19th International Congress on Modelling and Simulation, Perth, Australia, 12–16 December 2011
http://mssanz.org.au/modsim2011

Metamodelling in sustainable environmental management

M.G. Erechtchoukova [a] **and P.A. Khaiter** [a]

[a] *School of Information Technology, Faculty of Liberal Arts and Professional Studies, York University, Canada*
Email: marina@yorku.ca

Abstract: To evaluate decision's sustainability, it is necessary to determine and assess the values of current and future welfare outcomes which, in turn, depend on the current and predicted status of the environment. These tasks make the application of models and mathematical tools unavoidable and justify the necessity of quantitative indicators of sustainability in decision and policy making, since environmental models are aimed to produce the results which complement observations on environmental parameters where they cannot be obtained directly. At the same time, concerns raised by the scientists and practitioners in recent years led to a suggestion that the complexity of the environmental models is one of the main obstacles in their wider use by the stakeholders. Therefore, complexity reduction is an important task for the successful application of the environmental models in the practical environmental decision-making and management.

The term 'complexity' is used in conjunction with a computational algorithm in order to describe its efficiency during the runtime. The comparison of the complexity of any two models describing the same ecosystem must take into account the following model features: the total number of state variables included into the model, the total number of model parameters and the non-linear features of the model. It is worth noting that, in general, these three features are independent. While first two characteristics can be expressed through the complexity index, the third one cannot be easily quantified and added to the index. An obvious suggestion is that the implementation algorithms used to obtain the model solutions must also be considered in deciding on the comparative complexity of the models. In this case, the effect of non-linear terms could be taken into account, at least to some extent. Commonly used statistical approaches to building an emulator of a complex model include response surface method (RSM), neural network (NN) and kriging. In all three cases, the emulators are constructed using mathematical techniques which significantly differ from those used in the original models. This means that the emulators have sets of own parameters which do not overlap with the original model parameter sets in terms of their practical meaning or their values.

Environmental models can be used in environmental management within the following settings: (1) to test possible scenarios via "what-if" analysis; (2) to find an optimal or at least satisficing scenario via optimization methods; (3) to determine key factors for a case study at hand. The replacement of an original model by an emulator looks very attractive with one reservation: it is necessary to ensure that the replacement is valid. It is obvious that in general case equal or very close values of two functions in certain points do not guarantee that their derivatives will also have close values. This means that emulators of complex environmental models can be used in the tasks which require only values of model state variables to complete the investigation. If the problem calls for optimization methods, it is necessary to ensure that the emulator contains all relevant state variables permitting to find a solution, and only non-gradient methods can be recommended to find a solutions to avoid misleading results.

Keywords: *Model complexity, emulator, sensitivity analysis, cascading simulation framework*

1. INTRODUCTION

Sustaining the environment via maintaining its functions in order to provide economic and social needs becomes vitally important. The World Commission on Environment and Development (Brundtland, 1987) determined the concept of sustainable development as a "development that meets the needs of present [generations] without compromising the ability of future generations to meet their own needs". To evaluate decision's sustainability, it is necessary to determine and assess the values of current and future welfare outcomes which, in turn, depend on the current and predicted status of the environment. These tasks make application of models and mathematical tools unavoidable and explain the necessity to use quantitative indicators of sustainability in decision and policy making, since environmental models are aimed to produce the results which complement observations on environmental parameters when they cannot be obtained directly.

Environmental decision-making deals with ecological systems and the necessity to predict their behaviour. A lot has been said about the fact that ecosystems are complex, dynamic and spatially heterogenous objects, in which physical, chemical and biological processes are closely interrelated and interdependent (e.g., Antle *et al*., 2001; Levin, 1999). Sustainable management of natural resources and environmental systems requires an adequate consideration of various ecological and socio-economic services provided by ecosystems. An idea of sustainable environmental management is only possible if multiple goods and services generated by an ecosystem are properly identified, quantified, valuated, predicted and forwarded to decision-makers at the early stages of the process. This makes environmental models a key tool in the assessment. The diversity of stakeholders of environmental models including governmental authorities, researchers, IT specialists, NGOs, and private sector leads to a variety of expectations and perceptions regarding potential benefits and values of the information provided by environmental models (Mysiak et al., 2008). Concerns raised in the responses to a global questionnaire on the use of environmental models and decision support tools led to a suggestion that the complexity of the environmental models is one of the main obstacles in their wider use by the respondents (McIntosh and Diez, 2008). Therefore, complexity reduction is an important task for the successful application of the environmental models for the practical environmental decision-making and management.

The role of complex models has received due attention in modern literature (e.g., Reichert & Omlin, 1997; Van Ness & Scheffer, 2005). Environmental models of lower complexity are preferable for researchers and decision makers, since they allow for comprehensive analysis of a problem at hand and extensive simulation experiments. An application of a model of reduced complexity (an emulator) which is suitable for a decision making process is called metamodelling. A survey of emulation techniques has been undertaken by Simpson et al. (2001) who considered four metamodelling approaches: response surface method, neural networks, inductive learning, and kriging and provided some recommendations on using these approaches.

The paper analyses the issue of complexity for process-based environmental models, its possible definitions and approaches to the development of environmental models of reduced complexity. The study shows that the way of model complexity reduction depends on the nature of its usage and the required model analysis and that application of emulators can be validated only for specific types of simulation experiments.

2. MODEL COMPLEXITY

One of the approaches to environmental modeling rests on process-based models. Environmental indicators relevant to an investigated case study are selected and their spatial or temporal dynamics is imitated by describing natural processes affecting the indicators based on mathematical formulae. The set of indicators determines the number of model state variables and processes which must be taken into account. The processes contributing significantly to indicators' variability are included into a model using balance equations. It may call for additional processes to be added to the model and each process to be described by a particular mathematical term. Alternatively, several natural processes can be described by a single aggregated term.

It might seem that the more processes are taken into account and included into the model, the better simulation results describe the reality, i.e. the more precise the model is. At the same time, an increasing level of detail results in a more complex model. This logic leads to an intuitive understanding of model complexity which implies that the more complex model is the more state variables and model parameters are introduced into the model.

In computer science, the term 'complexity' is used in conjunction with a computational algorithm in order to describe its efficiency during runtime. Although an accurate evaluation of the algorithm performance needs estimates of the running time and required process space, the primary consideration is given to the number of basic operations the algorithm requires to process an input data set of a certain size. Such estimates are asymptotic. They help to decide on the applicability of the algorithm to a given data set, however, it is hard to use these estimates for comparing two different algorithms with similar complexity.

Since simulation models are computer programs, the concept of computational complexity can be employed to describe their efficiency. At the same time, these computer programs use sets of parameters which values are required for simulation runs and must be determined prior to simulation experiments. Most of these values are not available through direct observations and depend not only on a given case study, but also on mathematical tools used to model real world processes. The process of identifying the values of model parameters is another computational algorithm which can be even more sophisticated than the algorithm implementing the model itself. Thus, the process of model parameter identification must also be taken into account while defining the concept of 'model complexity'

2.1. Complexity index

An attempt to formalize the concept of model complexity was made by Snowling & Kramer (2001) where the complexity index was introduced. The index was aimed to take into account the model structure and the level of detail in the description of processes included into the model. The index counts the number of state variables, the number of processes included in the model, the number of parameters and number of arithmetic operations for each term of the model and for the entire model. The complexity index can be evaluated from a Petersen matrix (Petersen, 1965). The index can be used to compare models with different mathematical expressions to identify the more complex one. Unfortunately, the complexity index does not reflect the type of mathematical terms used in the model, since both linear and non-linear terms can be described by the same score, whereas non-linear models should obviously be considered as more complex compared to the linear ones.

2.2. Non-linear models

When an environmental model is built based on conservation laws, the same natural processes can be described by either linear or non-linear terms. The linear expressions result in the algorithms of lower complexity, while non-linear models often require iterations to obtain a solution. The non-linearity of expressions becomes even more important when models are built based on differential equations. For such models, the non-linearity introduces additional stationary points of equilibrium which change the stability portrait of the model solutions affecting their dynamic behavior. The model of a one-species population serves a good example. With constant rates of the population's reproduction and mortality, the model describes Malthus's Law. When these rates are represented as functions of the population's density, the model solutions become limited and may describe different types of population's dynamics (Svirezhev and Logofet, 1983). That is why, a simple replacement of non-linear terms for certain processes by their linear approximation can be done for applications, in which simulations actually interpolate values of state variables within a limited period of time. However, such substitution may not be valid for long term predictions where complex interactions of processes create notable effects on investigated state variables. Likewise, reactions of an investigated system to various perturbations cannot be fully described by only linear approximation of a model when interactions are imitated by non-linear terms.

The comparison of the complexity of any two models of the same ecosystem must take into account the following model features: the total number of state variables included in the model, the total number of model parameters and the non-linear features of the model. It is worth noting that, in general these three features are independent. While first two characteristics can be described using the complexity index, the third one cannot be easily quantified and added to the index. An obvious recommendation is that the implementation algorithms used for obtaining the model solutions must be also considered when deciding on the comparative complexity of the models. In this case, the effect of non-linear terms could be taken into account, at least to some extent.

3. BUILDING AN EMULATOR

When a decision making process requires large number of simulation experiments, models of reduced complexity look very appealing provided that they describe the real world systems as well as complex models do. In the traditional problem setting for engineering analysis, to build an emulator means to find an approximation $\boldsymbol{F}$ of the outputs $\mathbf{y}$ generated by the model $\boldsymbol{M}$ using inputs $\mathbf{x}$ such that $\boldsymbol{F}$ is more efficient to run and at the same time provides insights into the relationships between $\mathbf{x}$ and $\mathbf{y}$. Inputs $\mathbf{x}$ are considered as controlling factors which can be varied in order to change outputs. The most generic framework for metamodelling consists of three main steps: (1) to choose experimental design points $\mathbf{x}_d$ which determine the efficient set of computer runs for generating outputs; (2) to choose a model (an emulator) for the approximation; (3) to fit the model to generated outputs (Simpson et al., 2001). While such technical issues as non-stochastic nature of the simulation results, the dependency of experimental designs on the chosen emulator are well understood and their importance is appreciated, the current study focused on the selection of the type of an emulator from application perspectives.

With respect to environmental models, it is important to distinguish between model state variables $\mathbf{x}^S$ and model parameters $\mathbf{p}$, which values must be specified before simulation experiments. Although initial values of model state variables and values of model parameters affect the obtained solutions and both form input data set $\mathbf{x}$, their influence is analyzed differently. Model responses to variations of initial values of the model state variables are investigated through the model stability analysis. It is worth noting that only stable solutions of simulation models can be used for environmental management purposes. Model parameters usually reflect factors which are external to the system and control actions. The effect of changes in the model parameter values on the model solutions (or outputs) can be investigated through the sensitivity analysis, which helps to identify the most important model parameters for a given set of initial values of the model state variables.

Commonly used statistical approaches to building an emulator of a complex model include response surface method (RSM), neural network (NN) and kriging. In RSM, a lower order polynomial approximation is determined which minimizes the squares of errors or other fitting criteria. Neural networks are based on functions of a specific type and they are connected into a network of a particular architecture which can be thought as a regression analysis on the specific functions from statistical point of view (Cheng et al., 1994). Kriging is an optimal linear interpolation applied to a stochastic process. In all three cases, the emulators are constructed using mathematical techniques which significantly differ from those used in the original models. This means that the emulators have their own sets of parameters which do not overlap with the original model parameter sets in terms of their practical meaning or values.

3.1. Problem settings for application

Environmental models can be used in environmental management within different settings. The following three settings have been considered in the study: (1) to test possible scenarios via “what-if” analysis; (2) to find an optimal or at least satisficing scenario via optimization analysis; (3) to determine key factors for a case study at hand. There are publications reporting on the application of emulators in all three settings (e.g., Shahsavani and Grimvall, 2011; Makler-Pick et al., 2011). While the first type of analysis can undoubtedly be done either on the original model or on the emulator, in the second and third types, simulations may become computationally unaffordable due to large numbers of required simulation runs. Thus, optimization problems are solved based on algorithms which require iterative evaluation of corresponding goal functions or their derivatives until the solution is obtained. The third setting of the problem of environmental management is essentially based on model sensitivity analysis. Therefore, the replacement of an original model by an emulator looks very attractive with one reservation: it is necessary to ensure that the replacement is valid.

Setting 2 – optimization analysis

Strictly speaking, an optimization analysis requires articulation of a goal function ***G*** which distinguishes between deferent sets of values of the model parameters and initial values of state variables and helps to determine the optimal or satisficing solution. A constraint function restricting the search space can also be specified. The Lagrange multiplier method can be used to transform a constrained optimization problem to a non-constrained one. Thus, in the most generic case, the optimization analysis can be converted to the problem:

$$\min G(M(\mathbf{x}^S, \mathbf{p})) \,. \tag{1}$$

Given that evaluating of ***M*** even at a single point is computationally expensive, one may see the replacement of model ***M*** by its emulator ***F*** as an obvious and appropriate course of actions. However, it is necessary to consider the optimization algorithm employed to solve the problem (1).

While non-gradient optimization involves evaluation and comparison of the goal function ***G*** in representative points of the search space, gradient methods require calculation of partial derivatives of the goal function on the model state variables and/or parameters. The existing methods of emulator development use general criteria to approximate the solution of a complex model by a simplified function with a desired level of accuracy. It supports expectations that the emulator will generate values which are very close to those generated by the original model. As soon as model derivatives must be evaluated, emulators become of little help, since their reduced complexity does not allow for accurate evaluation of model derivatives.

The following issues must be taken into account. Depending on the way the emulator ***F*** has been derived, it may have a reduced set of state variables and its own parameter set which is different from the set **p** (e.g., Khaiter and Erechtchoukova, 2007). It is obvious that in general case

$$\frac{\partial G(M(\mathbf{x}^S, \mathbf{p}))}{\partial x_i^S} \neq \frac{\partial G(F(\mathbf{x}_{new}^S, \mathbf{q}))}{\partial x_i^S} \,, \tag{2}$$

where $\mathbf{x}_{new}^S$ is the subset of the original model state variables, which remain in the emulator, $\boldsymbol{q}$ is the set of emulator parameters.

Setting 3 – identification of key factors

Sensitivity analysis is an important step in a model development and validation (Jakeman et al., 2006). Along with that, sensitivity analysis allows for identification of the parameters which contribution to variability of the model solution exceeds others, thus, indicating which natural and anthropogenic factors are most important. Although the sensitivity analysis can be implemented based on local or global schemes, the evaluation of model output variations in response to a parameter perturbations is common for both schemes. Thus, the derivative $\partial \mathbf{y} / \partial p_j$ can be interpreted as a mathematical definition of the local sensitivity of the output $\boldsymbol{y}$ versus parameter p_j (Saltelli et al., 2008). As it has been mentioned above, the original model and its emulator have different parameter sets and obviously:

$$\frac{\partial G(M(\mathbf{x}^S, \mathbf{p}))}{\partial p_j} \neq \frac{\partial G(F(\mathbf{x}_{new}^S, \mathbf{q}))}{\partial p_j} \,. \tag{3}$$

Inequality (3) shows that the results of global sensitivity analysis implemented on the original model and on the emulators will also be different. Thus, any general recommendation to introduce an emulator to investigate the original model sensitivity cannot be valid.

4. AN ALTERNATIVE APPROACH TO COMPEXITY REDUCTION

Another commonly accepted approach to reduce model complexity suggests to separate main groups of processes and to model the groups using individual modules which all are linked together in a cascade. Interactions of processes from different groups are modeled by passing the simulation results on from one module to another (Ambrose et al., 1993; Argent et al., 2006). Individual modules in the cascading simulation framework support different forms of equations which can be chosen depending on data available for a given case study. Processes from different groups may also have different scales in time and space. A cascading simulation framework significantly reduces the computational time and space required for simulation runs. The modules can be investigated separately and sensitivity analysis can be implemented based on original mathematical expression (Erechtchoukova, 2005).

5. DISCUSSION AND CONCLUSIONS

Expressions (2) and (3) confirm the fact that equal or very close values of two functions in a certain point does not guarantee that their derivatives will also have close values. This means that emulators of complex environmental models can be used when only model outputs are required to complete the investigation. If the problem calls for optimization methods, it is necessary to ensure that the emulator contains all state variables permitting to find a solution, and only non-gradient methods can be recommended to find a solutions to avoid misleading results.

Strictly speaking, a complete sensitivity analysis of an original model based on an emulator constructed according to approaches mentioned above is hardly possible. Nevertheless, the attempts to evaluate model sensitivity using surrogate models are reported in the literature. With no intention to undermine the works done, it is important to stress out that such applications must be considered on a case-by-case basis. It is necessary to verify that derivatives of the investigated model and derivatives of its emulator have very close values at the design points $\mathbf{x}_d$.

In many cases, complex environmental models with large number of state variables generate prediction for indicators of environmental states. All these variables are necessary for simulation, but only a few of them correspond to indicators that are of interest or importance from the problem perspectives. It is advisable to investigate the part of the model corresponding to these relevant state variables and to determine sensitivity of these state variables to model parameters. The approach to model complexity reduction based on the cascading simulation framework can help to obtain realistic assessment of model sensitivity and to determine key factors affecting the outcomes of a project at hands. In any case, the choice of a technique for metamodelling must be based on a clear understanding of the options available for model usage and possible settings of simulation experiments.

ACKNOWLEDGMENTS

The authors are grateful to anonymous reviewers for their thoughtful comments and suggestions on the improvement of the manuscript.

REFERENCES

Ambrose, R.B., Wool, T.A. and Martin, J.L. (1993). The Water Quality Analysis Simulation Program, WASP5. Part A: Model documentation. Athens, GA: USEPA ERL.

Antle, J.M., Capalbo, S.M., Elliott, E.T., Hunt, H.W., Mooney, S. and Paustian, K.H. (2001). Research needs for understanding and predicting the behaviour of managed ecosystems: lessons from the study of agroecosystems. *Ecosystems*, 4: 723-735.

Argent, R.M., Voinov, A., Maxwell, T., Cuddy, S.M., Rahman, J.M., Seaton, S., Vertessy, R.A. and Braddock, R.D. (2006). Comparing modeling frameworks – A workshop approach. *Environmental Modelling and Software*, 21: 895-910.

Brundtland, G. (1987). *Our Common Future: The World Commission on Environment and Development*. Oxford University Press, Oxford.

Cheng, B. and Titterington, D.M. (1994). Neural Networks: A review from a statistical perspective. *Statistical Science*, 9(1): 2-54.

Erechtchoukova, M. G. (2005). Uncertainty transformation in ecological simulation models. In: Zerger, A. and Argent, R. (Eds.) MODSIM2005 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand, December 2005, pp: 2477-2483.

Jakeman, A.J., Letcher, R.A. and Norton, J.P. (2006). Ten iterative steps in development and evaluation of environmental models. *Environmental Modelling and Software*, 21: 602-614.

Khaiter, P.A. and Erechtchoukova, M.G. (2007). From complex to simple in environmental simulation modeling. In: Oxley, L. and Kulasiri, D. (Eds.) MODSIM 2007 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand, December 2007, pp. 2069 - 2075. ISBN: 978-0-9758400-4-7.

Levin, S.A. (1999). *Fragile dominion: complexity and the commons*, Reading, Mass: Perseus Books.

Makler-Pick, V., Gal, G., Gorfine G., Hipsey, M.R. and Carmel Y. (2011). Sensitivity analysis for complex ecological models – A new approach. *Environmental Modelling and Software*, 26: 124-134.

McIntosh, B.S. and Diez, E. (2008). Assessing the impact of environmental decision and information support tools. In: Sànchez-Marrè, M., Béjar, J., Comas, J., Rizzoli, A.E. and Guariso, (Eds.) 4th Int. Congress on Environmental Modelling and Software (iEMSs 2008), Barcelona, Catalonia: International Environmental Modelling and Software Society: 932-939.

Mysiak, J., Giupponi, C., Depietri, Y., and Colombini, G. (2008). A note on attitudes towards and expectation from the Decision Support Systems. In: Sànchez-Marrè, M., Béjar, J., Comas, J., Rizzoli, A.E. and Guariso, (Eds.) 4th Int. Congress on Environmental Modelling and Software (iEMSs 2008), Barcelona, Catalonia: International Environmental Modelling and Software Society: 925-931.

Petersen, E. (1965). *Chemical reaction analysis*. Prentice-Hall, Englewood Cliffs, New Jersey.

Reichert, P. and Omlin, M. (1997). On the usefulness of overparameterized ecological models. *Ecological Modelling*, 95: 289-299.

Saltelli, A., Ratto, M, Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Sainsana, M. and Tarantola, S. (2008). *Global Sensitivity Analysis. The Primer.* John Wiley & Sons, Chichester.

Simpson, T.W., Peplinski, J.D., Koch, P.N. and Allen, J.K. (2001). Metamodels for computer-based engineering design: survey and recommendations. *Engineering with Computers*, 17:129-150.

Snowling, S.D. and Kramer, J.R. (2001). Evaluating modeling uncertainty for model selection. *Ecological Modelling*, 138: 17-30.

Shahsavani, D., and Grimvall, A. (2011). Variance-based sensitivity analysis of model outputs using surrogate models. *Environmental Modelling and Software*, 26 (6): 723-730.

Svirezhev, Yu.M. and Logofet, D.O. (1983). *Stability of ecological communities.* Mir Publisher, Moscow.

Van Ness, E.H. and Scheffer, M. (2005). A strategy to improve the contribution of complex simulation models to ecological theory. *Ecological Modelling*, 185: 153-164.