



Dawson, C.W. and Mount, Nick J. and Abrahart, R.J. and Louis, J. (2014) Sensitivity analysis for comparison, validation and physical-legitimacy of neural network-based hydrological models. *Journal of Hydroinformatics*, 16 (2). pp. 407-424. ISSN 1464-7141

Access from the University of Nottingham repository:

<http://eprints.nottingham.ac.uk/28054/1/HYDRO-D-12-00222.pdf>

Copyright and reuse:

The Nottingham ePrints service makes this work by researchers of the University of Nottingham available open access under the following conditions.

- Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners.
- To the extent reasonable and practicable the material made available in Nottingham ePrints has been checked for eligibility before being made available.
- Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.
- Quotations or similar reproductions must be sufficiently acknowledged.

Please see our full end user licence at:

http://eprints.nottingham.ac.uk/end_user_agreement.pdf

A note on versions:

The version presented here may differ from the published version or from the version of record. If you wish to cite this item you are advised to consult the publisher's version. Please see the repository url above for details on accessing the published version and note that access may require a subscription.

For more information, please contact eprints@nottingham.ac.uk

30 **KEYWORDS** | generalised model, index flood, neural network, partial derivative, sensitivity

31 analysis, ungauged catchment

32

33 INTRODUCTION

34 This paper presents an approach for delivering greater meaning from the comparison of
35 artificial neural network (ANN) models with alternative modelling approaches in
36 hydrological studies. ANN-based hydrological models are most commonly applied as black-
37 box tools and the internal mechanisms by which the model output is generated are not
38 normally explored in hydrological terms. Used in this way, an ANN's primary purpose is the
39 optimisation of complex, non-linear relations between a specific set of hydrological input
40 and output data, and standard goodness-of-fit procedures may, therefore, be considered an
41 adequate basis by which to compare its performance to that of other models (Klemes, 1986;
42 Refsgaard and Knusden, 1996). Indeed, assessments of goodness-of-fit have been widely
43 used in comparative hydrological modelling studies to argue that ANN models can perform
44 as well as, or better than alternative modelling approaches (e.g. Shrestha and Nestmann,
45 2009; Mount and Abrahart, 2011). However, such arguments are informed solely by the
46 degree of optimisation that is achieved by each model. They say nothing about the means
47 by which different models achieve their performance and the relative merits of these
48 alternative means. Indeed, when ANN models are applied solely as black-boxes, their
49 potential relative to other modelling approaches can never be properly understood in a
50 generalised or transferrable manner because the extent to which their modelling
51 mechanisms conform to physically-based, hydrological domain knowledge remains untested
52 (Howes and Anderson, 1988; Sargent, 2011). Consequently, critical questions about
53 whether ANN modelling mechanisms are more or less reflective of real-world hydrological
54 processes than alternative models are seldom addressed directly (Minns and Hall, 1996;
55 Abrahart *et al.*, 2011), and the relative extent to which they are able to deliver hydrological

56 process insights (i.e. Caswell's (1976) model duality) is not normally evaluated. The purpose
57 of this paper is to present a method by which these questions may be addressed.

58 More informative approaches to model comparison are required that explicitly
59 consider the internal behaviours of the different models and assess them according to their
60 conformance with the logical, rational and physical expectations of the modeller (c.f.
61 Robinson, 1997). This process is termed model legitimisation and is discussed in a
62 philosophical context by Oreskes *et al.* (1994) and an applied, hydrological modelling
63 context by Mount *et al.* (in press). Sensitivity analysis (Hamby, 1994) is an important and
64 effective means by which the legitimacy of a hydrological model may be explored. It has
65 been widely applied in conceptual and physically-based modelling over several decades (e.g.
66 McCuen, 1973; Beven and Binley, 1992; Schulz and Huwe, 1999; Radwan *et al.*, 2004;
67 Pappenberger *et al.*, 2008; Mishra, 2009; Zhang *et al.*, 2012). A variety of approaches have
68 been used including local (e.g. Turanayi and Rabitz, 2000; Spruill *et al.*, 2000; Holvoet *et al.*,
69 2005; Hill and Tiedeman, 2007), regional (e.g. Spear and Hornberger, 1980) and global-scale
70 methods (Muleta and Nicklow, 2005; Salteli *et al.*, 2008). By contrast, sensitivity analysis
71 has not been widely adopted in ANN modelling studies beyond a few, isolated examples
72 (Sudheer, 2005; Nourani and Fard, 2012). This is presumably because the equations that
73 relate inputs and outputs in an ANN are considered complex, inaccessible and difficult to
74 interpret (Aytek *et al.*, 2008; Abrahart *et al.*, 2009), making exploration of model sensitivity
75 via direct analysis of the governing equations difficult. Nonetheless, recent progress has
76 been made (Yeung *et al.*, 2010) and relative sensitivity analysis techniques for ANNs have
77 made it possible to assess the internal, mechanistic legitimacy of such models (Abrahart *et*
78 *al.*, 2012b; Mount *et al.*, in press). However, the focus of these studies has so far been
79 restricted to mechanical considerations. The application of sensitivity analysis to evaluate

80 the physical legitimacy of ANN-based hydrological models, and thus the degree to which
81 they can be generalised and transferred, remains an outstanding task.

82 In this paper, we apply a sensitivity analysis method that can be used to compare the
83 physical legitimacy of ANN-based hydrological models and alternative model counterparts in
84 a direct manner. We exemplify the method by comparing the performance and physical
85 legitimacy of a pair of ANN-based models and an established generalised linear model
86 (GLM) for median flood magnitude prediction in ungauged catchments in the UK. First
87 order, partial derivatives of each model's response function are computed, interpreted and
88 used as a consistent means by which the physical legitimacy of each model can be evaluated
89 and compared. This focus on response function behaviour is distinctly different to past
90 efforts to assess the physical legitimacy of ANN models, which have traditionally explored
91 internal structural components, such as weights (Abrahart *et al.*, 1999; Olden and Jackson,
92 2002; Anctil *et al.*, 2004; Kingston *et al.*, 2003,2005,2006,2008) and units (Wilby *et al.*, 2003;
93 Jain *et al.*, 2004; Sudheer and Jain, 2004; See *et al.*, 2008; Fernando and Shamseldin, 2009;
94 Jain and Kumar, 2009). However, the uniqueness of ANN structures means that the
95 information derived from them cannot easily be compared directly with that derived from
96 alternative models with different internal structures - thus limiting the comparative value of
97 the information. To overcome this problem, we here assess the physical legitimacy of an
98 ANN's overall response function using a standard relative sensitivity-based method that can
99 be consistently and directly replicated across a range of alternative model types and that is
100 widely understood and accepted by hydrologists. Consequently, an evaluation of the
101 physical legitimacy of the means by which each model's performance is obtained
102 accompanies the usual assessments of output validity; enabling the extent to which each
103 model delivers a transferable, general solution to be considered.

104

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

107 The modelling of hydrological responses in ungauged catchments remains an important
108 focus of research for hydrologists, especially as the majority of the world's river catchments
109 remain ungauged or poorly gauged. In such catchments, the application of distributed
110 physically-based models and statistical approaches is hampered by a lack of input parameter
111 knowledge and datasets. Consequently, lumped models which relate broad physiographic,
112 hydrogeologic and climatologic catchment descriptors to flood frequency curves, have long
113 been recognised as offering potential (Rodriguez-Iturbe and Valdes, 1979; Grover *et al.*,
114 2002).

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

121 Four catchment descriptors are used in the standard UK methodology: 1) *AREA*
122 (catchment area in km²); 2) *SAAR* (standard-period average annual rainfall in mm); 3) *FARL*
123 (flood attenuation due to reservoirs and lakes); 4) *BFIHOST* (baseflow index derived from
124 *HOST* data; Boorman *et al.*, 1995).

125 These catchment descriptors can be thought of as physical controls of *QMED* potential.
126 *SAAR* controls the hydrological inputs to the catchment, *AREA* controls the scaling of the

127 catchment response, whilst *BFIHOST* and *FARL* control the degree of buffering of the input-
128 output signal.

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

136 ANN models are also very effective at optimising complex, non-linear relations in
137 hydrological data (American Society of Civil Engineers 2000a,b; Maier and Dandy, 2000;
138 Dawson and Wilby, 2001; Maier *et al.*, 2010; Abrahart *et al.*, 2010; 2012b) and a number of
139 studies have highlighted their potential in ungauged catchment prediction (Liong *et al.*,
140 1994; Muttiah *et al.*, 1997; Hall and Minns, 1998; Hall *et al.*, 2000; Dastorani and Wright,
141 2001; Dawson *et al.*, 2006; Dastorani *et al.*, 2010). Indeed, the UK relationship between
142 *QMED* and catchment descriptors has also been modelled using ANNs and been shown to
143 deliver comparable levels of fit when compared to GLMs (Dawson *et al.*, 2006). However, it
144 remains unclear whether the two modelling approaches are similarly comparable with
145 respect to their physical legitimacy. Models with greater physical legitimacy should be more
146 generally transferrable to new catchment settings. Therefore, determining the physical
147 legitimacy of each model is an important element in delivering a physically informed
148 evaluation of how robustly it can be expected to transfer from the gauged catchments upon
149 which it is developed, to the ungauged catchments in which it is intended to be applied.

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

159

160 **Data**

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

174 undisturbed catchments, a strong association exists between Baseflow Index (derived from
175 archived gauged daily mean flows) and *BFIHOST*. The relationships between *QMED* and
176 *AREA*, *SAAR* and *FARL* are positive, whilst that between *QMED* and *BFIHOST* is negative.

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

186

187 **Model development procedures**

188 Three models were developed for comparison.

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

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

201

$$202 \quad QMED_{GLM} = 8.6704AREA^{0.8568}0.1550\left(\frac{1000}{SAAR}\right)FARL^{3.3662}0.0380^{BFIHOST} \quad (1)$$

203

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

213

$$214 \quad f(x) = \frac{1}{1+e^{-x}} \quad (2)$$

215

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

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

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

236 ANN_A was developed using the approach described in Dawson *et al.* (2006) in which
237 a large number of candidate ANNs are trained on a random subset of the data, partitioned
238 according to a 60% calibration to 40% cross-validation ratio. Although there is no agreed
239 standard for splitting the data, this ratio is widely accepted in hydrological modelling
240 (Mount and Abrahart, 2011; See and Openshaw, 2000). 180 candidate models containing 2,
241 3, 4, 5, 6, 7, 8, 9, 10 hidden units were developed with each candidate being trained for up
242 to 20,000 epochs in steps of 1,000, using a learning rate of 0.1 and a momentum value of
243 0.9. Each candidate model was cross-validated using the remaining 40% as a means of
244 preventing overfitting (Giustolisi and Laucelli, 2005; Piotrowski and Napiorkowski, 2013).

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

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

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

265

266 **MODEL PERFORMANCE AND PHYSICAL LEGITIMACY ASSESSMENT**

267 **Model performance evaluation**

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

$$277 \quad RMSE = \sqrt{\frac{\sum_{i=1}^n (Q_i - \hat{Q}_i)^2}{n}} \quad (3)$$

$$278 \quad R^2 = \left[\frac{\sum_{i=1}^n (Q_i - \bar{Q})(\hat{Q}_i - \tilde{Q})}{\sqrt{\sum_{i=1}^n (Q_i - \bar{Q})^2 \sum_{i=1}^n (\hat{Q}_i - \tilde{Q})^2}} \right]^2 \quad (4)$$

$$279 \quad MSE = \frac{1}{n} \sum_{i=1}^n (Q_i - \hat{Q}_i)^2 \quad (5)$$

$$280 \quad MSLE = \frac{1}{n} \sum_{i=1}^n (\ln Q_i - \ln \hat{Q}_i)^2 \quad (6)$$

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

283

284 **Physical legitimacy**

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

300

$$301 \quad RS_i = \frac{\partial O}{\partial I_i} \cdot \frac{I_i}{O} \quad (7)$$

302

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

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

309

$$310 \quad \frac{\partial O}{\partial I_i} = \sum_{j=1}^n w_{ij} w_{jO} h_j (1 - h_j) O (1 - O) \quad (8)$$

311

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

315 One important difference between calculating partial derivatives for multiple input,
316 single output GLMs and ANN models should, however, be noted. When computing partial
317 derivatives of a GLM, there is no need to vary the values of the other inputs to investigate
318 the range of sensitivity responses under different input conditions. This is because GLMs
319 deliver a simple additive response function, such that the relative sensitivity for any one
320 variable will involve only that variable, given that all other parts of the expression will cancel
321 out, during the process of scaling the other variables. Hence, relative sensitivity values for
322 each input to the QMED_{GLM} model (Equation 1) can be computed according to Equations
323 (9)–(12). The final relative sensitivities of the QMED_{GLM} model are provided in Equations
324 (13)–(16).

325

$$326 \quad \frac{\partial QMED}{\partial AREA} = \frac{0.8568 QMED}{AREA} \quad (9)$$

$$327 \quad \frac{\partial QMED}{\partial SAAR} = \frac{1864.05 QMED}{SAAR^2} \quad (10)$$

328
$$\frac{\partial QMED}{\partial FARL} = \frac{3.3662 QMED}{FARL} \quad (11)$$

329
$$\frac{\partial QMED}{\partial BFIHOST} = -6.5385 QMED .BFIHOST \quad (12)$$

330
$$RS_{AREA} = 0.8568 \quad (13)$$

331
$$RS_{SAAR} = 1864.05 / SAAR \quad (14)$$

332
$$RS_{FARL} = 3.3662 \quad (15)$$

333
$$RS_{BFIHOST} = -6.5385 BFIHOST^2 \quad (16)$$

334

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

341

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

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

346 Step 3: Plot results and interpret the resultant graphs.

347

348 Thus, physically speaking, if variable states in our study are held at the 25th
 349 percentile (or the 75th percentile in the case of the inverse *BFIHOST* measure), the resultant
 350 scenario under test is representative of relatively small, dry catchments with high

351 permeability and high flood attenuation: i.e. low catchment *QMED* potential. Conversely,
352 when variables states are held at the 75th percentile (with *BFIHOST* at the 25th percentile),
353 the resultant scenario under test will be representative of relatively large, wet catchments
354 with low permeability and low attenuation: i.e. high catchment *QMED* potential.

355

356 **RESULTS**

357 **Independence**

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

371

372

373 **Model skill**

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

381 Although Figures 3, 4, and 5 provide an interpretive view of the accuracy of the three
382 models, Table 4 provides a more objective, numerical contrast by providing comparative
383 performance statistics for each of the models. It shows that while the ANN_B model is
384 undoubtedly the most accurate overall according to the RMSE and R² measures, the GLM is
385 more accurate at modelling low flood indices. Although there appears to be a significant
386 difference between the MSRE statistics of the GLM and the ANN_A model (0.19 and 16.12,
387 respectively) these results need to be treated with caution. A very basic model, that simply
388 predicts the index flood for every catchment as $1 \text{ m}^3 \text{ s}^{-1}$, results in a MSRE statistic of 0.93 –
389 better than both the ANN models and not too dissimilar from the GLM. One would not
390 seriously contemplate using such a simple model as a prediction of the index flood in an
391 ungauged catchment so it brings into question the suitability of the MSRE as an appropriate
392 measure of performance. It indicates that a model needs to make only a handful of errors at
393 lower levels (which may not be too far from the observed values) to result in a poor MSRE
394 result. This emphasises the importance of using multiple evaluation criteria and
395 understanding the limitations of individual error measures.

396 Although the scatter diagrams show reasonably similar performance at lower levels,
397 one or two over/under predictions have skewed the results. A more appropriate measure of

398 performance at lower levels is perhaps the MSLE used by Pokhrel *et al.* (2012), the results of
399 which are also presented in Table 4. In this case, although the GLM outperforms the ANN_A
400 and ANN_B models, the results are not too dissimilar. For the simple model (producing $1 \text{ m}^3\text{s}^{-1}$
401 for each case) the MSLE is calculated as 15.36 – significantly higher than the more complex
402 models. Given that the ANN_B performs reasonably well for low *QMED* values and better
403 than the GLM at large *QMED* values where prediction is normally more problematic, the
404 goodness-of-fit statistics suggest that ANN_B could be considered a reasonable alternative to
405 GLM.

406

407 **SENSITIVITY ANALYSIS AND PHYSICAL INTERPRETATION OF MODELS**

408 **GLM**

409 Relative sensitivity plots for the GLM are provided in Figure 6 are calculated using Equations
410 (13)–(16). *AREA* and *FARL* are both used as simple scaling variables in the model such that
411 the index flood magnitude increases proportionally for larger catchments with lower flood
412 attenuation. The model behaves in a manner that larger catchments produce consistently
413 larger floods, but the overall significance of this behaviour is relatively small. In a simplistic,
414 conceptual sense, this is physically legitimate behaviour and one would expect the
415 catchment area to act as a proportionally consistent driver of flood magnitude with a ratio
416 close to unity, as a larger catchment will have proportionally greater hydrological inputs.
417 Importantly, *FARL* as a driver, is shown to be around four times more important than *AREA*;
418 a pattern that perhaps highlights the overriding importance of in-channel buffering of flood
419 peaks by lakes and reservoirs in the model.

420 *SAAR* and *BFIHOST* function as more complex drivers of *QMED* and their relative
421 sensitivities vary considerably. Indeed, in certain data ranges each has the potential to

422 become the most influential driver of index flood magnitude. However, their specific
423 patterns of relative sensitivity prove difficult to legitimise in simplified, physical terms. The
424 proportionally greater sensitivity of index flood magnitude to increases in wetness in low
425 rainfall catchments, as opposed to ones possessing high rainfall, does not correspond well
426 with broad hydrological notions. The expectation would be to find low antecedent moisture
427 in low rainfall catchments to result in enhanced infiltration, reduced propensity for
428 Hortonian overland flow and correspondingly lower index flood sensitivity compared to
429 higher rainfall catchments. This suggests that there is a substantive runoff buffering
430 mechanism in wet catchments that is not present in dry ones. Whilst one may postulate that
431 factors such as different vegetation types in dry and wet catchments may buffer flood
432 responses differently, it is difficult to envisage their impact being sufficient to produce the
433 magnitude of difference observed in the relative sensitivity plot. Moreover, the pattern
434 appears counter to notions of antecedent moisture which would be expected to be lower in
435 dry catchments and, therefore, would act to proportionally reduce catchment runoff and
436 index flood magnitude.

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

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

447

448

449 **ANN_A**

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

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

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

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

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

510 For each of the model inputs the behaviour of the ANN_A model is seen to be
511 particularly influenced by the states of the input variables. When these are set to their
512 median values (i.e. indicative of median *QMED* potential), the majority of the relative
513 sensitivity plots indicate that the response function produces a model behaviour that can be
514 physically-legitimised. However, this legitimacy is less certain when other variables are set
515 at their 25th percentile values (i.e. indicative of low *QMED* potential) and completely breaks

516 down when set at their 75th percentile value (i.e. indicative of high *QMED* potential). Indeed,
517 under the latter condition, *AREA*, *FARL* and *SAAR* drive *QMED* in a manner that is particularly
518 difficult to explain in hydrological terms. Crucially then, a link can be made between the lack
519 of physical legitimacy in the model's behaviour in the upper and lower quartiles of the
520 solution space and a lack of coincident data points which exist there to constrain the form of
521 the ANN model.

522

523

524 **ANN_B**

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

535 Relative sensitivity of *QMED* to *FARL* behaves in a more constrained manner
536 than *AREA* or *SAAR*, ranging from +0.8 to -0.3 indicating the relative lack of sensitivity to this
537 variable in ANN_B. However, the sensitivity plots for low and median *QMED* potential show
538 both positive and negative responses at different data ranges. Indeed, these plots suggest
539 that in certain data ranges, a proportional decrease in flood attenuation will see a

540 proportional reduction in flood magnitude: a result that lacks physical legitimacy. The high
541 *QMED* potential plot is very similar to that of ANN_A

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

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

554

555 **Physical legitimacy**

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

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

573

574 **SUMMARY AND CONCLUSIONS**

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

586 However, the limitations of goodness-of-fit metrics are arguably more fundamental
587 if there is a requirement to compare the transferability of each model from one hydrological

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

605 To conclude it is clear that comparing ANN models to alternative approaches on the
606 basis of goodness-of-fit is insufficient, and that sensitivity analysis offers an important
607 means by which the physical legitimacy of ANN models can be compared with that of
608 counterpart models. Indeed, hydrological modellers using ANNs can and should be striving
609 to evaluate the physical legitimacy of their models as well as their performance. By applying
610 sensitivity analysis to ANN models a sense of trust is introduced that goes part of the way to
611 addressing one of the key issues in the international ANN river forecasting research agenda

612 of Abrahart *et al.* (2012a), specifically the need for advanced diagnostic techniques that can
613 help counter criticisms of the black-box nature of such models (e.g. Babovic, 2005). It is,
614 therefore, surprising that it remains almost entirely absent from ANN studies and highlights
615 the importance of a broader research agenda to develop robust, computational sensitivity
616 analysis methods across the range of data-driven techniques currently being used in
617 hydrological modelling. Such an agenda should include additional investigations that more
618 fully explore the impact of different architectural structures in ANN models especially the
619 potential bearing that internal complexity might have on the relative sensitivity of solutions
620 to particular types of hydrological modelling problem.

621

622 REFERENCES

- 623 Abrahart, R.J., Anctil, F., Coulibaly, P., Dawson, C.W., Mount, N.J., See, L.M., Shamseldin,
624 A.Y., Solomatine, D.P., Toth, E. & Wilby, R.L. 2012a Two decades of anarchy? Emerging
625 themes and outstanding challenges for neural network river forecasting. *Prog. Phys.*
626 *Geog.* **36** 480-513.
- 627 Abrahart, R.J., Dawson, C.W. & Mount, N.J. 2012b Partial derivative sensitivity analysis
628 applied to autoregressive neural network river forecasting. In: *Hydroinformatics 2012:*
629 *Proc. Tenth Int. Conf. on Hydroinformatics, Hamburg, Germany, 14-18 July 2012* (R.
630 Hinkelmann, M.H. Nasermoaddeli, S.Y. Liong, D. Savic, P. Fröhle & K.F. Daemrich, eds.)
631 [digital: eight page document]. [DOI?]
- 632 Abrahart, R.J., Ab Ghani, N. & Swan, J. 2009 Discussion of “An explicit neural network
633 formulation for evapotranspiration”. *Hydrolog. Sci. J.* **54** 382-388.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

675 Fernando, D.A.K. & Shamseldin, A.Y. 2009 Investigation of internal functioning of the radial-
676 basis-function neural network river flow forecasting models. *J. Hydrol. Eng.* **14** 286–
677 292.

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

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

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

686 Hall, M.J., Minns, A.W. & Ashrafuzzaman, A.K.M. 2000 Regionalisation and data mining in a
687 data scarce environment. *Proc. Seventh Natl Hydrol. Sym. Newcastle upon Tyne, UK, 4-*
688 *6 September 2000*, British Hydrological Society, London, pp 3.39-3.43.

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

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

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

696 Holvoet, K., van Griensven, A., Seuntjens, P. & Vanrolleghem, P.A. 2005 Sensitivity analysis
697 for hydrology and pesticide supply towards the river in SWAT. *Phys. Chem. Earth* **30**
698 518-526.

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

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

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

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

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

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

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

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

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

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

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

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

730 Klemes, V. 1986 Operational testing of hydrological simulation models. *Hydrolog. Sci. J.* **31**
731 13-24.

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

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

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

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

745 McCuen, R.H. 1973 The role of sensitivity analysis in hydrologic modelling. *J. Hydrol.* **18** 37-
746 53.

747 Minns, A.W. & Hall, M.J. 1996 Artificial neural networks as rainfall-runoff models. *Hyd. Sci. J.*
748 **41** 399-417.

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

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

754 Mount, N.J., Abrahart, R.J., Dawson, C.W. & Ab Ghani, N. 2012 The need for operational
755 reasoning in data-driven rating curve prediction of suspended sediment. *Hydrol. Proc.*
756 **26** 3982-4000.

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

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

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

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

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

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

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

773 Olden, J.D. & Jackson, D.A. 2002 Illuminating the ‘black box’: a randomization approach for
774 understanding variable contributions in artificial neural networks. *Ecol. Model.* **154**
775 135-150.

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

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

780 Piotrowski, A.P. & Napiorkowski, J.J. 2013 A comparison of methods to avoid overfitting in
781 neural networks training in the case of catchment runoff modelling. *J. Hydrol.* **476** 97-
782 111.

783 Pokhrel, P., Yilmaz, K. & Gupta, H. 2012 Multiple-criteria calibration of a distributed
784 watershed model using spatial regularization and response signatures. *J. Hydrol.* **418-**
785 **419** 49-60.

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

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

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

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

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

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

801 Sargent, R.G. 2011 Verification and validation of simulation models. In: *Proceedings of the*
802 *2011 Winter Simulation Conference, Inform Simulation Society*, 183-197.

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

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

807 See, L.M., Jain, A., Dawson, C.W. & Abraham, R.J. 2008 Visualisation of hidden neuron
808 behaviour in a neural network rainfall-runoff model. In: *Practical Hydroinformatics:*
809 *Computational Intelligence and Technological Developments in Water Applications,*
810 Springer, Berlin, 87–99.

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

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

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

817 Spruill, C.A., Workman, S.R. & Taraba, J.L. 2000 Simulation of daily and monthly stream
818 discharge from small watersheds using the SWAT model. *Am. Soc. Civ. Eng.* **43** 1431-
819 1439.

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

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

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

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

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

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

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

835

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

837

838 **FIGURE CAPTIONS**

839 Figure 1. Typical feed forward ANN structure

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

841 Figure 3. GLM versus *QMED*

842 Figure 4. ANN_A model versus *QMED*

843 Figure 5. ANN_B model versus *QMED*

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

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

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

847

848

849 **TABLE CAPTIONS**

850 Table 1. Statistical summary of catchment descriptors

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

853 ANNa

854 Table 3. Correlation matrix for model variables

855 Table 4. Numerical accuracy of different models under test

856 Table 5. Physical legitimacy of GLM and ANN models

857

858

859

Table 1. Statistical summary of catchment descriptors

860

	Median	Minimum	Maximum	25 th Percentile	75 th Percentile
<i>AREA (km²)</i>	148.70	1.63	4586.97	68.00	327.81
<i>BFIHOST</i>	0.47	0.20	0.97	0.40	0.57
<i>FARL</i>	0.99	0.65	1.00	0.96	1.00
<i>SAAR (mm)</i>	1096	558	2848	830	1375
<i>QMED</i>	43.54	0.14	992.85	12.92	117.71

861

862

863

864 Table 2. Network weights and biases. Input neurons I1 - I4 (AREA, BFIHOST, FARL, SAAR,

865 respectively); Hidden neurons H1 – H9; Output neuron O (QMED)

866 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

867

868 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

869

870 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

871

873
874
875

Table 3. Correlation matrix for model variables

876

	<i>AREA</i>	<i>BFIHOST</i>	<i>FARL</i>	<i>SAAR</i>	<i>QMED</i>
<i>AREA</i>	1.00	-0.02	-0.07	-0.05	0.76
<i>BFIHOST</i>		1.00	0.11	-0.42	-0.27
<i>FARL</i>			1.00	-0.39	-0.07
<i>SAAR</i>				1.00	0.24

877

878

879

Table 4. Numerical accuracy of different models under test

880

881

	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

882

883

884

885

Table 5. Physical legitimacy of GLM and ANN models

886

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
<i>AREA</i>	Low	} Yes	No	No
	Median		Yes	No
	High		No	No
<i>SAAR</i>	Low	} No	Yes	No
	Median		In Part	No
	High		No	No
<i>FARL</i>	Low	} Yes	In Part	No
	Median		In Part	No
	High		No	No
<i>BFIHOST</i>	Low	} No	No	No
	Median		In Part	No
	High		In Part	In Part

887

888

889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912

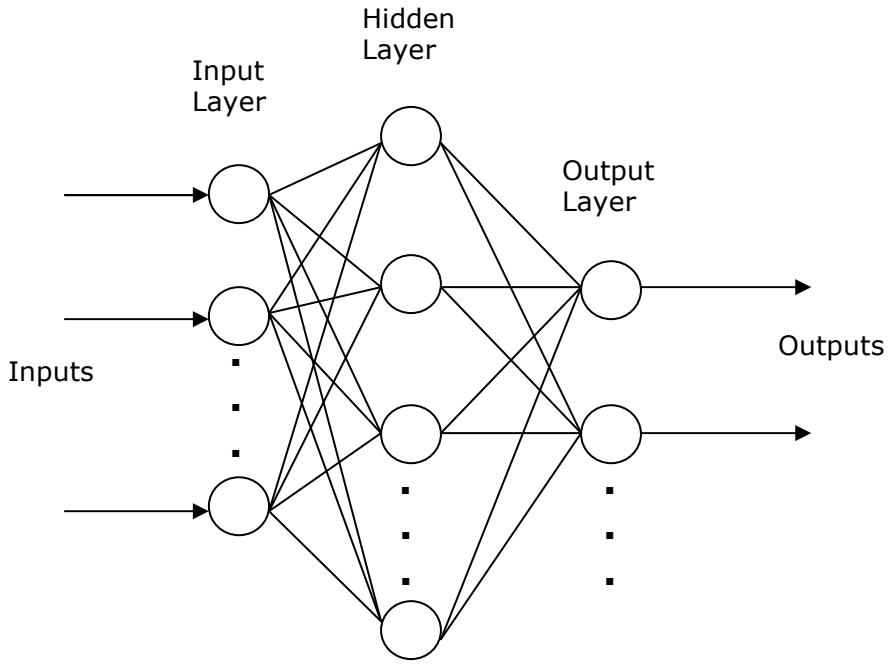
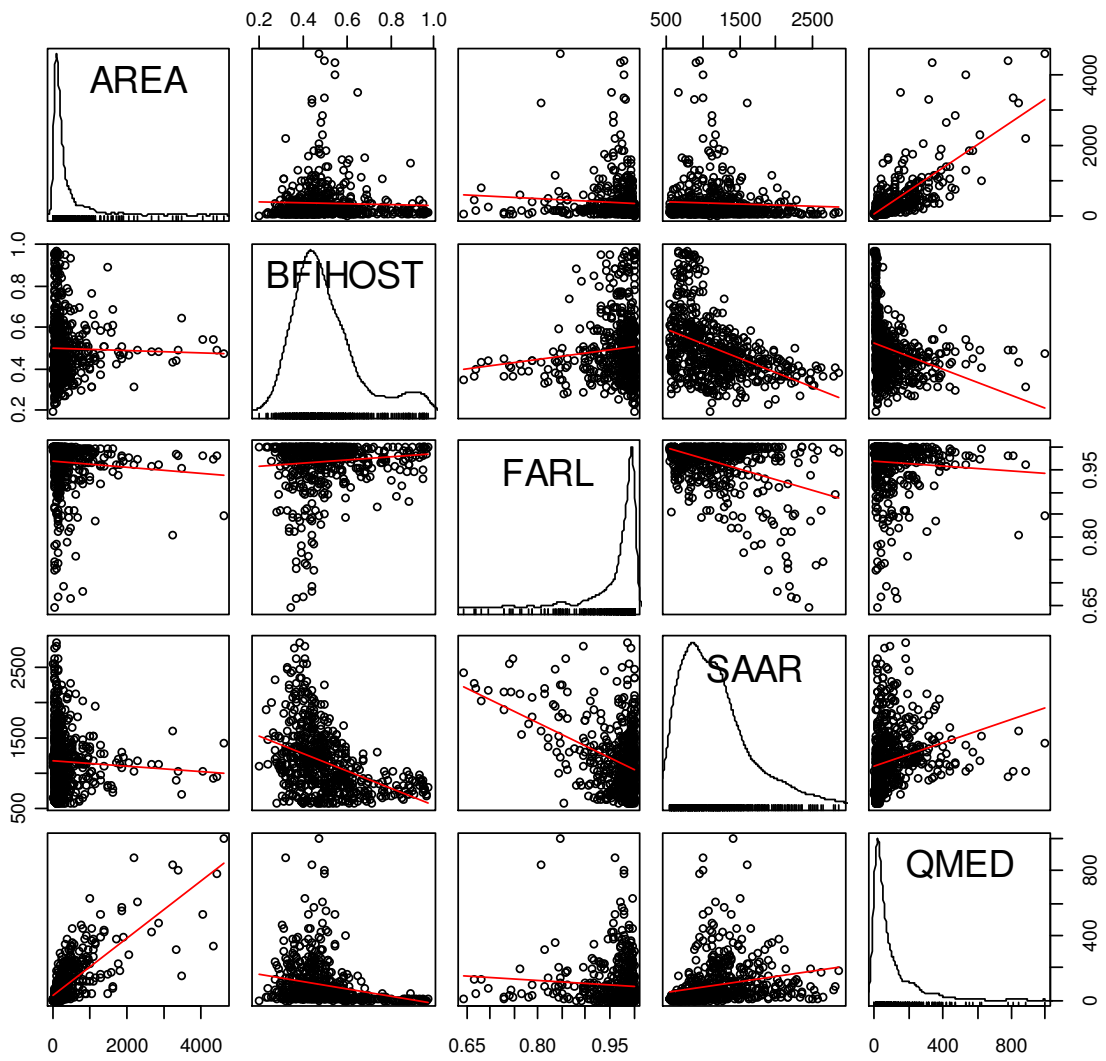


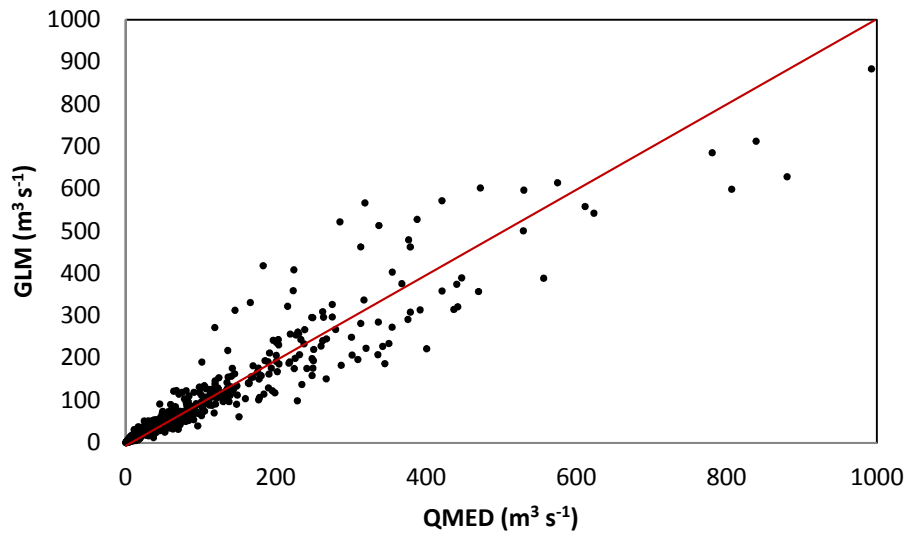
Figure 1. Typical feed forward ANN structure



913
 914
 915
 916

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

917
918
919
920
921

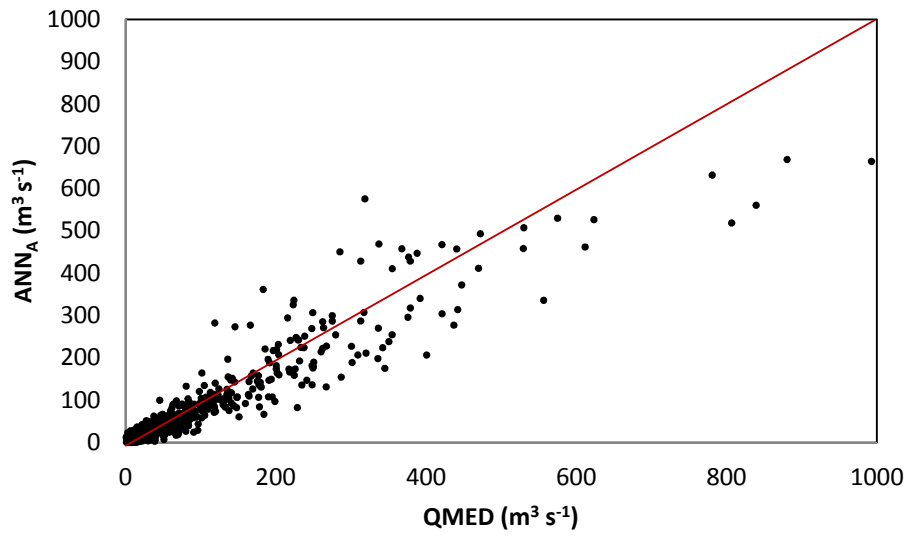


922
923
924
925

Figure 3. GLM versus *QMED*

926

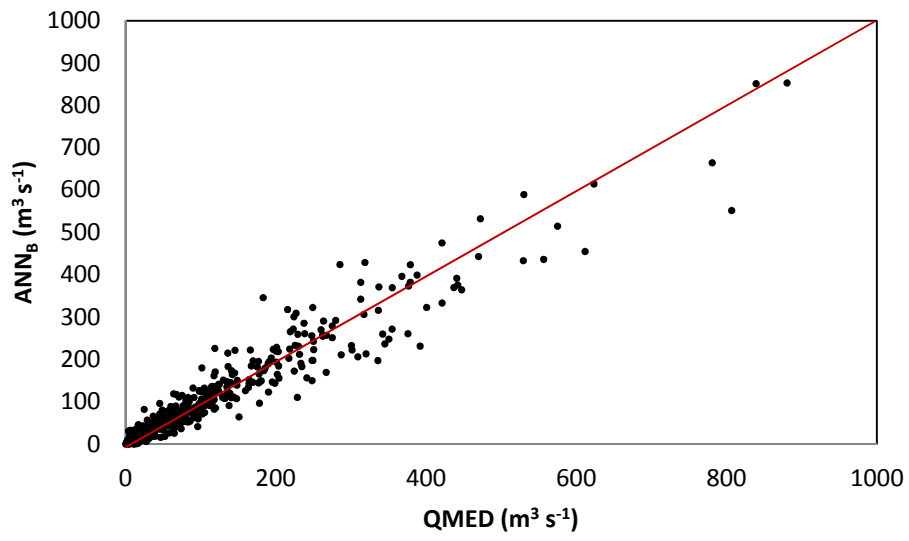
927
928
929



930
931
932
933

Figure 4. ANN_A model versus QMED

934
935



936
937
938
939

Figure 5. ANN_B model versus *QMED*

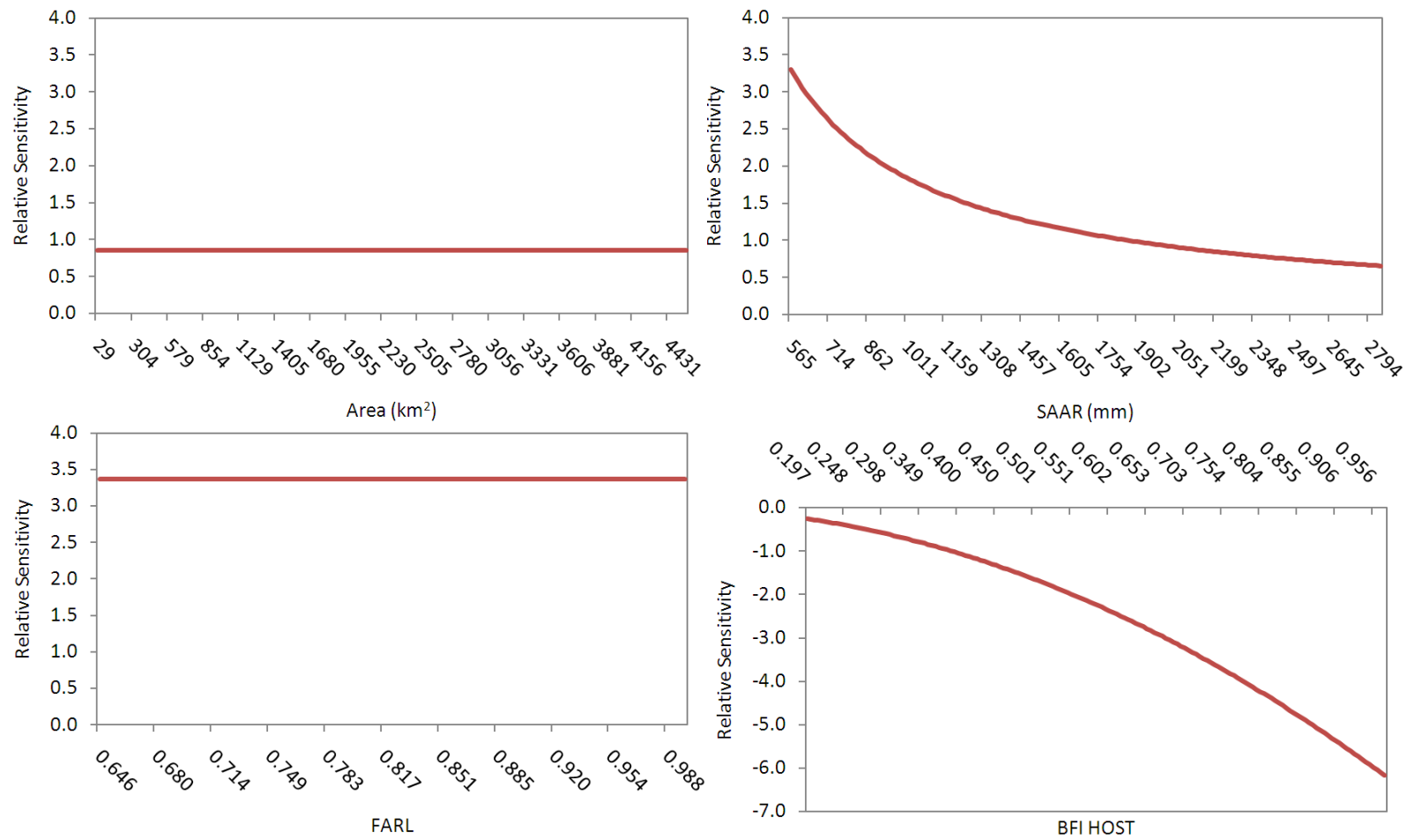


Figure 6. Relative sensitivity of *QMED* to model inputs: GLM

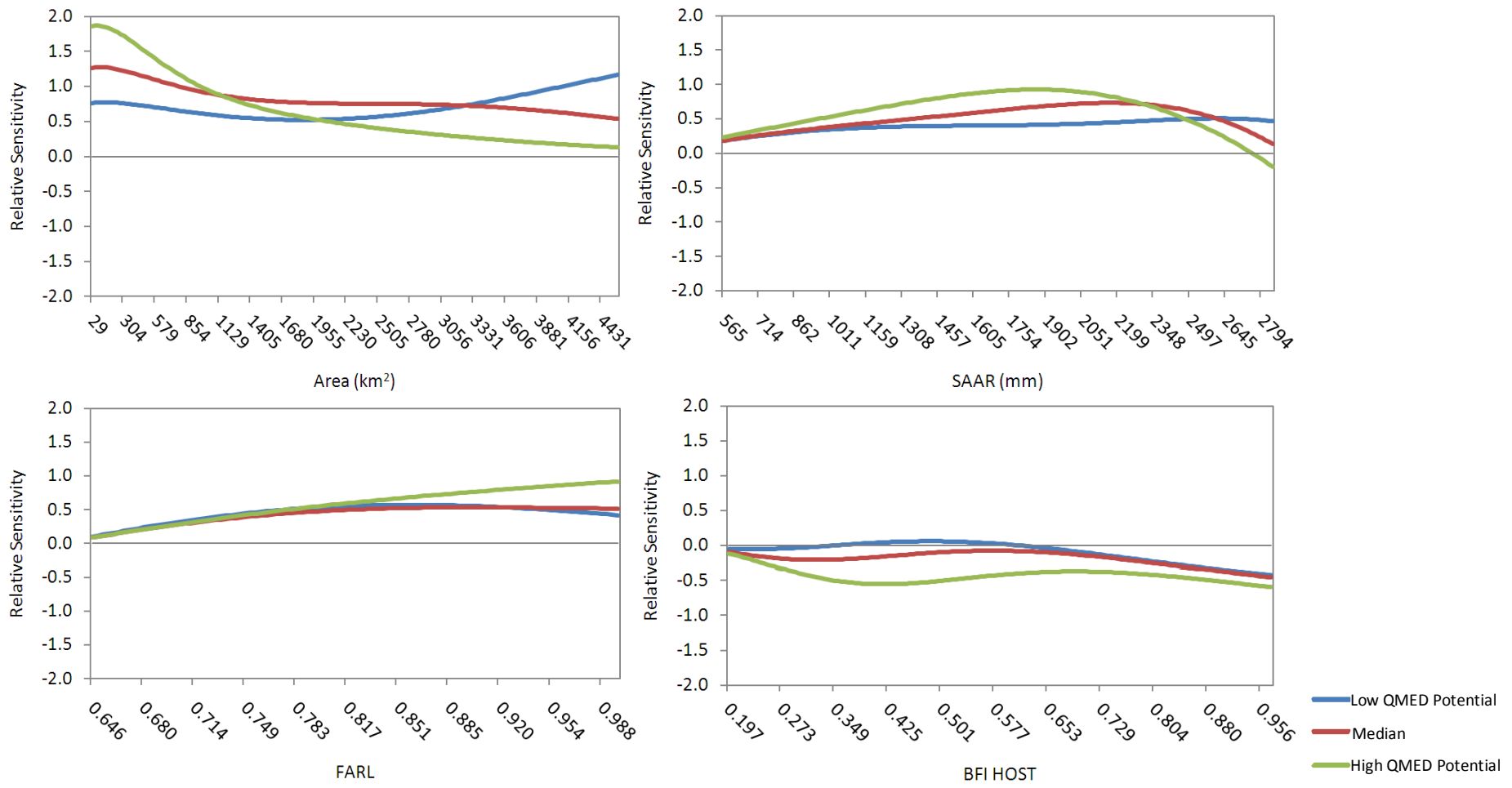


Figure 7. Relative sensitivity of *QMED* to model inputs: ANN_A

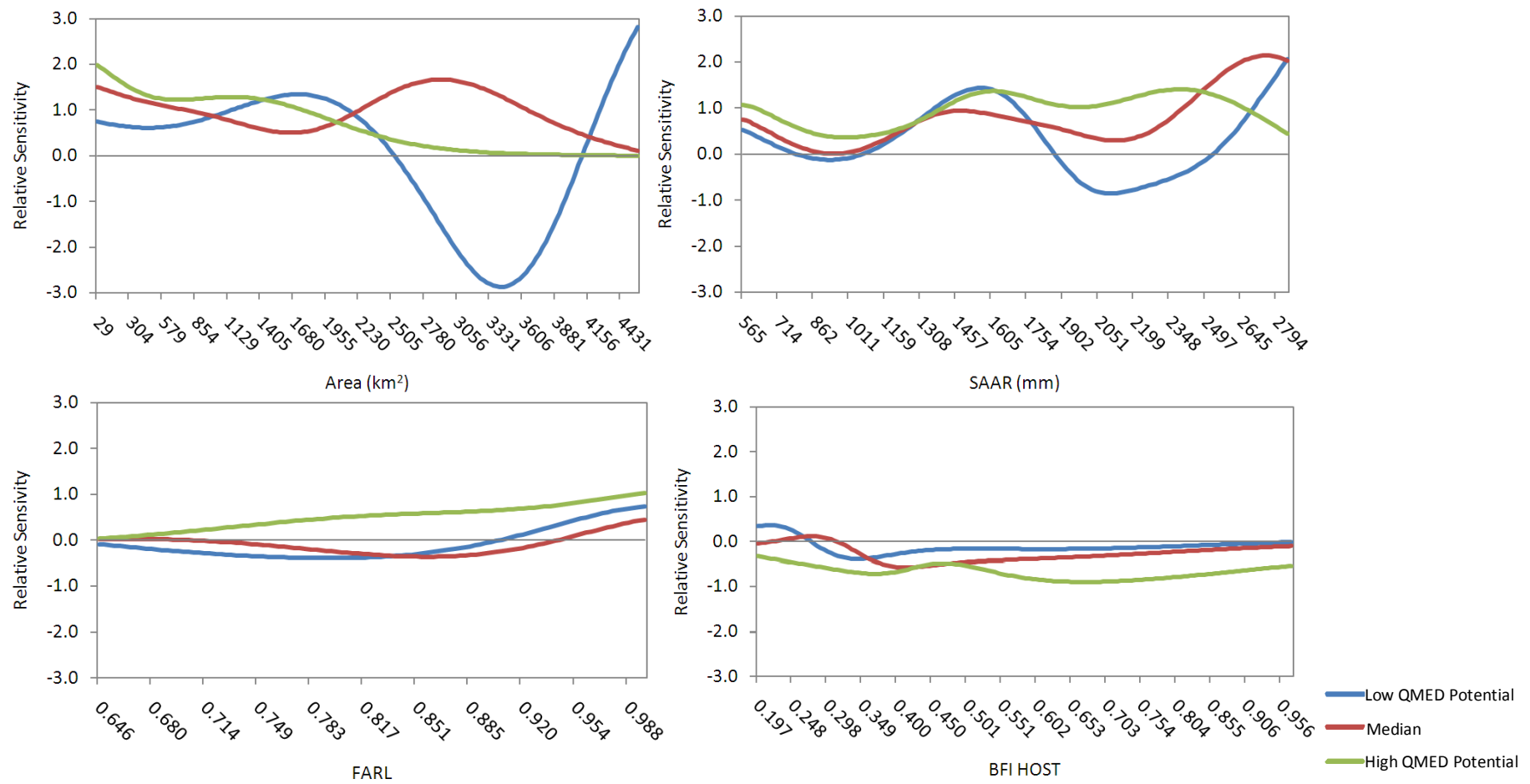


Figure 8. Relative sensitivity of *QMED* to model inputs: ANN_B