INVITED PAPER

Beyond validation: assessing the legitimacy of artificial neural network models

G.B. Humphrey [a], H.R. Maier [a], W. Wu [b], N.J. Mount [c], G.C. Dandy [a] and C.W. Dawson [d]

[a] School of Civil, Environmental & Mining Engineering, University of Adelaide, Adelaide, South Australia, [b] Department of Infrastructure Engineering, The University of Melbourne, Melbourne, VIC, [c] School of Geography, University of Nottingham, Nottingham, UK, [d]Department of Computer Science, Loughborough University, Loughborough, UK
Email: wenyan.wu@unimelb.edu.au

Abstract: Artificial neural network models have been used extensively for prediction and forecasting over the last 25 years. As the data used to develop ANNs contain important information about the physical processes being modelled, it is generally implied that a model that has been calibrated (trained) and performs well on an independent set of validation data represents the underlying physical processes of the system being modelled. However, this is not necessarily the case, most likely due to problems with equifinality, where different combinations of model parameters (e.g. connection weights) result in similar predictive performance. Consequently, there is also a need to check the behaviour of calibrated ANN models as part of the validation process, which is commonly referred to as structural, conceptual or scientific validation (Figure 1). This checks whether the input-output relationship captured by the model is plausible in accordance with *a priori* system understanding.

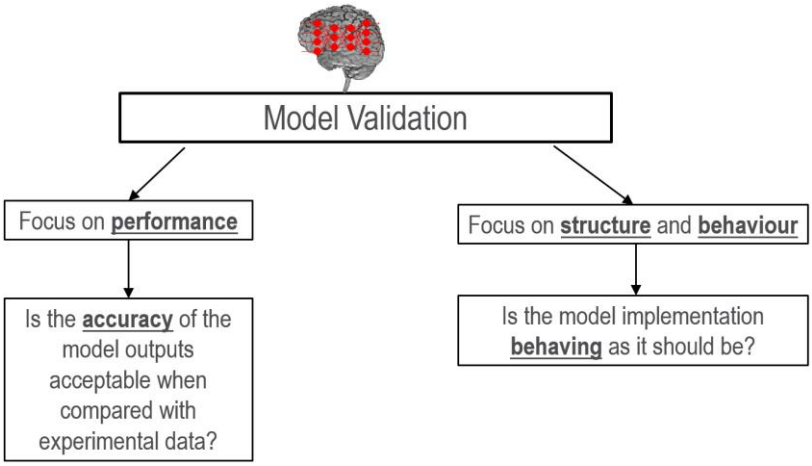

Figure 1. Importance of checking both predictive accuracy and model behaviour during ANN model validation processes

In this paper, the importance of considering structural validation is demonstrated. This is achieved by developing ANN models with different numbers of hidden nodes for two environmental modelling case studies from the literature namely, salinity forecasting in the River Murray in South Australia and the prediction of treated water turbidity at a water treatment plant based on raw water quality and the administered alum dose. The validation errors are then compared with corresponding model behaviours. This was done using the validann R-package, which caters to a range of structural validation approaches. Results show that ANN models producing the best fit to the data do not necessarily result in models that behave in accordance with underlying system understanding.

Keywords: *Artificial neural networks, multilayer perceptron, structural validation, process understanding, validann*

1. INTRODUCTION

Artificial neural network (ANN) models have been used extensively for prediction and forecasting for a range of water resources and environmental applications over the last few decades (e.g. Maier et al. 2010, Abrahart et al. 2012). In general, it is assumed that models that perform well on an independent validation data set can be used with confidence in practice (Wu et al. 2014). However, this is not the case for most other types of models, for which the ability of the model to behave in a way that is commensurate with underlying system understanding is considered an important part of the model validation process. This is referred to as structural (Power 1993), conceptual (Rykiel Jr 1996), or scientific validation (Biondi et al. 2012).

The fact that structural validation is generally ignored for ANN models can be particularly problematic, given that these types of models are not based on any underlying physics and can be prone to overfitting, given their potentially large number of model parameters (e.g. connection weights). Consequently, the objective of this paper is to assess the structural validity of ANN models with different structures for two environmental modelling case studies from the literature to test whether models with low validation errors are also structurally valid and whether the structural validity of models with similar validation errors is the same. This is achieved using the validann R-package (Humphrey et al. 2017), which caters to a number of approaches for checking the structural validity of ANN models. This paper presents an abridged overview of key methods discussed in Humphrey et al. (2017) to raise awareness of these techniques for a general audience.

The remainder of this paper is organised as follows. The methodology is introduced in Section 2, including the case studies and methods considered for checking the structural validity of ANN models. This is followed by results and discussion in Section 3 and summary and conclusions in Section 4.

2. METHODOLOGY

2.1. Case studies and ANN model development

Two case studies are considered, namely: (a) forecasting of salinity in the River Murray in Australia, and (b) prediction of the treated water quality at a water treatment plant, given a particular alum dose and the quality of water to be treated. Details of the case studies and model development process are given in Humphrey et al. (2017) and are summarized below.

The River Murray salinity (RMS) dataset has been studied extensively in the context of ANN development, where the aim has generally been to forecast salinity concentrations in the River Murray at Murray Bridge, South Australia, 14 days in advance. In line with previous studies, salinity at two upstream locations at lag 1 (Waikerie (WASt-1), Mannum (MASt-1)) and flow at one upstream location at lag 1 (Lock 7 (L7Ft-1)) were used as inputs for forecasting Murray Bridge salinity 14 days in advance (MBSt+13). Daily data between December 1986 and June 1992 were used for training and data from July 1992 to April 1998 for independent validation.

The southern Australian turbidity (SAT) dataset has been used in a number of previous ANN studies. In line with these studies, the inputs used for predicting treated water turbidity (TwTurbidity) were raw water turbidity (RwTurbidity), raw water pH (RwPh), raw water colour (RwColour), raw water ultraviolet absorbance at a wavelength of 254 nm (RwUvAbs254) and alum dose. Of the available data, 162 data samples (80%) were used for training and the remaining 40 samples (20%) were reserved for validation of the models.

For each case study, a number of multilayer perceptron (MLP) models were developed. Fifteen different ANN structures were considered, with the number of hidden nodes increasing from 1 to 15. Additionally, for each of the 15 network structures, the connection and bias weights were initialized five times with different random starting values uniformly distributed between -0.1 and 0.1, resulting in a total of 75 ANN models being developed for each case study. All ANNs were single hidden layer networks with hyperbolic tangent (tanh) hidden layer activations and a linear activation at the output.

All input data were standardised to have a mean of zero and standard deviation of one, while the target data were linearly rescaled between 0 and 1. The models were built in R (3.2.2) using the validann package (Humphrey et al., 2017), with the default BFGS optimisation algorithm used for training. All models were trained without cross-validation or early stopping for a maximum of 500 iterations using the default sum of square residuals as an objective function. Three of the best performing models, in terms of predictive validity, were selected for each case study and used to compare and contrast the corresponding validation errors and structural validity.

2.2. Assessment of structural validity

The purpose of structural validation is to check whether the input-output relationship captured by the trained models is plausible in accordance with *a priori* system understanding. While this approach does not determine whether the correct underlying relationship has been identified, it is helpful for identifying models that are not plausible from a physical perspective. Structural validation methods generally fall into two categories, including those that use values of the connection weights (parameters) of calibrated models directly to calculate the relative contribution of different inputs to the output and those that use sensitivity analysis (SA) approaches that examine the change in the model output as a result of input variation. In this study, two approaches from each category are considered, including Garson's algorithm and the modified connection weight (MCW) approach from the former category, and the profile SA method and the PaD method from the latter catgory.

Garson's algorithm (Garson 1991) is one of the earliest methods proposed for quantifying the relative importance (RI) of ANN inputs based on values of the connection weights of trained models. Using this method, input RI is calculated by partitioning the hidden-output layer connection weights into components associated with each input node using absolute values of the connection weights. Since absolute values of the weights are used, it is only possible to estimate the magnitude but not the direction of the input contributions (i.e. whether an input has a positive or negative effect on the output).

The MCW approach (Kingston et al. 2006a, 2010) is a modified version of the CW approach of Olden and Jackson (2002). As part of the latter, RI is computed based on an 'overall connection weight' between each input and the output, which in turn, is based on products of input-hidden and hidden-output connection weights for each input summed across all hidden nodes. In this approach, raw, rather than absolute, values of the weights are used, making it possible to estimate both the magnitude and direction of the input contributions. As part of the MCW approach, the raw input-hidden node weights are "squashed" using the hidden layer activation functions. This method is therefore only suitable for use with hidden layer activation functions that are symmetric about the origin, such as the hyperbolic tangent (tanh) function.

The Profile SA method, first described in Lek et al. (1995, 1996), involves successively varying each input variable across its range while keeping all others constant at their minimum, first quartile, median, third quartile, and maximum values; thus, producing five output profiles displaying variation in the output over the range of the input variable of interest. The median predicted responses across the five output profiles is also calculated, from which it is possible to assess the median behaviour of the model, given a range of different input values. In addition, the RI of each input is calculated based on the magnitude of the range of median output values produced by varying each input.

The PaD method (Dimopoulos et al. 1995, 1999) is another type of SA approach that involves computing partial derivatives of the model output with respect to each input variable in order to define the local rate of change of the output with respect to the corresponding input, while holding all other inputs fixed. Similar to the Profile method, this approach returns a profile of partial derivatives for each ANN input, which can be interpreted in a similar way to the coefficients in linear models, as well as a measure of input RI for each input. The sensitivity (partial derivatives) profiles enable model behaviour to be interpreted with respect to process rationality.

3. RESULTS AND DISCUSSION

3.1. Case study 1 – salinity forecasting

The validation errors for the three selected ANN models with different numbers of hidden nodes are shown in Table 1. As can be seen, the validation errors for these models are very similar based on all three error metrics considered. However, overall, Model 1, which has 13 hidden nodes and 66 model parameters, performs slightly better than the other two models and would most likely be selected as the preferred model based on validation error alone.

Table 1. Structure and performance of three selected ANN models developed for the salinity case study

	Model 1	**Model 2**	**Model 3**
Number of hidden nodes	13	5	3
Number of model parameters (weights)	66	26	16
Root Mean Squared Error (RMSE)	66.7	67.1	67.6
R^2	0.929	0.935	0.937
Coefficient of Efficiency (CE)	0.915	0.914	0.913

For this case study, it was known from previous work (Fernando et al. 2009) that WASt-1 is the most important variable for predicting MBSt+13, followed by MASt-1 and FL7t-1. The RI values obtained using Garson's method and the MCW method (Figure 2) indicate that only Model 2, which has 5 hidden nodes and 26 model parameters, is structurally valid. The directions of the relationships between inputs and outputs for this model are also correct, as it is known that the two upstream salinity variables should have a positive relationship with downstream salinity, whereas the opposite is the case for the flow input. The superior structural validity of Model 2 can also be seen in the comparison of the results of the PaD method for Models 1 and 2, as the relative sensitivities of the three input variables are much better aligned with this known behaviour of the system for this model (Figure 3).

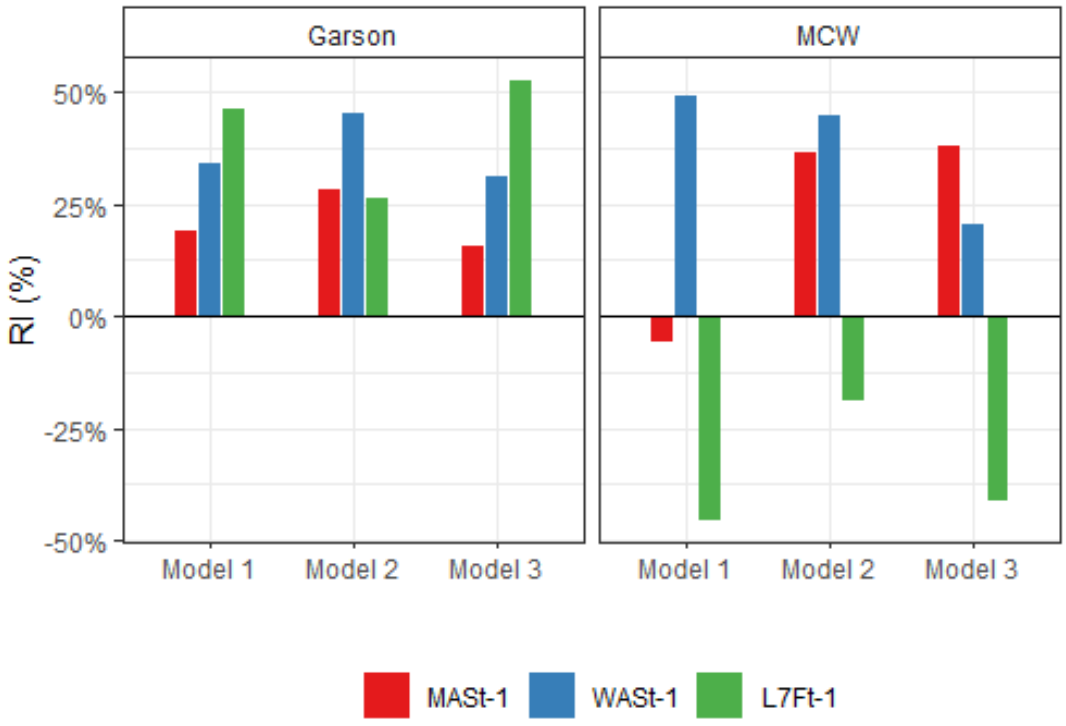

Figure 2. Structural validation of ANN models for the salinity case study using methods based on connection weight analysis

3.2. Case study 2 – treated water turbidity prediction

The validation errors for the three selected ANN models with different numbers of hidden nodes are shown in Table 2. As can be seen, the validation errors for these models are generally similar for all three error metrics considered. However, overall, Model 1, which has 14 hidden nodes and 99 model parameters, performs slightly better than the other two models and would most likely be selected as the preferred model based on validation error alone. In contrast, Model 3, which has one hidden node and eight model parameters, performs the worst on the validation data.

Table 2. Structure and performance of three selected ANN models developed for the turbidity case study

	Model 1	**Model 2**	**Model 3**
Number of hidden nodes	14	12	1
Number of model parameters (weights)	99	85	8
RMSE	0.29	0.31	0.32
R^2	0.81	0.80	0.78
CE	0.81	0.78	0.76

When assessing structural validity using the Profile method, a plausible model would be one that produces outputs roughly within the range of the observed data (TwTurbidity between approximately 0-6 NTU) and displays a reasonably monotonic relationship between each of the explanatory variables and TwTurbidity when all other explanatory variables are fixed. In addition, it would generally be expected that as the turbidity of the raw water (RwTurbidity) increases, the resulting turbidity of the treated water (TwTurbidity) would also increase for fixed values of all other explanatory variables. Likewise, the higher the UVA-254 of the raw water (RwUvAbs254), the higher the TwTurbidity would be expected to be, since UVA-254 is used as a surrogate for dissolved natural organic matter (NOM) concentration, which negatively impacts turbidity removal (alum reacts preferentially with dissolved NOM) (White et al. 1997).

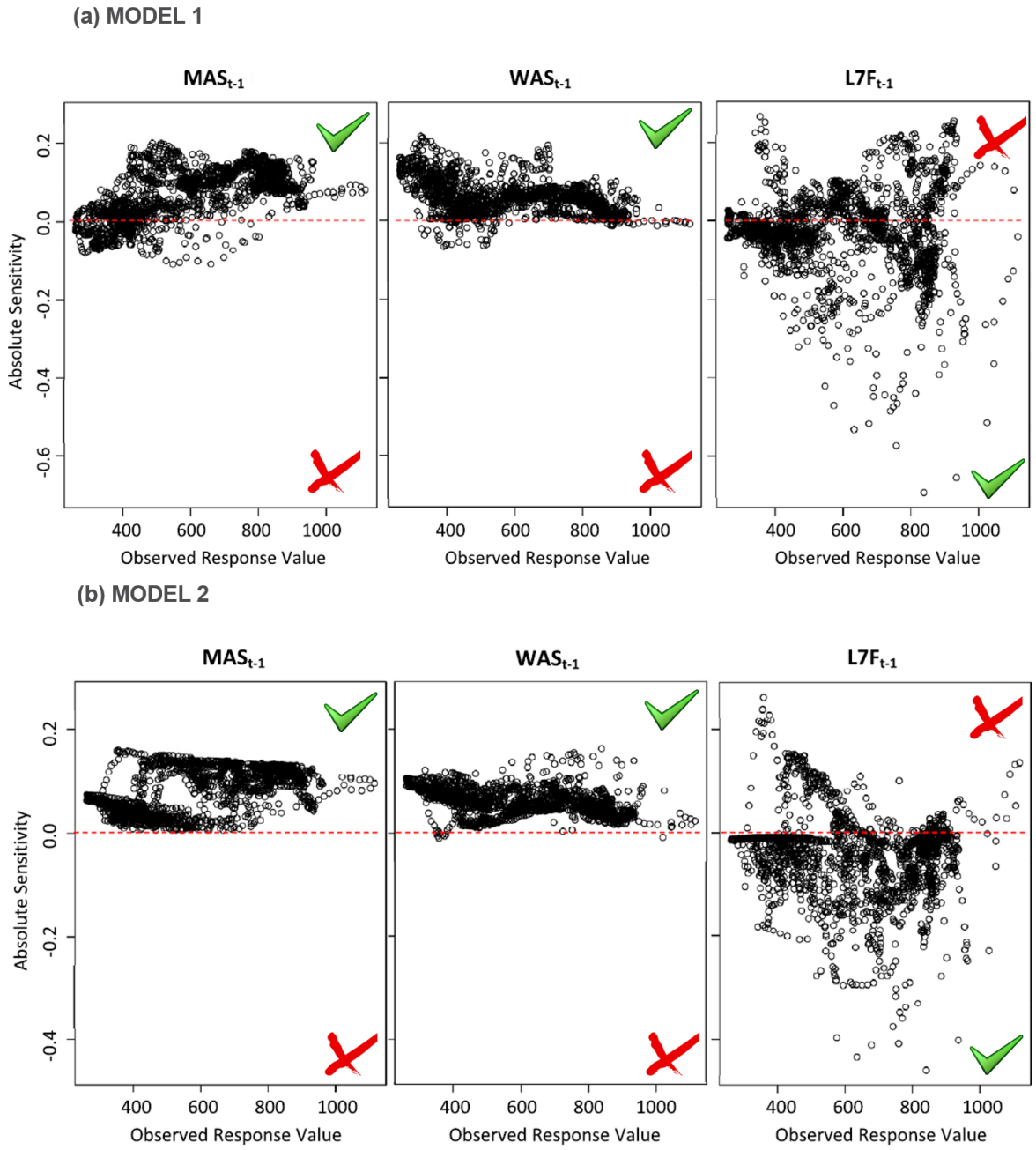

Figure 3. Structural validation of ANN models for the salinity case study using the PaD sensitivity analysis method, testing that sensitivity is positive (or negative) as expected

Based on the above understanding of system behaviour, the results of the Profile method indicate that Model 3 can be considered physically plausible, while this is not the case for Model 1. For example, the input-output relationships shown in Figure 4 (b) generated using Model 3 are in line with physical understanding; as expected, predicted TwTurbidity increases with increasing RwTurbidity and reduces with increasing alum dose. On the other hand, using Model 1, the model response to variation in these inputs is not only contradictory to the expected behaviour, as seen in Figure 4 (a), but negative values of TwTurbidity are produced for certain input values. For the remaining inputs not shown in Figure 4, the predicted TwTurbidity ranged between approximately 0-6.5 using Model 3, which is a plausible range for this variable given the ranges of the input variables considered. Consequently, Model 3 appears to be the most structurally valid. The threshold behaviour observed for Model 3 when increasing alum dose and fixing all other inputs at their maximum values is as would be expected, as it was observed by White et al. (1997) that a threshold alum dose is often required before a sharp reduction in turbidity is achieved.

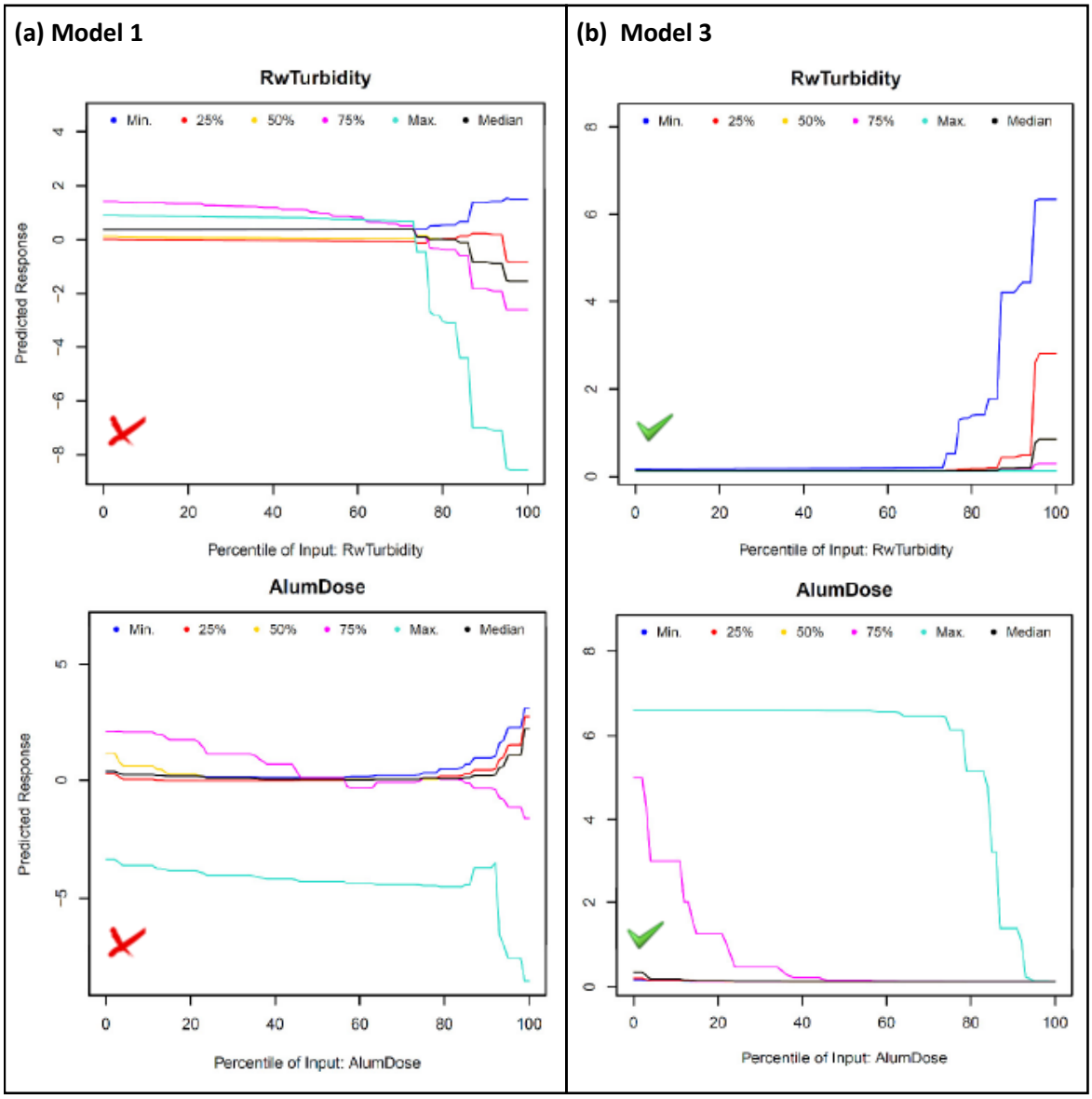

Figure 4. Structural validation of ANN models for the turbidity case study using the Profile sensitivity analysis method

4. SUMMARY AND CONCLUSIONS

Structural validity is generally not considered when developing ANN models (see e.g. Wu et al., 2014). Instead, ANN models are generally considered to be suitable for use when they perform adequately on an independent validation dataset. However, this can result in models that do not mimic the underlying physical processes being modelled and hence perform poorly in an operational setting. This was demonstrated in this paper for two environmental modelling case studies, where 75 ANN models with different numbers of hidden nodes and parameter (weight) initialisations were developed for each. While some models result in very similar performance in terms of validation error, the underlying relationships they had captured from the data were very different. This highlights the need to move beyond the sole use of validation errors to assess the validity of ANN models and to embrace structural validation as a standard component of the ANN model validation process.

ACKNOWLEDGMENTS

Wenyan Wu acknowledges support from ARC Discovery Early Career Researcher Award (DE210100117).

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., 2012. Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting. Progress in Physical Geography 36, 480-513. doi:10.1177/0309133312444943.

Biondi, D., Freni, G., Iacobellis, V., Mascaro, G., Montanari, A., 2012. Validation of hydrological models: Conceptual basis, methodological approaches and a proposal for a code of practice. Physics and Chemistry of the Earth, Parts A/B/C 4244, 70-76. doi:10.1016/j.pce.2011.07.037.

Dimopoulos, I., Chronopoulos, J., Chronopoulou-Sereli, A., Lek, S., 1999. Neural network models to study relationships between lead concentration in grasses and permanent urban descriptors in Athens city (Greece). Ecological Modelling 120, 157-165. doi:10.1016/S0304-3800(99)00099-X.

Dimopoulos, Y., Bourret, P., Lek, S., 1995. Use of some sensitivity criteria for choosing networks with good generalization ability. Neural Processing Letters 2, 1-4. doi:10.1007/bf02309007.

Fernando, T.M.K.G., Maier, H.R., Dandy, G.C., 2009. Selection of input variables for data driven models: An average shifted histogram partial mutual information estimator approach. Journal of Hydrology 367, 165-176. doi:10.1016/j.jhydrol.2008.10.019.

Garson, G.D., 1991. Interpreting neural-network connection weights. AI Expert 6, 46-51.

Humphrey G.B., Maier H.R., Wu W., Mount N.J., Dandy G.C., Abrahart R.J. and Dawson C.W., 2017. Improved validation framework and R-package for artificial neural network models, Environmental Modelling and Software, 92, 82-106, DOI: 10.1016/j.envsoft.2017.01.023.

Kingston, G.B., Maier, H.R., Lambert, M.F., 2006a. Forecasting cyanobacteria with Bayesian and deterministic artificial neural networks, in: IJCNN '06. International Joint Conference on Neural Networks, 2006. IEEE. pp.1304 4870-4877, doi:10.1109/ijcnn.2006.247166.

Kingston, G., Maier, H., Lambert, M., 2010. Bayesian Artificial Neural Networks: with Applications in Water Resources Engineering. VDM Verlag.

Lek, S., Belaud, A., Dimopoulos, I., Lauga, J., Moreau, J., 1995. Improved estimation, using neural networks, of the food consumption of fish populations. Marine and Freshwater Research 46, 1229-1236. doi:10.1071/MF9951229.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., Aulagnier, S., 1996. Application of neural networks to modelling nonlinear relationships in ecology. Ecological Modelling 90, 39-52. doi:10.1016/0304-3800(95)00142-5.

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. Environmental Modelling & Software 25, 891-909. doi:10.1016/j.envsoft.2010.02.003.

Olden, J.D., Jackson, D.A., 2002. Illuminating the \black box": a randomization approach for understanding variable contributions in artificial neural networks. Ecological Modelling 154, 135-150. doi:10.1016/S0304-3800(02)00064-9.

Power, M., 1993. The predictive validation of ecological and environmental models. Ecological Modelling 68, 33-50. doi:10.1016/0304-3800(93)90106-3.

Rykiel Jr, E.J., 1996. Testing ecological models: the meaning of validation. Ecological Modelling 90, 229-244. doi:10.1016/0304-3800(95)00152-2.

White, M.C., Thompson, J.D., Harrington, G.W., Singer, P.C., 1997. Evaluating criteria for enhanced coagulation compliance. Journal AWWA 89, 64-77.