22nd International Congress on Modelling and Simulation, Hobart, Tasmania, Australia, 3 to 8 December 2017
mssanz.org.au/modsim2017

Sensitivity analysis of constituent generation parameters of an integrated hydrological and water quality model using a GMDH polynomial neural network

F.R. Bennett [a], **B. Fentie** [a]

[a] *Queensland Department of Science, Information Technology and Innovation*
Email: frederick.bennett@dsiti.qld.gov.au

Abstract: Catchment water quality models are notoriously over-parametrised. Given this condition, it is useful to be able to identify which parameters have the greatest influence on the model results. In theory, this could be accomplished through a detailed first principles interrogation of the mathematical structure of the model in an abstract manner. This however, is impractical in most instances owing to the complexity of the models and *posteriori* methods of parameter sensitivity analysis are more conventional.

As with most aspects of large-scale modelling endeavours, a major consideration in choosing a technique for sensitivity analysis is efficiency and a compromise between computational effort and numerical accuracy is usually negotiated. ANOVA based sensitivity analysis methods are very popular as they offer a holistic survey of the parameter sensitivity by not only accounting for the response of the model output surface due to the activity of single parameters acting independently, but also due to the interaction between parameters. These global sensitivity indices are usually calculated by Monte Carlo simulation and may be too computationally demanding to be routinely applied in water quality modelling scenarios.

We demonstrate the application of the group method of data handling (GMDH) inductive, self-organising modelling method to the sensitivity analysis of constituent generation parameters of an integrated hydrological and water quality model. By using a modestly sized sample input-output dataset, a GMDH neural network is used to synthesise a sparse, random-sampling high dimensional model representation (RS-HDMR) that can be used to calculate first and second order Sobol sensitivity indices. This algorithm potentially leads to reductions in computational cost of 2-3 orders of magnitude over Monto Carlo simulation.

Although several other adaptive methods for efficiently constructing a sparse RS-HDMR have been reported in the literature, such as polynomial chaos expansions, the parameter selection and noise filtering characteristics of the GMDH network may result in more optimal HDMR expansion.

Keywords: *Sensitivity analysis, neural networks, water quality model*

1. INTRODUCTION

Reduced to their most fundamental form, water quality models are a set of equations that describe the processes that lead to the generation and transport of various constituents that contribute to instream loads or concentrations (James 1993). As these models increase in complexity to account for a broader range of landscape processes, the number of parameters informing the model can also increase substantially. It would stand to reason that an insight into which parameters contribute most significantly to variation in the model output is not only useful for an improved qualitative understanding of the behaviour of the model, it is also valuable for designing a model calibration strategy (Nossent, Elsen, and Bauwens 2011).

Sensitivity analysis provides a method for identifying a set of parameters that influences the output of a model or process in general and is regularly applied to environmental models (Pianosi et al. 2016), hydrological models (Song et al. 2015) and water quality models (Liu and Zou 2012; Khorashadi Zadeh et al. 2017; Manache and Melching 2004; Sincock, Wheater, and Whitehead 2003).

A number of approaches are available for studying how the sensitivity of a model output can be apportioned to various input parameters (Iooss and Lemaître 2015; Pianosi et al. 2016). Among the most ubiquitous of these are those that are structured around the functional decomposition of variance, particularly Sobol's method (Sobol 2001). The aim of variance based methods is to determine the magnitude of the variance of the model output attributable to the variance of each model parameter. In the Sobol analysis, variance associated with a single parameter or due to interactions between parameters is expressed as a Sobol sensitivity indices which represent fractions of unconditional variance of the model output.

The objective of this paper is to explore the application of a novel and efficient approach to the Sobol sensitivity analysis of a set of water quality model parameters effecting the generation and transport of fine sediment in a distributed stream network.

2. BACKGROUND

2.1. Water Quality Model

A Source Catchments model (Delgado et al. 2012; Kelley and O'Brien 2012) for the Burnett-Mary NRM region in Queensland which has been described in detail elsewhere (McCloskey et al. 2017) has been used for the current work. The Source Catchments model employs a set of fit for purpose component models for estimating the generation of constituents from various land use areas across the landscape. This current study is concerned with fine sediment generation and transport associated with the following processes:

- hillslope erosion occurring on grazing, forestry and conservation land use areas;
- streambank erosion;
- fine sediment settling and remobilisation in the channel; and
- deposition of fine sediment on the floodplain.

All of the listed processes are managed through a purpose built Dynamic SedNet plugin (Ellis and Searle 2014) which implements a daily time step Revised Soil Loss Equation (RUSLE) (Renard et al. 1991) for modelling hillslope erosion. Dynamic streambank erosion, channel and floodplain processes are discussed in Ellis and Searle and won't be further elaborated on here.

2.2. GMDH

The GMDH algorithm was first introduced by 1971 when Ivakhnenko described an inductive, deep learning method that would model the input-output relationship of a complex system using a multilayered perceptron-type network structure (Ivakhnenko 1971). The objective of the GMDH algorithm is the construction of a high-order Kolmogorov-Gabor polynomial of the form

$$Y(x_1,...,x_M) = a_0 + \sum_{i=1}^{M} a_i x_i + \sum_{i=1}^{M}\sum_{j=1}^{M} a_{ij} x_i x_j + \sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{M} a_{ijk} x_i x_j x_k + \cdots \quad (1)$$

which connects a vector of input variables $X(x_1, x_1, \cdots x_M)$ to the output variable Y by determining $A(a_1, a_2, \cdots a_M)$, the vector of summand coefficients (Farlow 1981).

The basic approach of GMDH is that each neuron in the network receives input from exactly two other neurons with the exception of the neurons representing the input layer. The two inputs, x_i and x_j are then combined to produce a partial descriptor based on the simple quadratic transfer function

$$y = a + bx_i + cx_j + dx_i^2 + ex_j^2 + fx_i x_j \tag{2}$$

where the coefficients *a..f* are determined statistically using a least squares regression and are unique for each transfer function. The coefficients can be thought of as analogous to weights found in other types of neural networks. The network of transfer functions is constructed one layer at a time. The first network layer consists of functions of each possible pair of *n* input variables (zeroth layer) resulting in $n \cdot (n\text{-}1)/2$ neurons. The second layer is created using inputs from the first layer and so on. The first network layer therefore consists of a set of quadratic functions of the input variables, the second layer involves fourth degree polynomials, the third layer includes eighth degree polynomials *etc*. A selection process is employed to limit the size of the network by culling neurons at each layer based on a performance criterion. The way in which this is done represents an important feature of the GMDH algorithm. The efficient PESS (predicted error sum of squares) criterion is used to rank neuron performance.

$$PESS = \frac{1}{N}\sum_{t=1}^{N}\left(y_t - f\left(x_t, \hat{a}_t\right)\right)^2 \tag{3}$$

Where $\hat{a}_t$ is the estimation of unknown parameters on the complete data set from which the t^{th} observation has been excluded. PESS is an external criterion but does not require the explicit subdivision of observed data into training and validation sets since it employs cross validation techniques internally. The sum of all N validations provides a measure of how consistent a model is when applied to new data, and thus helps to avoid overfitting. The best performing neurons in each layer can be selected based on their resulting PESS. The GMDH algorithm automatically terminates once the performance of the network begins to deteriorate.

Unlike other types of neural network approaches to modelling data, GMDH provides a fully portable, symbolic description of the final network or model in the form of a polynomial function.

Being a fully inductive process, GMDH presents some very appealing features such as:

- Fully automatic structural and parametric optimisation of the network.
- Optimal complexity of the model structure is found, adequate to the level of noise in data sample. For real problems, with noisy or short data, simplified optimal models are more accurate.
- The number of layers and neurons in hidden layers, model structure and other optimal neural networks parameters are determined automatically.
- It automatically finds interpretable relationships in data and selects effective input variables accordingly.
- It guarantees that the most accurate or unbiased models will be found – method does not miss the best solution during sorting of all variants (in the given class of functions).

2.3. Sensitivity analysis

Lambert and co-workers have recently shown that the GMDH can be used to construct a high dimensional model representation (HDMR) that can be used to calculate first and second order Sobol sensitivity indices (Lambert et al. 2016).

Assume that we have a function $Y(\mathbf{X})$ that is square-integrable and defined within the unit hypercube $[0,1]^M$, it is possible to decompose $Y(\mathbf{X})$ into a sum of elementary functions (Hoeffding 1992):

$$Y(x_1,...,x_M) = f_0 + \sum_{i=1}^{M} f_i(x_i) + \sum_{1\le i\le j\le M}^{M} f_{ij}(x_i, x_j) + \cdots + f_{12\ldots M}(x_1, x_2, \cdots, x_M) \tag{4}$$

where f_0 is a constant expressing the zeroth order effect which is simply the mean of all outputs. The first order term, $f_i(x_i)$ represents the influence of the individual inputs x_i acting independently on the output $Y(\mathbf{X})$

. $f_{ij}(x_i x_j)$ are functions associated with the interactive influence on $Y(\mathbf{X})$ from x_i and x_i acting together. As the series progresses, higher order terms represent the correlation effects between larger ranges of input variables. This meta-representation of the original function $Y(\mathbf{X})$ is often referred to as a high dimensional model representation (HDMR) (Rabitz et al. 1999). Typically, HDMR expansions truncated at second order have been sufficient to describe many high dimensional systems. There are various methods for determining Sobol sensitivity indices from HDMR which won't be generally discussed in this paper, rather we will focus our attention on the random sampling HDMR (RS-HDMR) method. In order to reduce the overall sampling effort, RS-HDMR can employ different analytical basis functions, such as orthonormal polynomials, cubic B splines, and polynomials to approximate the RS-HDMR component functions. Only one set of random input−output samples is necessary to determine all the RS-HDMR component functions, and a few hundred samples may give a satisfactory approximation, regardless of the dimension of the input variable space (Li, Wang, and Rabitz 2002).

The RS-HDMR approach approximates the lower order component functions by an expansion of an orthonormal polynomial basis set as

$$f_i(x_i) \approx \sum_{r=1}^{k} \alpha_r^i \varphi_r(x_i)$$
$$f_{ij}(x_i, x_j) \approx \sum_{p=1}^{l}\sum_{q=1}^{m} \beta_{pq}^{ij} \varphi_p(x_i)\varphi_q(x_j) \tag{5}$$

where *k*, *l* and *m* are the predefined polynomial orders

A sparse representation of Eq. (4) can now be rewritten as (terminating at order 2)

$$Y(x_1,...,x_M) = c_0 + \sum_{i=1}^{M}\sum_{r=1}^{k} \alpha_r^i \varphi_r(x_i) + \sum_{1\le i\le j\le M}^{M}\sum_{p=1}^{l}\sum_{q=1}^{m} \beta_{pq}^{ij} \varphi_p(x_i)\varphi_q(x_j) \tag{6}$$

The decomposition coefficients in (6) can be used to calculate the partial variances as :

$$D_i = \sum_{r=1}^{k} (\alpha_r^i)^2 \tag{7}$$

$$D_{ij} = \sum_{p=1}^{l}\sum_{q=1}^{m} (\beta_{pq}^{ij})^2 \tag{8}$$

And the first and second order Sobol sensitivity indices can be calculated as follows:

$$S_i = \frac{D_i}{D} \tag{9}$$

$$S_{ij} = \frac{D_{ij}}{D} \tag{10}$$

where *D* is the total variance of $Y(\mathbf{X})$. Ideally, the sensitivity indices will sum to unity such that

$$\sum_i S_i + \sum_{i,j} S_{ij} = 1 \tag{11}$$

3. METHODOLOGY

This study is based on modelled fine sediment loads at the locations of the 136014A gauging station in the Burnett catchment. The contributing are for this drainage point is 33,020 km^2.

A set of model parameters have been chosen to represent each geophysical process influencing the generation and delivery of fine sediment to the stream. Most of these parameters are considered as variable in the model context and can be used for model calibration. Although the sediment dry bulk density parameter, ρ_s, is a well-defined geophysical property, it is proportional to the contribution of streambank erosion to the total delivered sediment load and is therefore a useful proxy representing this process in the sensitivity analysis. To help ensure the evenness of sampling over the range of the parameters given in Table 1, a 257 sample nearly orthogonal latin hypercube (Gu and Yang 2013; Cioppa and Lucas 2007; Kleijnen et al. 2005) was constructed to define the set of input vectors, **X**. No further model evaluations were required to calculate the Sobol sensitivities for these parameters.

Table 1. Model parameters and range of values used in sensitivity analysis

Parameter	Description	Range of values
HSDR	Hillslope sediment delivery ratio	0-15 (%)
GSDR	Gully sediment delivery ratio	0-15 (%)
ρ_s	Sediment dry bulk density	0-3 (t/m^3)
V_p	Floodplain deposition settling velocity	10^{-7}-10^{-4} (m/s)
ω_{mob}	Channel average terminal fall velocity for fine sediment remobilisation	0-1 (m/s)
ω_{dep}	Channel average terminal fall velocity for fine sediment deposition	10^{-6}-10^{-4} (m/s)

For the RS-HDMR, the variables x_i first need to be rescaled by some suitable transformation such that $0 \le x_i \le 1$. A set of synthetic parameters are constructed using orthogonal polynomial functions resulting in an expanded input basis $X_{p,i}$ where p refers to the polynomial order such that $X_{p,i} = \varphi_p(x_i)$ where $\varphi_p(x_i)$ is the shifted Legendre function $\tilde{P}_p(x_i)$ given by the terms:

$$\begin{aligned}
\tilde{P}_1(x) &= \sqrt{3}(2x-1) \\
\tilde{P}_2(x) &= 6\sqrt{5}\left(x^2 - x + \frac{1}{6}\right) \\
\tilde{P}_3(x) &= 20\sqrt{7}\left(x^3 - \frac{3}{2}x^2 + \frac{3}{5}x - \frac{1}{20}\right) \\
&\vdots \\
\tilde{P}_n(x) &= (-1)^n\sqrt{2n+1}\sum_{k=0}^{p}\binom{p}{k}\binom{p+k}{k}(-x)^k
\end{aligned} \tag{12}$$

or alternatively using a variation of Rodrigue's formula (Horner 1965)

$$\tilde{P}_n(x) = \frac{\sqrt{2n+1}}{n!}\frac{d^n}{dx^n}(x^2 - x)^n \tag{13}$$

These shifted Legendre polynomials obey the following orthonormal condition over the interval $0 \le x \le 1$

$$\int_0^1 \tilde{P}_m(x)\tilde{P}_n(x)dx = \delta_{mn} \tag{14}$$

The synthesised expanded input data set along with the output vector Y can now be used to estimate the optimal coefficients α_r^i and β_{pq}^{ij} using the GMDH alogorithm and deriving (6) from (1). Armed with these coefficients, partial variances and Sobol indices can be calculated according to equations (7) to (10).

4. DISCUSSION AND CONCLUSIONS

The sensitivity indices of the input parameters listed in Table 1 have been estimated using the GMDH RS-HDMR method with the results are provided in Table 2.

A 5 year period ranging from 1st July 2007 to 30th June 2012 has been considered for the sensitivity analysis. Within this timeframe, years 1, 2, 3 and 5 can be regarded as "typical" weather years with modelled sediment loads averaged over the parameter range deviating marginally from their respective average of around 9.33×10^6 kg/y. Year 4, which represents the year commencing 1st July 2010 is clearly an exceptional case with an average modelled fine sediment load of 241.5×10^6 kg/y which is a greater than 20 fold increase in comparison to the synoptic years. The large movement of sediment in Year 4 can be attributed to a single extraordinary flood event in the Burnett catchment occurring in late December 2010.

Table 2. Average annual sediment loads and sensitivity indices. Only indices ≥ 0.01 are included. Year 1 refers to the period 1st July 2007 – 31st June 2008.

	Time Period					
Value/Parameter	Year 1	Year 2	Year 3	Year 4	Year 5	Average
Average Load (10^6 kg/y)	9.8	5.4	11.6	241.5	10.5	55.7
S1 (HSDR)	0.62	0.74	0.63	0.07	0.69	0.14
S2 (GSDR)	0.03	0.02	0.06	0.02	0.03	0.02
S3 (ρ_s)	0.33	0.24	0.31	0.70	0.27	0.68
S4 (V_p)	0.00	0.00	0.00	0.02	0.00	0.03
S5 (ω_{mob})	0.02	0.00	0.00	0.12	0.01	0.11
S3,5	0.00	0.00	0.00	0.08	0.00	0.03

It can been seen in Table 2 that an effect of the December 2010 event is a significant increase in the S3 index and reduction in the S1 index and this distortion carries over to the 5 year average indices. Only weak second order effects coupling streambank erosion and channel remobilisation are reported, and this effect only becomes significant in Year 4 with an index of 8%.

The sensitivity analysis indicates that when aggregating the 5 year set of modelling results through averaging, a strong bias is introduced due to the unusual conditions in Year 4. These results, where the influence of a single event overwhelms the contributions from the remainder of the time series, would have important implications for parameter calibration. One strategy that may account for this in a balanced way would be to conduct a multi-objective optimisation which would allow for more flexibility in identifying the most suitable set of model parameters (Shafii and De Smedt 2009).

Further work on the GMDH RS-HDMR Sobol sensitivity analysis algorithm is currently underway. The method described here has been implemented as a Python script and is being developed into a standalone package and a more detailed summary of the technical aspects and applications of the method will be reported elsewhere.

REFERENCES

Cioppa, T. M. and T. W. Lucas (2007). "Efficient Nearly Orthogonal and Space-Filling Latin Hypercubes." *Technometrics* 49 (1): 45–55.

Delgado, P., P. Kelley, N. Murray, and A. Satheesh (2012). e*Water Source: User Guide*, eWater CRC (Australia).

Ellis, R. J. & R. D. Searle (2014). "Dynamic SedNet Component Model Reference Guide,Concepts and Algorithms Used in Source Catchments Customisation Plugin for Great Barrier Reef Catchment Modelling." Bundaberg, Queensland, Australia: Queensland Department of Science, Information Technology, Innovation and the Arts.

Farlow, S. J. (1981). "The GMDH Algorithm of Ivakhnenko." *The American Statistician* 35 (4): 210-15.

Gu, L., and J. Yang (2013). "Construction of Nearly Orthogonal Latin Hypercube Designs." *Metrika* 76 (6): 819–30.

Hoeffding, W. (1992). "A Class of Statistics with Asymptotically Normal Distribution." In *Breakthroughs in Statistics*, edited by Samuel Kotz and Norman L. Johnson, 308–34. New York, NY: Springer New York.

Horner, J. M. (1965). "Generalized Rodriques Formula Solutions for Certain Linear Differential Equations." *Transactions of the American Mathematical Society* 115: 31–42.

Iooss, B., and P. Lemaître (2015). "A Review on Global Sensitivity Analysis Methods." In *Uncertainty Management in Simulation-Optimization of Complex Systems*, edited by Gabriella Dellino and Carlo Meloni, 59:101–22. Boston, MA: Springer US.

Ivakhnenko, A. G. (1971). "Polynomial Theory of Complex Systems." *IEEE Transactions on Systems, Man, and Cybernetics* SMC-1 (4): 364–78.

James, A. (1993). "An Introduction to Water Quality Modelling. 2nd Edition." *An Introduction to Water Quality Modelling. 2nd Edition*.

Kelley, P., and A. O'Brien (2012). e*Water Source: Scientific Reference Guide*. eWater CRC (Australia).

Kleijnen, J. P. C., S. M. Sanchez, T. W. Lucas, and T. M. Cioppa (2005). "State-of-the-Art Review: A User's Guide to the Brave New World of Designing Simulation Experiments." *INFORMS Journal on Computing* 17 (3): 263–89.

Lambert, R. S.C., F. Lemke, S. S. Kucherenko, S. Song and N. Shah (2016). "Global Sensitivity Analysis Using Sparse High Dimensional Model Representations Generated by the Group Method of Data Handling." *Mathematics and Computers in Simulation* 128 (October): 42–54.

Li, G., S. Wang and H. Rabitz (2002). "Practical Approaches To Construct RS-HDMR Component Functions." *The Journal of Physical Chemistry A* 106 (37): 8721–33.

Liu, D. and Z. Zou (2012). "Sensitivity Analysis of Parameters in Water Quality Models and Water Environment Management." *Journal of Environmental Protection* 03 (08): 863–70. doi:10.4236/jep.2012.328101.

Manache, G., and C. S. Melching (2004). "Sensitivity Analysis of a Water-Quality Model Using Latin Hypercube Sampling." *Journal of Water Resources Planning and Management* 130 (May): 232–42.

McCloskey, G.L., D. Waters, R. Baheerathan, S. Darr, C. Dougall, R. Ellis, B. Fentie, and L. Hateley (2017). "Modelling Pollutant Load Changes Due to Improved Management Practices in the Great Barrier Reef Catchments: Updated Methodology and Results – Technical Report for Reef Report Cards 2015." Brisbane, Queensland: Queensland Department of Natural Resources and Mines.

Nossent, J., P. Elsen, and W. Bauwens (2011). "Sobol' Sensitivity Analysis of a Complex Environmental Model." *Environmental Modelling & Software* 26 (12): 1515–25.

Francesca, P., K. Beven, J. Freer, J. W. Hall, J, J. Rougier, D. B. Stephenson, and T. Wagener (2016). "Sensitivity Analysis of Environmental Models: A Systematic Review with Practical Workflow." *Environmental Modelling & Software* 79 (May): 214–32.

Herschel, F., F. Ö. F. Aliş, J. Shorter, and K. Shim. (1999). "Efficient Input—output Model Representations." *Computer Physics Communications* 117 (1–2): 11–20.

Renard, K. G., G. R. Foster, G. A. Weesies, and J. P. Porter (1991). "RUSLE: Revised Universal Soil Loss Equation." *Journal of Soil and Water Conservation* 46 (1): 30–33.

Shafii, M. and F. De Smedt (2009). "Multi-Objective Calibration of a Distributed Hydrological Model (WetSpa) Using a Genetic Algorithm." *Hydrology and Earth System Sciences* 13 (11): 2137–49.

Sincock, A. M., H. S. Wheater, and P. G. Whitehead (2003). "Calibration and Sensitivity Analysis of a River Water Quality Model under Unsteady Flow Conditions." *Journal of Hydrology* 277 (3–4): 214–29.

Sobol, I.M. (2001). "Global Sensitivity Indices for Nonlinear Mathematical Models and Their Monte Carlo Estimates." *Mathematics and Computers in Simulation* 55 (1–3): 271–80.

Song, X., J. Zhang, C. Zhan, Y. Xuan, M. Ye, and C. Xu (2015). "Global Sensitivity Analysis in Hydrological Modeling: Review of Concepts, Methods, Theoretical Framework, and Applications." *Journal of Hydrology* 523 (April): 739–57.

Zadeh, F. K., J. Nossent, F. Sarrazin, F. Pianosi, A. van Griensven, T. Wagener, and W. Bauwens (2017). "Comparison of Variance-Based and Moment-Independent Global Sensitivity Analysis Approaches by Application to the SWAT Model." *Environmental Modelling & Software* 91 (May): 210–22.