Sensitivity analysis for comparison, validation and physical-legitimacy of neural network-based hydrological models

C.W. Dawson[1], N.J. Mount[2], R.J. Abrahart[2]* and J. Louis[3]

[1] Department of Computer Science, Loughborough University, Loughborough, LE11 3TU, UK; [2]School of Geography, University of Nottingham, Nottingham, NG7 2RD, UK; [3]School of Computing and Mathematics, Charles Stuart University, Locked Bag 588, Wagga Wagga, NSW 2678, Australia. *Corresponding author E-mail: bob.abrahart@nottingham.ac.uk

SHORT TITLE: Sensitivity analysis for comparison, validation and physical-legitimacy

ABSTRACT

This paper addresses the difficult question of how to perform meaningful comparisons between neural network-based hydrological models and alternative modelling approaches. Standard, goodness-of-fit metric approaches are limited since they only assess numerical performance and not physical legitimacy of the means by which output is achieved. Consequently, the potential for general application or catchment transfer of such models is seldom understood. This paper presents a partial derivative, relative sensitivity analysis method as a consistent means by which the physical legitimacy of models can be evaluated. It is used to compare the behaviour and physical rationality of a generalised linear model and two neural network models for predicting median flood magnitude in rural catchments. The different models perform similarly in terms of goodness-of-fit statistics, but behave quite distinctly when the relative sensitivities of their parameters are evaluated. The neural solutions are seen to offer an encouraging degree of physical legitimacy in their behaviour, over that of their generalised linear modelling counterpart, particularly when overfitting is constrained. This indicates that neural solutions are preferable models for transferring into ungauged catchments. Thus, the importance of understanding both model performance and physical legitimacy when comparing neural models with alternative modelling approaches is demonstrated.

KEYWORDS | generalised model, index flood, neural network, partial derivative, sensitivity analysis, ungauged catchment

INTRODUCTION

This paper presents an approach for delivering greater meaning from the comparison of artificial neural network (ANN) models with alternative modelling approaches in hydrological studies. ANN-based hydrological models are most commonly applied as black-box tools and the internal mechanisms by which the model output is generated are not normally explored in hydrological terms. Used in this way, an ANN's primary purpose is the optimisation of complex, non-linear relations between a specific set of hydrological input and output data, and standard goodness-of-fit procedures may, therefore, be considered an adequate basis by which to compare its performance to that of other models (Klemes, 1986; Refsgaard and Knusden, 1996). Indeed, assessments of goodness-of-fit have been widely used in comparative hydrological modelling studies to argue that ANN models can perform as well as, or better than alternative modelling approaches (e.g. Shrestha and Nestmann, 2009; Mount and Abrahart, 2011). However, such arguments are informed solely by the degree of optimisation that is achieved by each model. They say nothing about the means by which different models achieve their performance and the relative merits of these alternative means. Indeed, when ANN models are applied solely as black-boxes, their potential relative to other modelling approaches can never be properly understood in a generalised or transferrable manner because the extent to which their modelling mechanisms conform to physically-based, hydrological domain knowledge remains untested (Howes and Anderson, 1988; Sargent, 2011). Consequently, critical questions about whether ANN modelling mechanisms are more or less reflective of real-world hydrological processes than alternative models are seldom addressed directly (Minns and Hall, 1996; Abrahart *et al.*, 2011), and the relative extent to which they are able to deliver hydrological process insights (i.e. Caswell's (1976) model duality) is not normally evaluated. The purpose of this paper is to present a method by which these questions may be addressed.

More informative approaches to model comparison are required that explicitly consider the internal behaviours of the different models and assess them according to their conformance with the logical, rational and physical expectations of the modeller (c.f. Robinson, 1997). This process is termed model legitimisation and is discussed in a philosophical context by Oreskes *et al.* (1994) and an applied, hydrological modelling context by Mount *et al.* (in press). Sensitivity analysis (Hamby, 1994) is an important and effective means by which the legitimacy of a hydrological model may be explored. It has been widely applied in conceptual and physically-based modelling over several decades (e.g. McCuen, 1973; Beven and Binley, 1992; Schulz and Huwe, 1999; Radwan *et al.*, 2004; Pappenberger *et al.*, 2008; Mishra, 2009; Zhang *et al.*, 2012). A variety of approaches have been used including local (e.g. Turanayi and Rabitz, 2000; Spruill *et al.*, 2000; Holvoet *et al.*, 2005; Hill and Tiedeman, 2007), regional (e.g. Spear and Hornberger, 1980) and global-scale methods (Muleta and Nicklow, 2005; Salteli *et al.*, 2008). By contrast, sensitivity analysis has not been widely adopted in ANN modelling studies beyond a few, isolated examples (Sudheer, 2005; Nourani and Fard, 2012). This is presumably because the equations that relate inputs and outputs in an ANN are considered complex, inaccessible and difficult to interpret (Aytek *et al.*, 2008; Abrahart *et al.*, 2009), making exploration of model sensitivity via direct analysis of the governing equations difficult. Nonetheless, recent progress has been made (Yeung *et al.*, 2010) and relative sensitivity analysis techniques for ANNs have made it possible to assess the internal, mechanistic legitimacy of such models (Abrahart *et al.*, 2012b; Mount *et al.*, in press). However, the focus of these studies has so far been restricted to mechanical considerations. The application of sensitivity analysis to evaluate the physical legitimacy of ANN-based hydrological models, and thus the degree to which they can be generalised and transferred, remains an outstanding task.

In this paper, we apply a sensitivity analysis method that can be used to compare the physical legitimacy of ANN-based hydrological models and alternative model counterparts in a direct manner. We exemplify the method by comparing the performance and physical legitimacy of a pair of ANN-based models and an established generalised linear model (GLM) for median flood magnitude prediction in ungauged catchments in the UK. First order, partial derivatives of each model's response function are computed, interpreted and used as a consistent means by which the physical legitimacy of each model can be evaluated and compared. This focus on response function behaviour is distinctly different to past efforts to assess the physical legitimacy of ANN models, which have traditionally explored internal structural components, such as weights (Abrahart *et al.*, 1999; Olden and Jackson, 2002; Anctil *et al.*, 2004; Kingston *et al.*, 2003,2005,2006,2008) and units (Wilby *et al.*, 2003; Jain *et al.*, 2004; Sudheer and Jain, 2004; See *et al.*, 2008; Fernando and Shamseldin, 2009; Jain and Kumar, 2009). However, the uniqueness of ANN structures means that the information derived from them cannot easily be compared directly with that derived from alternative models with different internal structures - thus limiting the comparative value of the information. To overcome this problem, we here assess the physical legitimacy of an ANN's overall response function using a standard relative sensitivity-based method that can be consistently and directly replicated across a range of alternative model types and that is widely understood and accepted by hydrologists. Consequently, an evaluation of the physical legitimacy of the means by which each model's performance is obtained accompanies the usual assessments of output validity; enabling the extent to which each model delivers a transferable, general solution to be considered.

COMPARING GLM AND ANN-BASED MODELS FOR UNGAUGED CATCHMENT PREDICTION IN THE UK

The modelling of hydrological responses in ungauged catchments remains an important focus of research for hydrologists, especially as the majority of the world's river catchments remain ungauged or poorly gauged. In such catchments, the application of distributed physically-based models and statistical approaches is hampered by a lack of input parameter knowledge and datasets. Consequently, lumped models which relate broad physiographic, hydrogeologic and climatologic catchment descriptors to flood frequency curves, have long been recognised as offering potential (Rodriguez-Iturbe and Valdes, 1979; Grover *et al.*, 2002).

The standard UK method (Natural Environment Research Council, 1975; Vogel and Kroll, 1992; Schrieber and Demuth, 1997) models the relationship between the median of the annual flood series (*QMED*) and a set of regionalised catchment descriptors for rivers in the national, gauged network. The modelled relationship is then applied to ungauged catchments and used to estimate *QMED*, which is subsequently multiplied by a standard, dimensionless growth curve to estimate flood frequency (Institute of Hydrology, 1999).

Four catchment descriptors are used in the standard UK methodology: 1) *AREA* (catchment area in km^2); 2) *SAAR* (standard-period average annual rainfall in mm); 3) *FARL* (flood attenuation due to reservoirs and lakes); 4) *BFIHOST* (baseflow index derived from HOST data; Boorman *et al.*, 1995).

These catchment descriptors can be thought of as physical controls of *QMED* potential. *SAAR* controls the hydrological inputs to the catchment, *AREA* controls the scaling of the catchment response, whilst *BFIHOST* and *FARL* control the degree of buffering of the input-output signal.

Of central importance to the above method is the model that is used to relate *QMED* and the catchment descriptors. These relationships are non-linear and not well represented by standard multiple linear regression. Therefore, the most recent UK method described applies a range of non-linear transformations within a generalised linear modelling (GLM) framework (Kjeldsen *et al.*, 2008; Kjeldsen and Jones, 2009; Kjeldsen and Jones, 2010). The end product is a non-linear regression equation (see Equation 1) from which *QMED* can be estimated directly from the four catchments descriptors.

ANN models are also very effective at optimising complex, non-linear relations in hydrological data (American Society of Civil Engineers 2000a,b; Maier and Dandy, 2000; Dawson and Wilby, 2001; Maier *et al.*, 2010; Abrahart *et al.*, 2010; 2012b) and a number of studies have highlighted their potential in ungauged catchment prediction (Liong *et al.*, 1994; Muttiah *et al.*, 1997; Hall and Minns, 1998; Hall *et al.*, 2000; Dastorani and Wright, 2001; Dawson *et al.*, 2006; Dastorani *et al.*, 2010). Indeed, the UK relationship between *QMED* and catchment descriptors has also been modelled using ANNs and been shown to deliver comparable levels of fit when compared to GLMs (Dawson *et al.*, 2006). However, it remains unclear whether the two modelling approaches are similarly comparable with respect to their physical legitimacy. Models with greater physical legitimacy should be more generally transferrable to new catchment settings. Therefore, determining the physical legitimacy of each model is an important element in delivering a physically informed evaluation of how robustly it can be expected to transfer from the gauged catchments upon which it is developed, to the ungauged catchments in which it is intended to be applied.

In the following sections, the importance of evaluating both model performance and physical legitimacy in ANN model comparisons is exemplified by contrasting the performance and legitimacy of the standard GLM method for *QMED* prediction with two different ANN-based model counterparts. Its use as an example is particularly appropriate because the model inputs and outputs are all physical-based measurements, meaning that patterns observed in inputs and output relations can be interpreted directly in physical terms, also the number of model inputs is relatively small, the first order partial derivatives can be computed for the GLM and directly compared with those of the ANN-based models, and the results of the analysis have real-world relevance and application.

Data

A GLM model and two counterpart ANN models for *QMED* estimation are developed for comparison, with the model inputs conforming to the four used in the standard UK methodology. These inputs were extracted from a pre-filtered set of HiFlows-UK rural catchment data, available at (http://www.environment-agency.gov.uk/hiflows/97503.aspx). *AREA* values are derived from the Centre for Ecology and Hydrology's Integrated Hydrological Digital Terrain Model (based on a 50m grid) and represent surface catchment area projected onto a horizontal plane, draining to the gauging station (Marsh and Hannaford, 2008: 5). *SAAR* values are derived from UK precipitation records over the standard period 1961-1990. *FARL* provides a guide to the degree of flood attenuation attributable to reservoirs and lakes above the gauging station. The index ranges from zero (complete attenuation) to one (no attenuation) with values < 0.8 representing a substantial influence on flood response. *BFIHOST* is derived from the HOST (Hydrology of Soil Types) soil data classification and ranges from zero (impermeable) to one (completely permeable). In undisturbed catchments, a strong association exists between Baseflow Index (derived from archived gauged daily mean flows) and *BFIHOST*. The relationships between *QMED and AREA*, *SAAR and FARL* are positive, whilst that between *QMED* and *BFIHOST* is negative.

The data from which our models are derived are almost identical to those from which the GLM that is published in the revitalised UK Flood Estimation Handbook (Kjeldsen *et al.*, 2008) has been developed, and full particulars of the Hi-Flows UK data set can be found in this handbook. A statistical summary of our dataset is provided in Table 1. Some minor discrepancies exist between the data used in this study and that used by Kjeldsen *et al.* (2008) due to our use of the public release version of HiFlows-UK 3.02 rather than the pre-release version originally used. Specifically, our dataset comprises 597 rural catchment records rather than the 602 used previously, and we use an unadjusted flood attenuation variable.

Model development procedures

Three models were developed for comparison.

1. $QMED_{GLM}$ – a GLM developed on all 597 catchment records, using the methodology outlined in Kjeldsen *et al.* (2008).
2. ANN_A – an optimised ANN, selected from 180 candidate solutions of varying complexity and training iterations according to both its goodness-of-fit performance and avoidance of evident overfitting.
3. ANN_B – a purposely over-trained version of ANN_A in which the number of training iterations was artificially extended to deliver an overfitted solution. It is included as a means of exemplifying the impact of ANN overfitting on the physical legitimacy of a network response function.

$QMED_{GLM}$ was developed in accordance with the method of Kjeldsen *et al.* (2008). Despite the minor differences in the dataset noted above, the resultant regression equation (Equation 1) remains almost identical to Kjeldsen's original:

$$QMED_{GLM} = 8.6704 AREA^{0.8568} 0.1550^{\left(\frac{1000}{SAAR}\right)} FARL^{3.3662} 0.0380^{BFIHOST} \quad (1)$$

ANN_A and ANN_B comprise a Multi-Layer Perceptron (MLP), with one hidden layer, trained using error back propagation (Rumelhart *et al.*, 1986). The basic structure of these networks is shown schematically in Figure 1. The ANN consists of a number of units or neurons arranged in three layers (although additional hidden layers can be incorporated). The units in the input layer distribute the inputs to the units in the hidden layer, which in turn pass their outputs to the output layer (usually consisting of a single output neuron). Each neuron consists of a weighted set of inputs and an activation function – typically the logistic sigmoid function (Equation 2). The output from a single unit is calculated by applying this sigmoid function to the weighted sum of its inputs.

$$f(x) = \frac{1}{1+e^{-x}} \quad (2)$$

Training such networks using back propagation involves presenting the ANN with training data, calculating the error of the network's output with respect to the observed values, propagating this error backwards through the network and adjusting the input weights to the neurons accordingly (to reduce this error). This process must be repeated many times, making minor adjustments to the weights of each cycle (or epoch), until the

ANN begins to map input values to the correct output response. The amount by which the weights are adjusted each time can be manipulated by using a learning rate multiplier. Readers that are unfamiliar with ANN concepts, structures and training methods are referred to Kattan *et al.* (2011) or Nelson (2011).

The simplicity of this ANN has enabled the development of computational methods for delivering first-order partial derivatives of its response function (Hashem, 1992), which we subsequently use as the basis for our comparative assessment of model legitimacy (see Section 3). This standard ANN has been successfully used in many hydrological studies in the past (Abrahart *et al.*, 2012a) and provides an established non-linear modelling benchmark for ANN studies and a starting point against which more novel approaches can subsequently be compared (Mount *et al.*, 2012). Whilst it is recognised that more advanced ANN structures might arguably deliver some additional optimisation advantages, the computational methods required to quantify their response function partial derivatives, and hence deliver directly comparable assessments of their physical legitimacy, are not readily available. Their use is thus avoided in this study.

ANN_A was developed using the approach described in Dawson *et al.* (2006) in which a large number of candidate ANNs are trained on a random subset of the data, partitioned according to a 60% calibration to 40% cross-validation ratio. Although there is no agreed standard for splitting the data, this ratio is widely accepted in hydrological modelling (Mount and Abrahart, 2011; See and Openshaw, 2000). 180 candidate models containing 2, 3, 4, 5, 6, 7, 8, 9, 10 hidden units were developed with each candidate being trained for up to 20,000 epochs in steps of 1,000, using a learning rate of 0.1 and a momentum value of 0.9. Each candidate model was cross-validated using the remaining 40% as a means of preventing overfitting (Giustolisi and Laucelli, 2005; Piotrowski and Napiorkowski, 2013).

Overfitting of each candidate solution was evaluated according to its cross-validation scores, and the candidate solution displaying the best optimisation performance, whilst avoiding apparent overfitting, was selected as the final model.

ANN_A has nine hidden units, and is trained for 4000 epochs. ANN_B, which we adopt as an example of an overfitted ANN, is structurally identical to ANN_A. However its training epochs have been artificially extended to ten times that of ANN_A (i.e. 40,000 epochs) to promote overfitting. The network unit weights and biases are provided in Table 2 and are used as the inputs to Equation 8, from which relative sensitivity can be computed.

It should be noted that the GLM and ANN models utilise the available data records differently during model development. Whilst the GLM uses all 597 records to define the model, each candidate ANN uses only the first 400 records to refine the model, and the remaining 197 records to constrain it via cross-validation. Indeed, the apparent inconsistency with which the GLM and ANN models use the available data could be cited as an argument to negate the fairness of a direct comparison between them. However, this stance fails to credit that both models do use all of the data in the model development process; they just use it in a characteristically different manner that reflects the fundamental differences between each method. In this sense, the models are comparable; not because they use the same data in the same way, but rather because each one's use of the data is equally appropriate and justifiable in the context of its own model development method.

MODEL PERFORMANCE AND PHYSICAL LEGITIMACY ASSESSMENT

Model performance evaluation

Each model's performance was evaluated using standard goodness-of-fit metrics to deliver output validation. To ensure a consistent approach the metrics were generated using HydroTest (http://www.hydrotest.org.uk), a standardised, open access web site that performs the required numerical calculations (Dawson *et al.* 2007,2010). Each model's performance is evaluated using *RMSE* (root mean squared error) and R^2 (R-squared – the coefficient of determination) providing an overall measure of model performance; *MSRE* (mean squared relative error) and *MSLE* (mean squared logarithmic error) providing two additional measures of performance which place greater emphasis on errors occurring in lower magnitude predictions. These comparative performance statistics are defined as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Q_i - \hat{Q}_i)^2}{n}} \tag{3}$$

$$R^2 = \left[\frac{\sum_{i=1}^{n}(Q_i - \overline{Q})(\hat{Q} - \tilde{Q})}{\sqrt{\sum_{i=1}^{n}(Q_i - \overline{Q})^2 \sum_{i=1}^{n}(\hat{Q}_i - \tilde{Q})^2}} \right]^2 \tag{4}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Q_i - \hat{Q}_i)^2 \tag{5}$$

$$MSLE = \frac{1}{n}\sum_{i=1}^{n}(\ln Q_i - \ln \hat{Q}_i)^2 \tag{6}$$

where Q_i is observed index flood value *i* (of *n* values), $\hat{Q}_i$ is the modelled value *i*, $\bar{Q}$ is the mean of the observed data, and $\tilde{Q}$ is the mean of the modelled data.

Physical legitimacy

Following the recent studies of Abrahart *et al.* (2012b) and Mount *et al*. (in press), the physical legitimacy of each model was assessed by means of relative, first-order partial derivative sensitivity analysis (see Hamby, 1994 for an overview of sensitivity analysis approaches). Partial derivative sensitivity analysis elucidates the patterns of influence that each model input has on the output (and *vice versa*) across the output range, thus revealing the internal behaviour of the model response function. First order derivatives reveal the separate behaviours associated with each model input. When using partial derivatives in model comparison studies, it is necessary to standardise derivative values to rates to avoid the difficulties associated with comparing absolute values derived from different inputs with different ranges (Nourani and Fard, 2012). Patterns of relative sensitivity can then be used to directly compare the internal response function behaviour of different models, and legitimacy of these behaviours can then be evaluated according to how well the relative sensitivity patterns conform to the logical, rational and physical expectations of the modeller. The relative sensitivity (RS_i) of the output from a model (O) with respect to input (I_i) can be calculated as:

$$RS_i = \frac{\partial O}{\partial I_i} \cdot \frac{I_i}{O} \qquad (7)$$

Partial derivatives can be computed for ANNs via the application of a backward chaining partial differentiation rule as outlined in Hashem (1992). Adapted from Hashem's more general rule, for an ANN with sigmoid activation functions (i.e. of standard type, as used in our case study), one hidden layer, *i* input units, *j* hidden units and one output unit

(O), the partial derivative of a network's output can be calculated with respect to each of its inputs as:

$$\frac{\partial O}{\partial I_i} = \sum_{j=1}^{n} w_{ij} w_{jO} h_j (1 - h_j) O (1 - O) \qquad (8)$$

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.

One important difference between calculating partial derivatives for multiple input, single output GLMs and ANN models should, however, be noted. When computing partial derivatives of a GLM, there is no need to vary the values of the other inputs to investigate the range of sensitivity responses under different input conditions. This is because GLMs deliver a simple additive response function, such that the relative sensitivity for any one variable will involve only that variable, given that all other parts of the expression will cancel out, during the process of scaling the other variables. Hence, relative sensitivity values for each input to the $QMED_{GLM}$ model (Equation 1) can be computed according to Equations (9)–(12). The final relative sensitivities of the $QMED_{GLM}$ model are provided in Equations (13)–(16).

$$\frac{\partial QMED}{\partial AREA} = \frac{0.8568\, QMED}{AREA} \qquad (9)$$

$$\frac{\partial QMED}{\partial SAAR} = \frac{1864.05\, QMED}{SAAR^2} \qquad (10)$$

$$\frac{\partial QMED}{\partial FARL} = \frac{3.3662\,QMED}{FARL} \quad (11)$$

$$\frac{\partial QMED}{\partial BFIHOST} = -6.5385\,QMED\,.BFIHOST \quad (12)$$

$$RS_{AREA} = 0.8568 \quad (13)$$

$$RS_{SAAR} = 1864.05 / SAAR \quad (14)$$

$$RS_{FARL} = 3.3662 \quad (15)$$

$$RS_{BFIHOST} = -6.5385\ BFIHOST^2 \quad (16)$$

The same is not true for ANNs, which are not constrained to produce simple, additive response functions. When computing partial derivatives for an ANN it is therefore necessary to isolate the pattern of relative sensitivity of each input variable in turn by holding the other inputs at fixed values so that the patterns of sensitivity associated with each variable can be interpreted within the context of the other variable states. To this end we adopt a simple three-step methodology.

Step 1: Compute 25th percentile, median and 75th percentile values for each input variable in the data set.

Step 2: Holding all other variables at either 25th percentile, median or 75th percentile, vary each input variable in turn from across the range of observed values.

Step 3: Plot results and interpret the resultant graphs.

Thus, physically speaking, if variable states in our study are held at the 25th percentile (or the 75th percentile in the case of the inverse *BFIHOST* measure), the resultant scenario under test is representative of relatively small, dry catchments with high permeability and high flood attenuation: i.e. low catchment *QMED* potential. Conversely, when variables states are held at the 75th percentile (with *BFIHOST* at the 25th percentile), the resultant scenario under test will be representative of relatively large, wet catchments with low permeability and low attenuation: i.e. high catchment *QMED* potential.

RESULTS

Independence

Figure 2 and Table 3 present an overview of the data showing the relationships that exist between each of the five variables. *AREA* is not correlated with any of the other three parameters (correlation coefficient ranging from -0.07 to -0.02). There is a negative correlation between *SAAR* and *BFIHOST* (correlation coefficient of -0.42) and a similar strength negative relationship between *SAAR* and *FARL* (correlation coefficient of -0.39). The only positive correlation is that between *BFIHOST* and *FARL* (correlation coefficient of 0.11). These weak relationships indicate a reasonable degree of linear independence between the four variables. The strength of the linear relationship between each of the parameters and *QMED* ranges from a correlation coefficient score of 0.76 for *AREA* to -0.07 for *FARL*. The strong linear relationship between *QMED* and *AREA*, contrasts with the relative sensitivity scores presented later in this paper for the multiple linear regression model, and in so doing emphasises the additional insights provided by sensitivity analysis over basic statistical measures.

Model skill

Figures 3– 5 present scatter diagrams of observed versus modelled index flood values for the GLM, ANN_A and ANN_B models. The full dataset is depicted in each scatter plot. Figures 3 and 4 reveal comparable amounts of predictive skill for the GLM and ANN_A model. Both plots, indeed, appear to show a reasonable degree of model performance at lower levels, but typically under-estimate the higher magnitude flood events. In contrast the ANN_B model appears to perform well across the range of flood event magnitudes and seems very close to correctly modelling the two largest flood events.

Although Figures 3, 4, and 5 provide an interpretive view of the accuracy of the three models, Table 4 provides a more objective, numerical contrast by providing comparative performance statistics for each of the models. It shows that while the ANN_B model is undoubtedly the most accurate overall according to the RMSE and R^2 measures, the GLM is more accurate at modelling low flood indices. Although there appears to be a significant difference between the MSRE statistics of the GLM and the ANN_A model (0.19 and 16.12, respectively) these results need to be treated with caution. A very basic model, that simply predicts the index flood for every catchment as 1 m^3 s^{-1}, results in a MSRE statistic of 0.93 – better than both the ANN models and not too dissimilar from the GLM. One would not seriously contemplate using such a simple model as a prediction of the index flood in an ungauged catchment so it brings into question the suitability of the MSRE as an appropriate measure of performance. It indicates that a model needs to make only a handful of errors at lower levels (which may not be too far from the observed values) to result in a poor MSRE result. This emphasises the importance of using multiple evaluation criteria and understanding the limitations of individual error measures.

Although the scatter diagrams show reasonably similar performance at lower levels, one or two over/under predictions have skewed the results. A more appropriate measure of performance at lower levels is perhaps the MSLE used by Pokhrel *et al.* (2012), the results of which are also presented in Table 4. In this case, although the GLM outperforms the ANN_A and ANN_B models, the results are not too dissimilar. For the simple model (producing 1 m^3s^{-1} for each case) the MSLE is calculated as 15.36 – significantly higher than the more complex models. Given that the ANN_B performs reasonably well for low *QMED* values and better than the GLM at large *QMED* values where prediction is normally more problematic, the goodness-of-fit statistics suggest that ANN_B could be considered a reasonable alternative to GLM.

SENSITIVITY ANALYSIS AND PHYSICAL INTERPRETATION OF MODELS

GLM

Relative sensitivity plots for the GLM are provided in Figure 6 are calculated using Equations (13)–(16). *AREA* and *FARL* are both used as simple scaling variables in the model such that the index flood magnitude increases proportionally for larger catchments with lower flood attenuation. The model behaves in a manner that larger catchments produce consistently larger floods, but the overall significance of this behaviour is relatively small. In a simplistic, conceptual sense, this is physically legitimate behaviour and one would expect the catchment area to act as a proportionally consistent driver of flood magnitude with a ratio close to unity, as a larger catchment will have proportionally greater hydrological inputs. Importantly, *FARL* as a driver, is shown to be around four times more important than *AREA*; a pattern that perhaps highlights the overriding importance of in-channel buffering of flood peaks by lakes and reservoirs in the model.

SAAR and *BFIHOST* function as more complex drivers of *QMED* and their relative sensitivities vary considerably. Indeed, in certain data ranges each has the potential to become the most influential driver of index flood magnitude. However, their specific patterns of relative sensitivity prove difficult to legitimise in simplified, physical terms. The proportionally greater sensitivity of index flood magnitude to increases in wetness in low rainfall catchments, as opposed to ones possessing high rainfall, does not correspond well with broad hydrological notions. The expectation would be to find low antecedent moisture in low rainfall catchments to result in enhanced infiltration, reduced propensity for Hortonian overland flow and correspondingly lower index flood sensitivity compared to higher rainfall catchments. This suggests that there is a substantive runoff buffering mechanism in wet catchments that is not present in dry ones. Whilst one may postulate that factors such as different vegetation types in dry and wet catchments may buffer flood responses differently, it is difficult to envisage their impact being sufficient to produce the magnitude of difference observed in the relative sensitivity plot. Moreover, the pattern appears counter to notions of antecedent moisture which would be expected to be lower in dry catchments and, therefore, would act to proportionally reduce catchment runoff and index flood magnitude.

Similarly, the sensitivity of the index flood to catchment permeability is counter to basic physical principles with index floods seen to be an order of magnitude more sensitive to a unit change in permeability in a highly permeable catchment when compared with the same proportional change in an impermeable one. Whilst the overall negative relative sensitivity of *QMED* to *BFIHOST* is conceptually legitimate, the specific pattern is difficult to legitimise physically as is the magnitude of the relative sensitivity observed relative to that of the other variables.

The sensitivity analysis thus indicates only partial physical legitimacy of the GLM, with the pattern of sensitivity of *QMED* to *SAAR* and *BFIHOST* being particularly difficult to rationalise.

ANN_A

Relative sensitivity plots for the ANN_A model are provided in Figure 7. Importantly, none of the plots exhibit the extreme, localised sensitivity variability that one would expect from an over-fitted model (see ANN_B below), which in the context of the model skill statistics reported above, suggests ANN_A offers a reasonable solution. ANN_A is characterised by generally lower relative sensitivity values in comparison to those observed for the GLM, coupled with enhanced complexity in the sensitivity responses across each variable's data range, the form of which is strongly influenced by the values of the other variables.

The relatively high sensitivity of *QMED* to *AREA* highlights the central importance of catchment size as a determinant of index flood magnitude in this model. This pattern of behaviour is an approximate counterpart of the GLM plot. Relative sensitivity remains roughly consistent at a value close to 1 and *AREA* is seen to act as a scaling variable in a physically-legitimate manner. However, the same degree of legitimacy is not observed in either the low or high *QMED* potential plots. Here opposing trends in the relative sensitivity are observed. When all other inputs are set to high *QMED* potential, proportional changes in catchment area of small catchments is seen to have almost 10 times the impact on *QMED* than the same proportional change in large catchments. The pattern reverses when inputs are set to low *QMED* potential. This model behaviour is very difficult to legitimise in physical terms.

Low values associated with *BFIHOST* highlight the general insensitivity of *QMED* to catchment permeability in this model. As expected, *BFIHOST* has a generally negative influence on *QMED* such that as permeability increases, *QMED* reduces. A general increase in *QMED*'s sensitivity to *BFIHOST* is observed as the other inputs are set to increasing levels of *QMED* potential. This indicates an increased importance of permeability as a constraint on index flood magnitude in catchments with high potential for generating large index floods. However, the very low magnitude of the sensitivities observed makes it difficult to draw any clear conclusions about the physical legitimacy of the patterns observed beyond the fact that *BFIHOST* is clearly not a particularly important driver of *QMED*.

In contrast to the GLM, *FARL* acts as a relatively modest driver of *QMED*, indicating that the ANN_A model is less heavily influenced by in-channel controls of peak discharge magnitude than the GLM. In simplistic physical terms, one would expect a reduction in flood attenuation to drive a proportional increase in *QMED*, and the positive relative sensitivity plots confirm this basic assumption. However, the precise form of the sensitivity relationship between *QMED* and *FARL* is more difficult to legitimise. The GLM represents the relationship as one of simple scaling and this same basic pattern exists for low and median *QMED* potential plots across medium to high *FARL* data ranges (i.e. medium to low levels of attenuation) where relative sensitivity is consistently about 0.5. However, at lower *FARL* data ranges the proportional response of *QMED* to change in *FARL* reduces substantially to 0.1. When other inputs are set to high *QMED* potential, the decreasing trend is consistent across all *FARL* ranges. This is less easily rationalised and is most likely attributable to the scarcity of catchments with low *FARL* values in the data resulting in a lack of data constraint on the form of the ANN model covering this data range, irrespective of the values of the other inputs.

The pattern of sensitivities observed for *SAAR* can only be partially legitimised in generalised physical terms. At a very simplistic level, the scaling behaviour of *SAAR* observed in the low *QMED* potential plot is perhaps reasonable given that proportionally wetter catchments should indeed result in proportionally greater floods. However, the patterns observed in the median and high *QMED* potential plots possess elements that are both physically rational and irrational. The increasing sensitivity to *SAAR* at low and mid data ranges could feasibly be explained in terms of antecedent moisture. Indeed, the on-average lower antecedent moisture in dry catchments could be expected to result in a smaller proportion of the rainfall contributing to runoff; leading to reduced hydrograph flashiness and proportionally lower *QMED* sensitivity to *SAAR* in dryer catchments. Similarly, the decline in sensitivity in the upper data ranges could be argued to be due to the fact that the catchment is already so wet that any additional rainfall makes relatively little difference to the index flood. However, this explanation ignores the role of overland, Hortonian flow in saturated, wet catchments which one would expect to drive an increase in the relative sensitivity in the upper data ranges. Finally, the negative relative sensitivity observed in the extreme upper ranges of the high *QMED* potential plot is physically-irrational as it suggests that proportionally increasing the catchment wetness will reduce the proportional response in *QMED*; in extreme cases even resulting in a reduction in *QMED*.

For each of the model inputs the behaviour of the ANN_A model is seen to be particularly influenced by the states of the input variables. When these are set to their median values (i.e. indicative of median *QMED* potential), the majority of the relative sensitivity plots indicate that the response function produces a model behaviour that can be physically-legitimised. However, this legitimacy is less certain when other variables are set at their 25th percentile values (i.e. indicative of low *QMED* potential) and completely breaks down when set at their 75th percentile value (i.e. indicative of high *QMED* potential). Indeed, under the latter condition, *AREA*, *FARL* and *SAAR* drive *QMED* in a manner that is particularly difficult to explain in hydrological terms. Crucially then, a link can be made between the lack of physical legitimacy in the model's behaviour in the upper and lower quartiles of the solution space and a lack of coincident data points which exist there to constrain the form of the ANN model.

ANN$_B$

Relative sensitivity plots for the ANN$_B$ model are provided in Figure 8. This ANN model is intentionally over-fitted and the impact of this over-fitting is clearly seen in the relative sensitivity plots. The degree of local variability in relative sensitivity is highly exaggerated when compared to ANN$_A$ with variables switching between both negative and positive responses in *QMED* at different data ranges. *QMED* responds to *AREA* and *SAAR* (the most influential drivers in the model) in an irrational manner with high magnitude, localised variation in relative sensitivity being particularly characteristic of the patterns observed. The relative sensitivity plots of *QMED* to *AREA* and *SAAR* are characterised by complex polynomial forms with no consistent trends in the relationship. The patterns observed are indicative of data over-fitting and lack any physical legitimacy.

Relative sensitivity of *QMED* to *FARL* behaves in a more constrained manner than *AREA* or *SAAR*, ranging from +0.8 to -0.3 indicating the relative lack of sensitivity to this variable in ANN$_B$. However, the sensitivity plots for low and median *QMED* potential show both positive and negative responses at different data ranges. Indeed, these plots suggest that in certain data ranges, a proportional decrease in flood attenuation will see a proportional reduction in flood magnitude: a result that lacks physical legitimacy. The high *QMED* potential plot is very similar to that of ANN_A

Relative sensitivity of *BFIHOST* to *QMED* is very muted with this variable being an almost irrelevant driver of index flood magnitude when other variables are set to low and median *QMED* potential. Localised complexity in the relative sensitivity is observed, particularly across low *BFIHOST* values where low and median *QMED* potential plots switch between positive and negative relative sensitivity values in a physically-irrational manner. The high *QMED* potential plot is perhaps more rational as it displays a flatter, negative response which indicates a negative scaling behaviour.

In contrast with ANN_A, local variation in relative sensitivity for *AREA* and *SAAR* becomes highly exaggerated when other variables are held at their low *QMED* potential values. This again highlights difficulties of fitting a 'bottom heavy' physically-legitimate ANN model, through upper regions of a solution space that lack sufficient coincident higher magnitude data points to constrain the form of the model.

Physical legitimacy

The broad physical legitimacy of the different model sensitivity plots are compared in Table 5. It is clear that none of the models behave in a manner that can be physically rationalised for all input variables. The GLM displays a basic level of physical legitimacy in the behaviour of *AREA* and *FARL* but this is lacking for *SAAR* and *BFIHOST* drivers. ANN_A displays varying degrees of physical legitimacy in the sensitivity between *QMED* and each of the input variables, with the least rational responses occurring when other variables are set to the high *QMED* potential values. However, in all cases, when other variables are set to their median values, the relative sensitivities of the ANN are physically legitimate at least in part.

Indeed, in this sense ANN_A arguably performs better than its GLM counterpart albeit delivering slightly less favourable goodness-of-fit. ANN_B is over-fitted and the patterns observed in its relative sensitivity plots cannot be legitimised in a physical sense. However, this lack of model legitimacy is in contrast to the goodness-of-fit statistics which indicate ANN_B to be the best model. Thus, developing techniques that can deliver a clear physical or mechanistic interpretation of input relative sensitivity analysis patterns in ANN modelling scenarios represents an important consideration for future research. Indeed, the presented results serve as a clear demonstration of the dangers associated with evaluating models on the basis of statistical performance validation approaches alone.

SUMMARY AND CONCLUSIONS

This paper has addressed the difficult question of how to make meaningful comparisons between artificial neural network-based hydrological models and alternative modelling approaches. Comparisons which are based solely on goodness-of-fit metrics (i.e. the standard black-box approach presented in much of the literature) are very limited because they only consider model performance and not the means by which the performance is obtained. The commonly encountered limitation of metric equifinality, in which metric scores for the models being compared are insufficiently different to enable conclusive differentiation of the best or preferred model, is evident in our results. Our example of median flood modelling provides a clear demonstration of this with the fit scores obtained by the ANN and GLM models delivering inconclusive evidence about relative overall model performance.

However, the limitations of goodness-of-fit metrics are arguably more fundamental if there is a requirement to compare the transferability of each model from one hydrological context to another. In such cases, the physical legitimacy of each model must also be evaluated and compared in a direct manner. Models used in ungauged catchment prediction are a good example of those that must ultimately be transferred, and that therefore require evaluation of their physical legitimacy. This study has presented a consistent means by which the physical legitimacy of ANN models can be evaluated and compared with alternative modelling approaches. The application of relative sensitivity analysis in our median flood modelling example has enabled the physical legitimacy of two ANN-based models to be compared directly with the GLM counterpart used as standard in the UK. Tables 4 and 5 provide clear evidence that a general ANN modelling approach can deliver models as good as the GLM approach currently used in the UK Flood Estimation Handbook, both in terms of their performance and their legitimacy. Whilst the paper does not purport to be a competition between ANNs and GLMs, in this isolated case the evidence does lend some support to the view that ANN-based models may have some advantages over their GLM counterparts. However, one can only build good physically-legitimate ANN models if ample data of sufficient quality exist, and if the model development process is sound. It is also evident from this evaluation that ANN solutions can only deliver physical legitimacy if issues such as overfitting are avoided.

To conclude it is clear that comparing ANN models to alternative approaches on the basis of goodness-of-fit is insufficient, and that sensitivity analysis offers an important means by which the physical legitimacy of ANN models can be compared with that of counterpart models. Indeed, hydrological modellers using ANNs can and should be striving to evaluate the physical legitimacy of their models as well as their performance. By applying sensitivity analysis to ANN models a sense of trust is introduced that goes part of the way to addressing one of the key issues in the international ANN river forecasting research agenda of Abrahart *et al.* (2012a), specifically the need for advanced diagnostic techniques that can help counter criticisms of the black-box nature of such models (e.g. Babovic, 2005). It is, therefore, surprising that it remains almost entirely absent from ANN studies and highlights the importance of a broader research agenda to develop robust, computational sensitivity analysis methods across the range of data-driven techniques currently being used in hydrological modelling. Such an agenda should include additional investigations that more fully explore the impact of different architectural structures in ANN models especially the potential bearing that internal complexity might have on the relative sensitivity of solutions to particular types of hydrological modelling problem.

REFERENCES

Abrahart, R.J., Anctil, F., Coulibaly, P., Dawson, C.W., Mount, N.J., See, L.M., Shamseldin, A.Y., Solomatine, D.P., Toth, E. & Wilby, R.L. 2012a Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting. *Prog. Phys. Geog.* **36** 480-513.

Abrahart, R.J., Dawson, C.W. & Mount, N.J. 2012b Partial derivative sensitivity analysis applied to autoregressive neural network river forecasting. In: *Hydroinformatics 2012: Proc. Tenth Int. Conf. on Hydroinformatics, Hamburg, Germany, 14-18 July 2012* (R. Hinkelmann, M.H. Nasermoaddeli, S.Y. Liong, D. Savic, P. Fröhle & K.F. Daemrich, eds.) [digital: eight page document]. [DOI?]

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

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

Abrahart, R.J., See, L.M., Dawson, C.W., Shamseldin, A.Y. & Wilby, R.L. 2010 Nearly Two Decades of Neural Network Hydrologic Modeling. In: *Advances in Data-Based Approaches for Hydrologic Mod. and For.* (B. Sivakumar & R. Berndtsson, eds). World Scientific Publishing, Hackensack, NJ, USA, 267-347.

Abrahart, R.J., See, L.M. & Kneale, P.E. 1999 Using pruning algorithms and genetic algorithms to optimise neural network architectures and forecasting inputs in a neural network rainfall-runoff model. *J. Hydroinform*. **1** 103-114.

American Society of Civil Engineers 2000a Artificial neural networks in hydrology. I: Preliminary concepts. *ASCE J. Hydrol. Eng.* **5** 115-123.

American Society of Civil Engineers 2000b Artificial neural networks in hydrology. II: Hydrologic applications. *ASCE J. Hydrol. Eng.* **5** 124-137.

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

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

Babovic, V. 2005 Data mining in hydrology. *Hydrol. Proc.* **19** 1511-1515.

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

Boorman, D.B., Hollis, J.M. & Lilly, A. 1995 *Hydrology of Soil Types: a hydrologically-based classification of the soils of the United Kingdom*. Institute of Hydrology Report 126, Institute of Hydrology, Wallingford, UK.

Caswell, H. 1976 The validation problem. In: *Systems Analysis and Simulation in Ecology*, Vol. IV., Academic Press, New York, 313-325.

Dastorani, M.T., Talebi, A. & Dastorani, M. 2010 Using neural networks to predict runoff from ungauged catchments. *Asian J. App. Sci.* **3** 399-410.

Dastorani, M.T. & Wright, N.G. 2001 Application of artificial neural networks for ungauged catchment flood prediction. *Floodplain Management Association Conference*, San Diego, CA, March 2001.

Dawson, C.W., Abrahart, R.J. & See, L.M. 2007 HydroTest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. *Environ. Mod. Software* **22** 1034–1052.

Dawson, C.W., Abrahart, R.J. & See, L.M. 2010 HydroTest: further development of a web resource for the standardised assessment of hydrological models. *Environ. Mod. Software.* **25** 1481-1482.

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

Dawson, C.W. & Wilby, R.L. 2001 Hydrological modelling using artificial neural networks. *Prog. Phys. Geog.* **25** 80-108.

Fernando, D.A.K. & Shamseldin, A.Y. 2009 Investigation of internal functioning of the radial-basis-function neural network river flow forecasting models. *J. Hydrol. Eng.* **14** 286–292.

Giustolisi, O. & Laucelli, D. 2005 Improving generalization of artificial neural networks in rainfall-runoff modelling. *Hydrol. Sci. J.* **50** 439-457.

Grover, P.L., Burn, D.H. & Cunderlik, J.M. 2002 A comparison of index flood estimation procedures for ungauged catchments. *Can. J. Civ. Eng.* **29** 734-741.

Hall, M.J. & Minns, A.W. 1998 Regional flood frequency analysis using artificial neural networks. In: *Hydroinformatics'98: Proc. Third Int. Conf. on Hydroinformatics, Copenhagen, Denmark, 24-26 August 1998* (V. Babovic & L.C. Larsen, eds.). A.A. Balkema, Rotterdam, The Netherlands. **2**, 759-763.

Hall, M.J., Minns, A.W. & Ashrafuzzaman, A.K.M. 2000 Regionalisation and data mining in a data scarce environment. *Proc. Seventh Natl Hydrol. Sym. Newcastle upon Tyne, UK, 4-6 September 2000*, British Hydrological Society, London, pp 3.39-3.43.

Hamby, D. M. 1994 A review of techniques for parameter sensitivity analysis of environmental models. *Environ. Monit. Assess.* **32** 135-154.

Hashem, S. 1992 Sensitivity Analysis for Feedforward Artificial Networks with Differentiable Activation Functions. *Proc. Int. Joint Conf. Neural Networks, Baltimore, M.D., USA, 7-11 June 1992*, IEEE, N.J., USA. **1**, pp 419-424.

Hill, M.C. & Tiedeman, C.R. 2007 *Effective Groundwater Model Calibration with Analysis of Sensitivities, Predictions, and Uncertainty*, Wiley, New York.

Holvoet, K., van Griensven, A., Seuntjens, P. & Vanrolleghem, P.A. 2005 Sensitivity analysis for hydrology and pesticide supply towards the river in SWAT. *Phys. Chem. Earth* **30** 518-526.

Howes, S. & Anderson, M. G. 1988 *Computer simulation in geomorphology*. In: *Modeling Geomorphological Systems*, John Wiley and Sons Ltd, Chichester.

Institute of Hydrology 1999 *Flood Estimation Handbook* (5 Volumes). Institute of Hydrology, Wallingford, UK.

Jain, A. & Kumar, S. 2009 Dissection of trained neural network hydrologic model architectures for knowledge extraction. *Water Resour. Res*. **45** W07420, DOI:10.1029/2008WR007194.

Jain, A., Sudheer, K.P. & Srinivasulu, S. 2004 Identification of physical processes inherent in artificial neural network rainfall runoff models. *Hydro. Pro.* **18** 571–581.

Kattan, A., Abudullah, R. & Geem, Z.W. 2011 Artificial Neural Network Training & Software Implementation Techniques. Nova Science Publishers, New York.

Kingston, G.B., Maier, H.R. & Lambert, M.F. 2003 Understanding the mechanisms modelled by artificial neural networks for hydrological prediction. In: *Modsim 2003 – International Congress on Modelling and Simulation*, *Modelling and Simulation Society of Australia and New Zealand Inc, Townsville, Australia, 14-17th July*, **2**, 825-830.

Kingston, G.B., Maier, H.R. & Lambert, M.F. 2005 Calibration and validation of neural networks to ensure physically plausible hydrological modelling. *J. Hydrol*. **314** 158-176.

Kingston, G.B., Maier, H.R. & Lambert, M.F. 2006 A probabilistic method to assist knowledge extraction from artificial neural networks used for hydrological prediction. *Math. Comput. Model.* **44** 499-512.

Kingston, G.B., Maier, H.R. & Lambert, M.F. 2008 Bayesian model selection applied to artificial neural networks used for water resources modelling. *Water Resour. Res.* **44** W04419.

Kjeldsen, T.R. & Jones, D.A. 2009 An exploratory analysis of error components in hydrological regression modelling. *Water Resourc. Res.* **45** W02407 DOI:10.1029/2007WR006283.

Kjeldsen, T.R. & Jones, D.A. 2010 Predicting the index flood in unguaged UK catchments: on the link between data transfer and spatial model error structure. *J. Hydrol.* **387**1-9.

Kjeldsen, T.R., Jones, D.A. & Bayliss, A.C. 2008 *Improving the FEH statistical procedures for flood frequency estimation.* Science Report Number SC050050, Environment Agency, Bristol, UK.

Klemes, V. 1986 Operational testing of hydrological simulation models. *Hydrolog. Sci. J.* **31** 13-24.

Liong, S.Y., Nguyen, V.T.V., Chan, W.T. & Chia, Y.S. 1994 Regional estimation of floods for ungauged catchments with neural networks In: *Developments in Hydraulic Engineering and their Impact on the Environment*, *Proceedings Ninth Congress of Asian and Pacific Division of the International Association for Hydraulic Research , Singapore, 24–26 August 1994*, (H-F. Cheong, N.J. Shankar, E-S. Chan & W-J. Ng, eds.), pp 372-378.

Maier, H.R. & Dandy, G.C. 2000 Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environ. Mod. Software* **15** 101-123.

Maier, H.R., Jain, A., Dandy, G.C. & Sudheer, K.P. 2010 Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. *Environ. Mod. Software* **25** 891-909.

Marsh, T.J. & Hannaford, J. 2008 *UK Hydrometric Register.* Hydrological Data UK Series, Centre for Ecology and Hydrology, Wallingford, UK.

McCuen, R.H. 1973 The role of sensitivity analysis in hydrologic modelling. *J. Hydrol.* **18** 37-53.

Minns, A.W. & Hall, M.J. 1996 Artificial neural networks as rainfall-runoff models. *Hyd. Sci. J.* **41** 399-417.

Mishra, S. 2009 Uncertainty and sensitivity analysis techniques for hydrologic modelling. *J. Hydroinform*. **11** 282-296.

Mount, N.J. & Abrahart, R.J. 2011 Load or concentration, logged or unlogged? Addressing ten years of uncertainty in neural network suspended sediment prediction. *Hydrol. Proc.* **25** 3144-3157.

Mount, N.J., Abrahart, R.J., Dawson, C.W. & Ab Ghani, N. 2012 The need for operational reasoning in data-driven rating curve prediction of suspended sediment. *Hydrol. Proc.* **26** 3982-4000.

Mount, N.J., Dawson, C.W. & Abrahart, R.J. 2013 Legitimising neural network river forecasting models: a new data-driven mechanistic modelling framework. *Hydrol. Earth Syst. Sci.* In press.

Muleta, M. K. & Nicklow, J. W. 2005 Sensitivity and uncertainty analysis coupled with automatic calibration for a distributed watershed model. *J. Hydrol*. **306** 127–145.

Muttiah, R.S., Srinivasan, R. & Allen, P.M. 1997 Prediction of two-year peak stream discharges using neural networks. *J. Am. Wat. Res. Assoc.* **33** 625-630.

Natural Environment Research Council 1975 *Flood Studies Report*. Natural Environment Research Council, London, UK.

Nelson, R.W. 2011 *New developments in artificial neural networks research*. Nova Science Publishers, New York.

Nourani, V. & Fard, M.S. 2012 Sensitivity analysis of the artificial neural network outputs in simulation of the evaporation process at different climatologic regimes. *Adv Eng. Software* **47** 127-146.

Nourani, V., Komasi, M. & Alami, M. 2012 Hybrid wavelet-genetic programming approach to optimize ANN modeling of rainfall-runoff process. *J. Hydrol. Eng.* **17** 724-741.

Olden, J.D. & Jackson, D.A. 2002 Illuminating the ‘black box’: a randomization approach for understanding variable contributions in artificial neural networks. *Ecol. Model.* **154** 135-150.

Oreskes, N., Shrader-Frechette, K. & Belitz, K. 1994 Verification, validation and confirmation of numerical models in the Earth Sciences. *Science* **263** 641-646.

Pappenberger, F., Beven, K.J., Ratto, M. & Matgen, P. 2008 Multi-method global sensitivity analysis of flood inundation models. *Adv. Water Resour.* **31** 1-14.

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.

Pokhrel, P., Yilmaz, K. & Gupta, H. 2012 Multiple-criteria calibration of a distributed watershed model using spatial regularization and response signatures. *J. Hydrol.* **418-419** 49-60.

Radwan, M., Willems, P. & Berlamont, J. 2004 Sensitivity and uncertainty analysis for river quality modelling. *J. Hydroinform*. **6** 83-99.

Refsgaard, J.C. & Knudsen, J. 1996 Operational validation and intercomparison of different types of hydrological models. *Water Resour. Res*. **32** 2189-2202.

Robinson, S. 1997 Simulation model verification and validation: increasing the users' confidence. In: *Proceedings of the 1997 Winter Simulation Conference*, Atlanta, Georgia, 53-59.

Rodriguez-Iturbe, I. & Valdes, J.B. 1979 The geomorphologic structure of hydrologic response. *Water Resourc. Res.* **15** 1409-1420.

Rumelhart, D.E., Hinton, G.E. & Williams, R.J. 1986 Learning internal representations by error propagation. In: *Parallel Distributed Processing: Explorations in the Microstructures of Cognition Vol. 1* (D.E. Rumelhart, & J.L. McClelland, eds). MIT Press, Cambridge, MA, USA, pp 318–362.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M. & Tarantola, S. 2008 *Global Sensitivity Analysis*. *The primer*. Wiley, Chichester, 304.

Sargent, R.G. 2011 Verification and validation of simulation models. In: *Proceedings of the 2011 Winter Simulation Conference, Informs Simulation Society*, 183-197.

Schrieber, P. & Demuth, S. 1997 Regionalization of low flows in southwest Germany. *Hydrol. Sci. J.* **42** 845-858.

Schulz, K. & Huwe, B. 1999 Uncertainty and sensitivity analysis of water transport modelling in a layered soil profile using fuzzy set theory. *J. Hydroinform*. **1** 127-138.

See, L.M., Jain, A., Dawson. C.W. & Abrahart, R.J. 2008 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.

See, L.M. & Openshaw, S. 2000 A hybrid multi-model approach to river level forecasting. *Hydrol. Sci. J.* **45** 523–536.

Shrestha, R.R. & Nestmann, F. 2009 Physically-based and data-driven models and propagation of uncertainties in flood prediction, *J. Hydrolog. Eng.* **14** 1309-1319.

Spear, R.C. & Hornberger, G.M. 1980 Eutrophication in Peel Inlet, II, Identification of critical uncertainties via generalized sensitivity analysis. *Water Resour. Res.* **14** 43-49.

Spruill, C.A., Workman, S.R. & Taraba, J.L. 2000 Simulation of daily and monthly stream discharge from small watersheds using the SWAT model. *Am. Soc. Civ. Eng.* **43** 1431-1439.

Sudheer, K.P. 2005 Knowledge extraction from trained neural network river flow models. *ASCE J. Hydrol. Eng.* **10** 264-269.

Sudheer, K.P. & Jain, A. 2004 Explaining the internal behaviour of artificial neural network river flow models. *Hydro. Proc.* **18** 833–844.

Turanayi, T. & Rabitz, H. 2000 Local methods. In: *Sensitivity Analysis*, Wiley Series in Probability and Statistics, Wiley, Chichester.

Vogel, R.M. & Kroll, C.N. 1992 Regional geohydrologic-geomorphic relationships for the estimation of low-flow statistics. *Water Resourc. Res.* **28** 2451-2458.

Wilby, R.L., Abrahart, R.J. & Dawson, C.W. 2003 Detection of conceptual model rainfall-runoff processes insider. An artificial neural network. *Hydrol. Sci. J.* **48** 163–181.

Yeung, D.S., Cloete, I., Shi, D. & Ng, W.W.Y. 2010 *Sensitivity Analysis for Neural Networks*. Springer-Verlag, Berlin / Heidelberg, Germany.

Zhang, X., Hormann, G., Fohrer, N. & Gao, J. 2012 Estimating the impacts and uncertainty of changing spatial input data resolutions on streamflow situations in two basins. *J. Hydroinform*. **14** 902-917.

First received 20 November 2012; accepted in revised form 29 May 2013. Available online.

FIGURE CAPTIONS

Figure 1. Typical feed forward ANN structure

Figure 2. Scatter plot matrix of model variable with linear regression lines fitted

Figure 3. GLM versus *QMED*

Figure 4. ANN_A model versus *QMED*

Figure 5. ANN_B model versus *QMED*

Figure 6. Relative sensitivity of *QMED* to model inputs: GLM

Figure 7. Relative sensitivity of *QMED* to model inputs: ANN_A

Figure 8. Relative sensitivity of *QMED* to model inputs: ANN_B

TABLE CAPTIONS

Table 1. Statistical summary of catchment descriptors

Table 2. Network weights and biases. Input neurons I1 - I4 (AREA, BFIHOST, FARL, SAAR, respectively); Hidden neurons H1 – H9; Output neuron O (QMED)

ANNa

Table 3. Correlation matrix for model variables

Table 4. Numerical accuracy of different models under test

Table 5. Physical legitimacy of GLM and ANN models

Table 1. Statistical summary of catchment descriptors

	Median	Minimum	Maximum	25^{th} Percentile	75^{th} Percentile
AREA (km^2)	148.70	1.63	4586.97	68.00	327.81
BFIHOST	0.47	0.20	0.97	0.40	0.57
FARL	0.99	0.65	1.00	0.96	1.00
SAAR (mm)	1096	558	2848	830	1375
QMED	43.54	0.14	992.85	12.92	117.71

Table 2. Network weights and biases. Input neurons I1 - I4 (AREA, BFIHOST, FARL, SAAR, respectively); Hidden neurons H1 – H9; Output neuron O (QMED)

ANNa

		Weight			Weight			Weight			Weight			Weight
I1	H1	2.112	I2	H1	1.287	I3	H1	-1.858	I4	H1	-4.078	H1	O	-2.004
I1	H2	-0.211	I2	H2	-0.392	I3	H2	-1.591	I4	H2	-0.154	H2	O	-0.797
I1	H3	2.907	I2	H3	-6.502	I3	H3	2.196	I4	H3	4.048	H3	O	4.901
I1	H4	-1.170	I2	H4	2.792	I3	H4	-0.347	I4	H4	-3.403	H4	O	-1.904
I1	H5	0.245	I2	H5	-0.337	I3	H5	-2.473	I4	H5	0.521	H5	O	-1.001
I1	H6	0.009	I2	H6	-1.236	I3	H6	-1.627	I4	H6	0.087	H6	O	-0.533
I1	H7	-13.412	I2	H7	-4.484	I3	H7	1.478	I4	H7	2.806	H7	O	-7.586
I1	H8	-1.236	I2	H8	0.008	I3	H8	-0.782	I4	H8	-0.284	H8	O	-0.921
I1	H9	-6.588	I2	H9	-2.458	I3	H9	0.998	I4	H9	1.157	H9	O	-3.972

ANNb

		Weight			Weight			Weight			Weight			Weight
I1	H1	-1.877	I2	H1	20.295	I3	H1	0.185	I4	H1	-14.475	H1	O	-2.575
I1	H2	-16.987	I2	H2	-3.354	I3	H2	1.693	I4	H2	2.498	H2	O	-13.556
I1	H3	-3.798	I2	H3	-0.008	I3	H3	-2.085	I4	H3	-7.115	H3	O	4.112
I1	H4	5.559	I2	H4	-0.845	I3	H4	1.849	I4	H4	-18.273	H4	O	-4.311
I1	H5	-2.996	I2	H5	4.687	I3	H5	-6.742	I4	H5	6.914	H5	O	-1.337
I1	H6	8.318	I2	H6	-8.377	I3	H6	2.917	I4	H6	8.574	H6	O	4.750
I1	H7	8.324	I2	H7	-3.983	I3	H7	-3.674	I4	H7	10.392	H7	O	3.969
I1	H8	11.702	I2	H8	-19.838	I3	H8	-2.518	I4	H8	16.069	H8	O	-2.763
I1	H9	1.210	I2	H9	-3.488	I3	H9	-3.777	I4	H9	6.853	H9	O	-3.085

Biases

Neuron	Bias ANNa	Bias ANNb
H1	-0.596	-0.708
H2	-0.175	-1.927
H3	-3.240	0.049
H4	-0.315	-1.594
H5	0.413	2.982
H6	-0.098	-7.794
H7	-1.459	-0.996
H8	-0.508	0.627
H9	-0.720	0.278
O	0.282	1.707

Table 3. Correlation matrix for model variables

	AREA	*BFIHOST*	*FARL*	*SAAR*	*QMED*
AREA	1.00	-0.02	-0.07	-0.05	0.76
BFIHOST		1.00	0.11	-0.42	-0.27
FARL			1.00	-0.39	-0.07
SAAR				1.00	0.24

Table 4. Numerical accuracy of different models under test

	GLM	**ANN_A**	**ANN_B**
RMSE ($m^3\ s^{-1}$)	43.09	47.49	33.18
R^2	0.89	0.88	0.94
MSRE	0.19	16.12	1.91
MSLE	0.13	0.51	0.33

Table 5. Physical legitimacy of GLM and ANN models

Input Variable	*QMED* potential of other catchment variables	Does the pattern of sensitivity response conform to conceptual notions of physically-rationality?		
		GLM	ANN_A	ANN_B
AREA	Low		No	No
	Median	Yes	Yes	No
	High		No	No
SAAR	Low		Yes	No
	Median	No	In Part	No
	High		No	No
FARL	Low		In Part	No
	Median	Yes	In Part	No
	High		No	No
BFIHOST	Low		No	No
	Median	No	In Part	No
	High		In Part	In Part

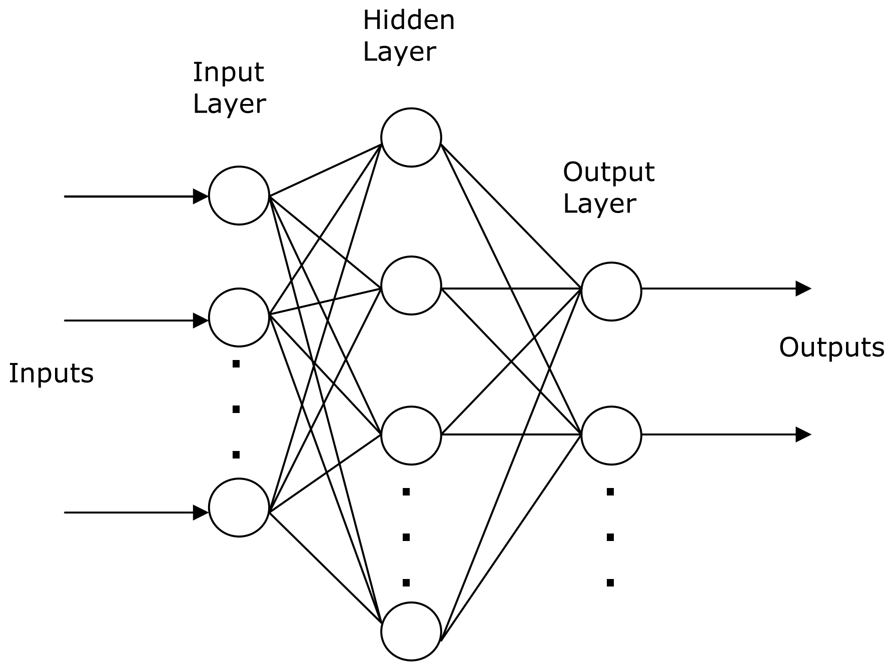

Figure 1. Typical feed forward ANN structure

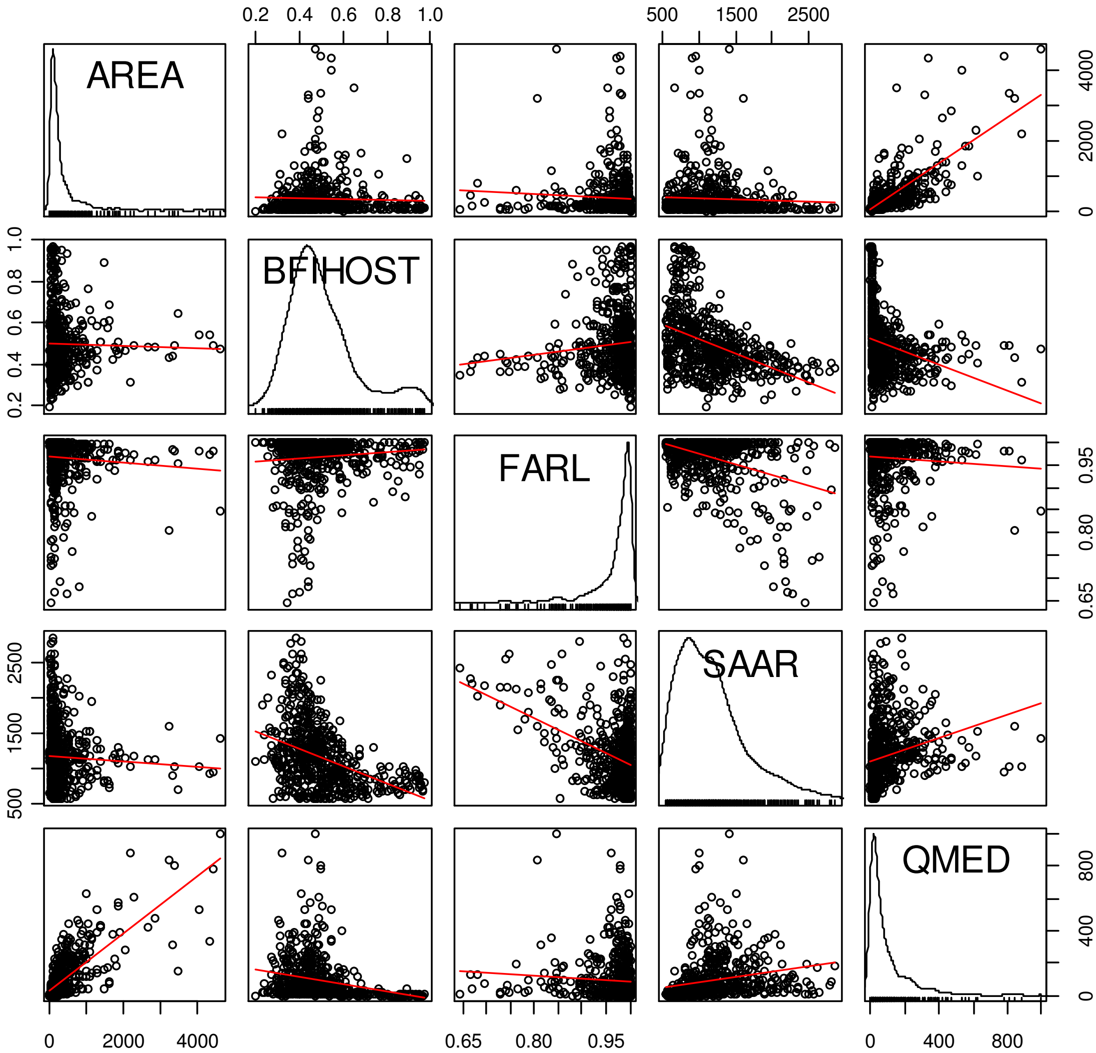

Figure 2. Scatter plot matrix of model variable with linear regression lines fitted

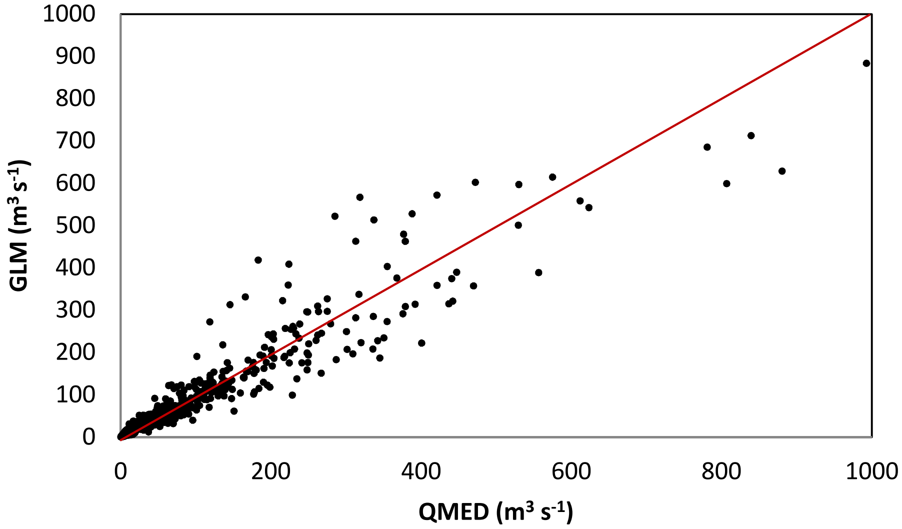

Figure 3. GLM versus *QMED*

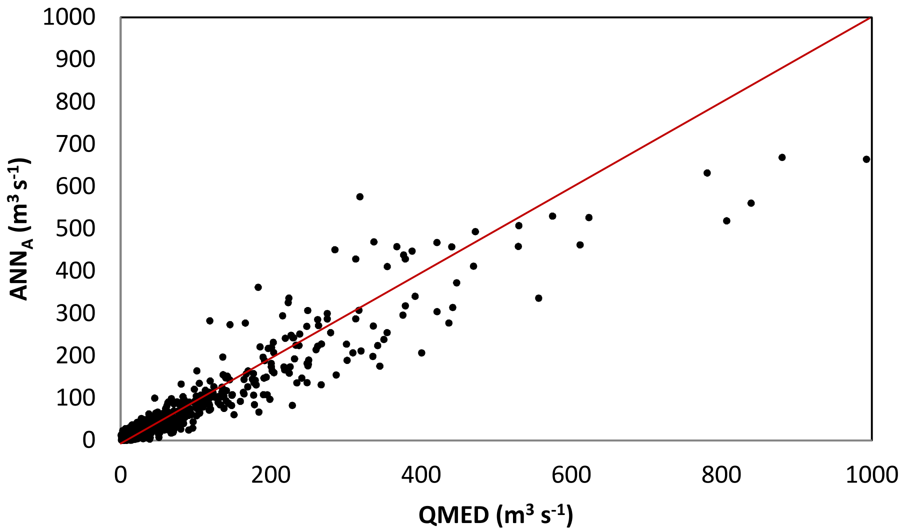

Figure 4. ANN$_A$ model versus *QMED*

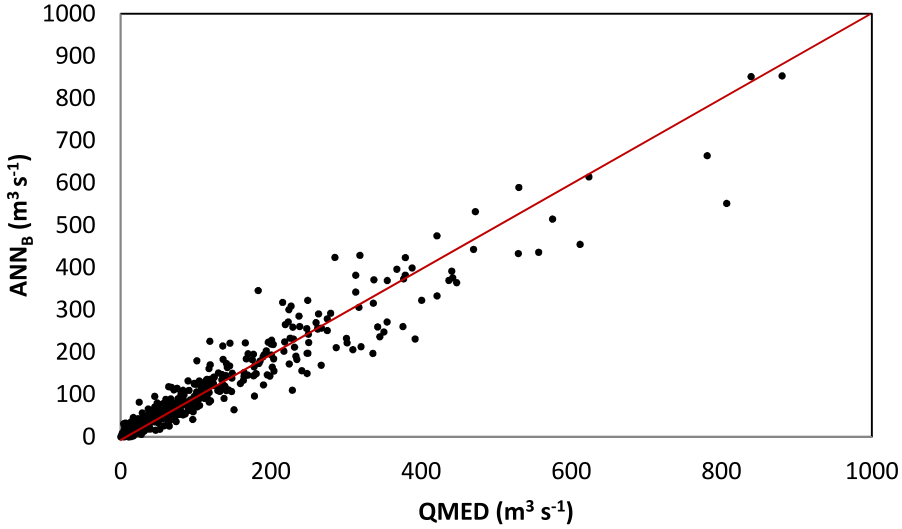

Figure 5. ANN$_B$ model versus *QMED*

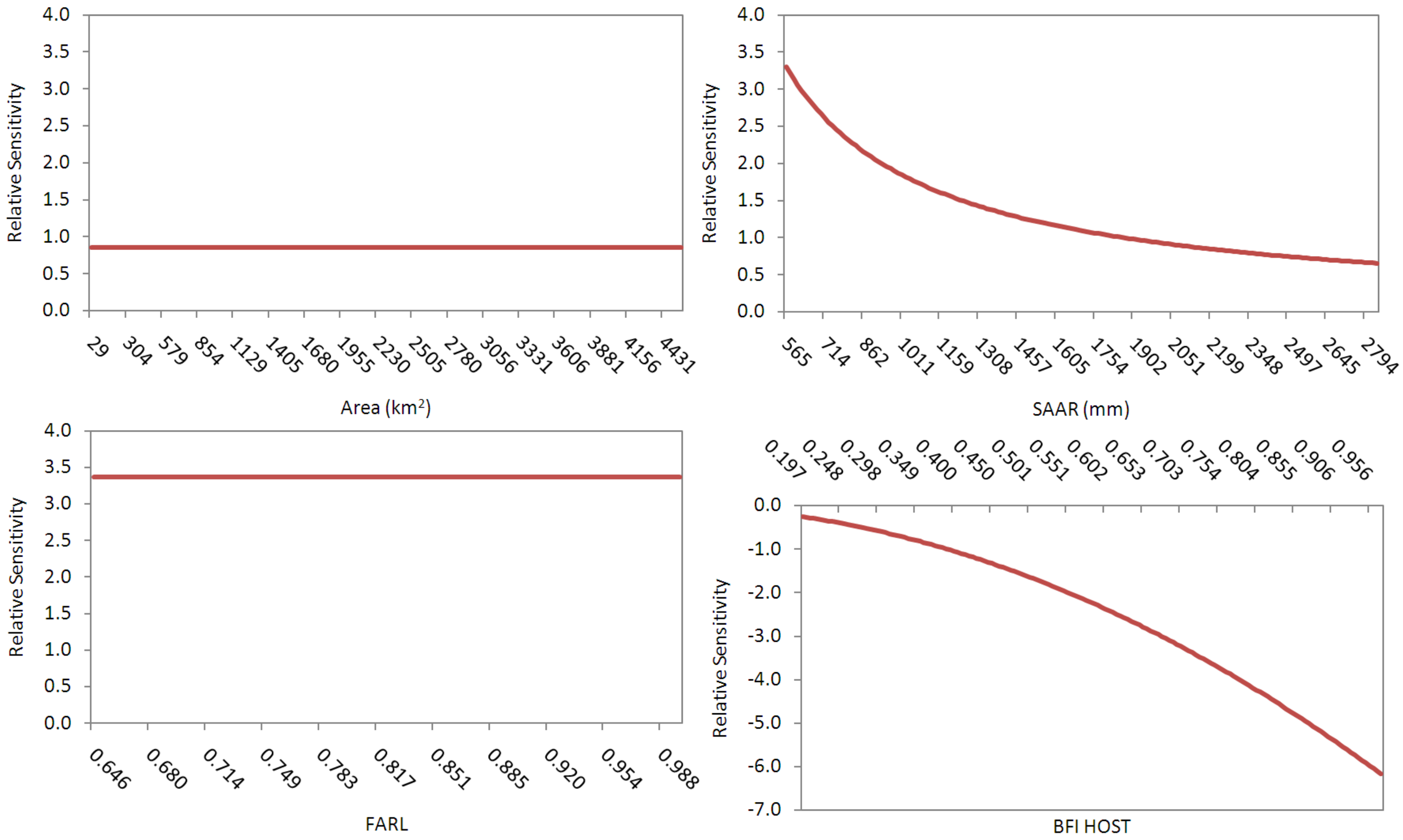

Figure 6. Relative sensitivity of *QMED* to model inputs: GLM

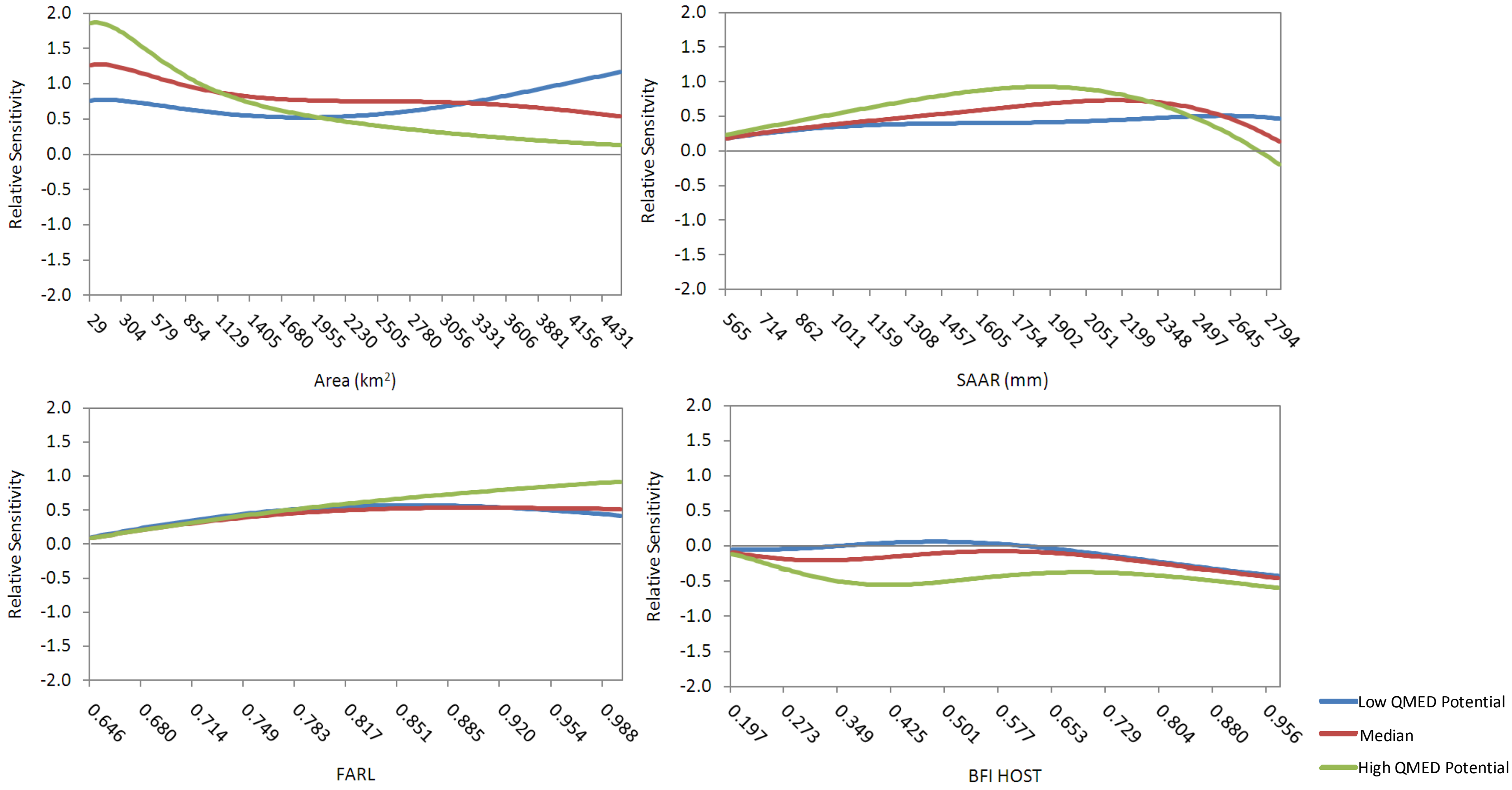

Figure 7. Relative sensitivity of *QMED* to model inputs: ANN_A

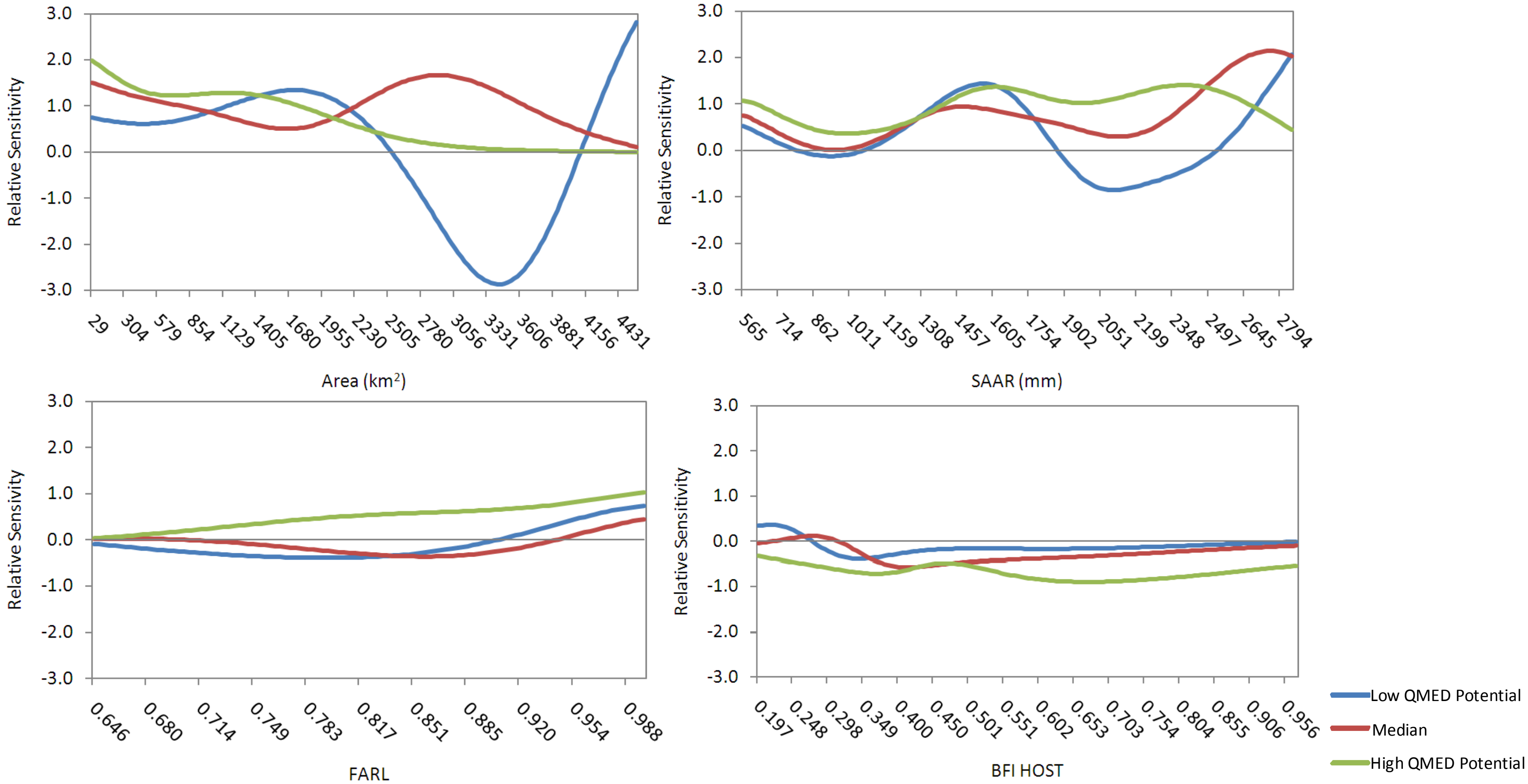

Figure 8. Relative sensitivity of *QMED* to model inputs: ANN_B