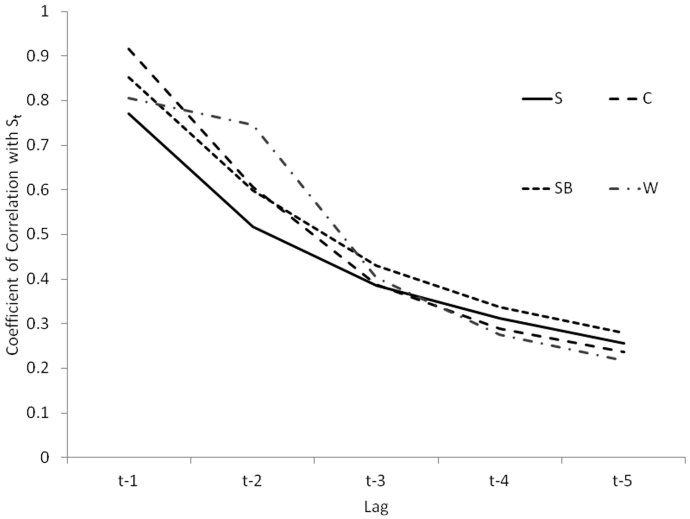

Fig. 6. Lag analysis for the four gauging stations.

HESSD
10, 145–187, 2013

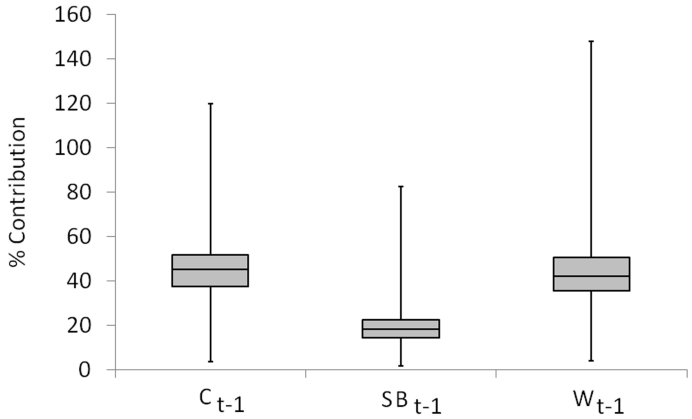

Fig. 7. Proportional contributions of lagged upstream inputs to discharge forecast at Skelton.

HESSD
10, 145–187, 2013

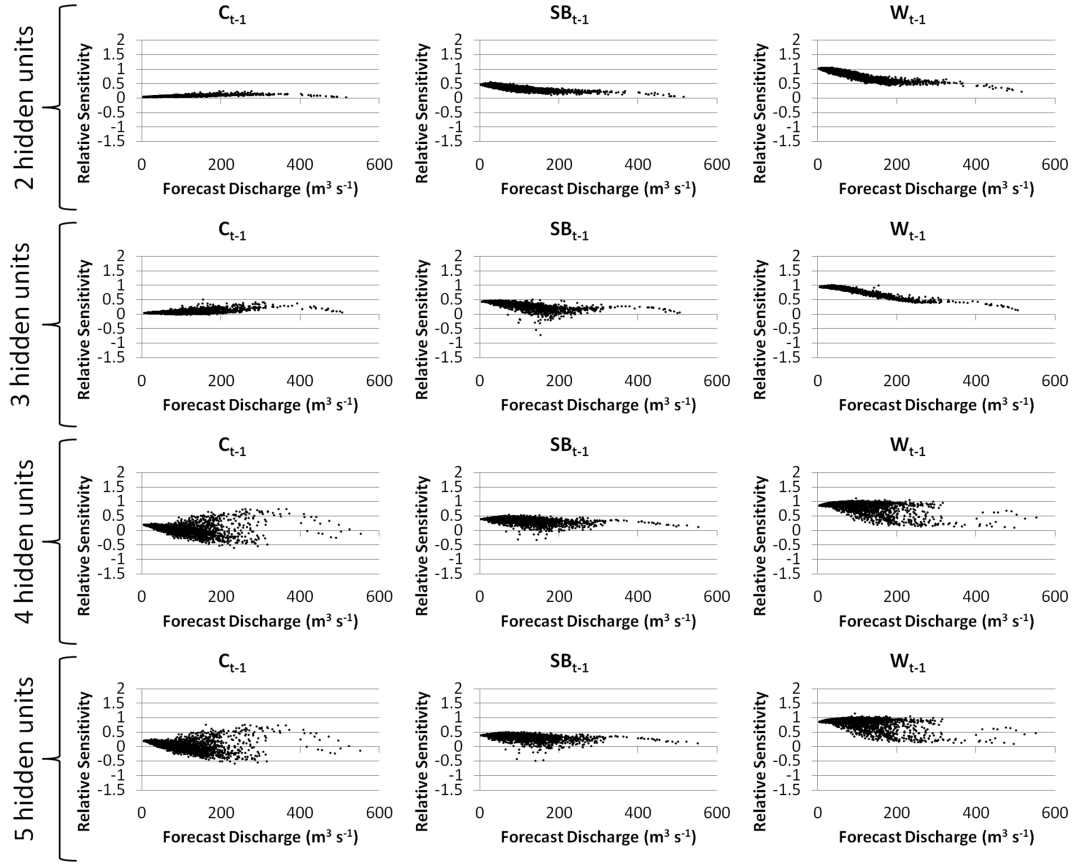

Fig. 8. Global-local relative sensitivity plots for all candidate models in Scenario A.

HESSD
10, 145–187, 2013

Legitimising neural network river forecasting models

N. J. Mount et al.

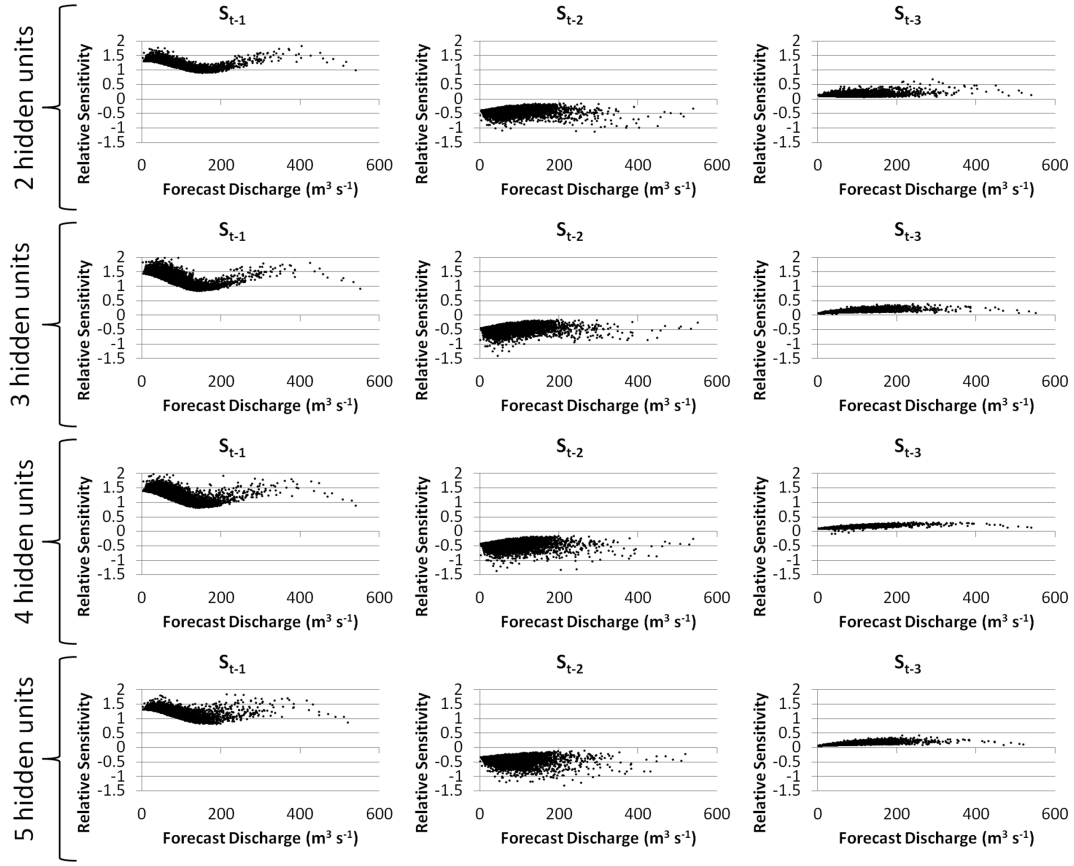

Fig. 9. Global-local relative sensitivity plots for all candidate models in Scenario B.

HESSD
10, 145–187, 2013

Legitimising neural network river forecasting models

N. J. Mount et al.