Environmental Modelling & Software 34 (2012) 51–66

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

A comparison of eight metamodeling techniques for the simulation of N_2O fluxes and N leaching from corn crops

Nathalie Villa-Vialaneix [a,b,*], Marco Follador [c], Marco Ratto [d], Adrian Leip [c]

[a] IUT de Perpignan (Dpt STID, Carcassonne), Univ. Perpignan, Via Domitia, France
[b] Institut de Mathématiques de Toulouse, Université de Toulouse, France
[c] European Commission, Institute for Environment and Sustainability, CCU, Ispra, Italy
[d] European Commission, Econometrics and Applied Statistics Unit, Ispra, Italy

ARTICLE INFO

Article history:
Received 18 October 2010
Received in revised form
15 April 2011
Accepted 1 May 2011
Available online 8 June 2011

Keywords:
Metamodeling
Splines
SVM
Neural network
Random forest
N_2O flux
N leaching
Agriculture

ABSTRACT

The environmental costs of intensive farming activities are often under-estimated or not traded by the market, even though they play an important role in addressing future society's needs. The estimation of nitrogen (N) dynamics is thus an important issue which demands detailed simulation based methods and their integrated use to correctly represent complex and non-linear interactions into cropping systems. To calculate the N_2O flux and N leaching from European arable lands, a modeling framework has been developed by linking the CAPRI agro-economic dataset with the DNDC-EUROPE bio-geo-chemical model. But, despite the great power of modern calculators, their use at continental scale is often too computationally costly. By comparing several statistical methods this paper aims to design a metamodel able to approximate the expensive code of the detailed modeling approach, devising the best compromise between estimation performance and simulation speed. We describe the use of two parametric (linear) models and six non-parametric approaches: two methods based on splines (ACOSSO and SDR), one method based on kriging (DACE), a neural networks method (multilayer perceptron, MLP), SVM and a bagging method (random forest, RF). This analysis shows that, as long as few data are available to train the model, splines approaches lead to best results, while when the size of training dataset increases, SVM and RF provide faster and more accurate solutions.

1. Introduction

The impact of modern agriculture on the environment is well documented (Power, 2010; Tilman et al., 2002; Scherr and Sthapit, 2009; FAO, 2007, 2005; Singh, 2000; Matson et al., 1997). Intensive farming has a high consumption of nitrogen, which is often inefficiently used, particularly in livestock production systems (Leip et al., in press a; Webb et al., 2005; Oenema et al., 2007; Chadwick, 2005). This leads to a large surplus of nitrogen which is lost to the environment. Up to 95% of ammonia emission in Europe have their origin in agricultural activities (Kirchmann et al., 1998; Leip et al., 2011) contributing to eutrophication, loss of biodiversity and health problems. Beside NH_3, nitrate leaching below the soil root zone and entering the groundwater poses a particular problem for the quality of drinking water (van Grinsven et al., 2006). Additionally, agricultural sector is the major source of anthropogenic emissions of N_2O from the soils, mainly as a consequence of the application of mineral fertilizer or manure nitrogen (Del Grosso et al., 2006; Leip et al., in press b, 2005; European Environment Agency, 2010). N_2O is a potent greenhouse gas (GHG) contributing with each kilogram emitted about 300 times more to global warming than the same mass emitted as CO_2, on the basis of a 100-years time horizon (Intergovernmental Panel on Climate Change, 2007).

Various European legislations attempt to reduce the environmental impact of the agriculture sector, particularly the Nitrates Directive (European Council, 1991) and the Water Framework Directive (European Council, 2000). Initially, however, compliance to these directives was poor (Oenema et al., 2009; European Commission, 2002). Therefore, with the last reform of the Common Agricultural Policy (CAP) in the year 2003 (European Council, 2003), the European Union introduced a compulsory Cross-Compliance (CC) mechanism to improve compliance with 18 environmental, food safety, animal welfare, and animal and plant health standards (Statutory Management Requirements, SMRs) as well as with requirements to maintain farmlands in good

* Corresponding author. Institut de Mathématiques de Toulouse, Université Paul Sabatier, 118 route de Narbonne, F-31062 Toulouse cedex 9, France. Tel.: +33 5 61 55 63 58.
E-mail address: nathalie.villa@math.univ-toulouse.fr (N. Villa-Vialaneix).

1364-8152/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2011.05.003

agricultural and environmental condition (Good Agricultural and Environment Condition requirements, GAECs), as prerequisite for receiving direct payments (European Union Commission, 2004, 2009; European Council, 2009; Dimopoulus et al., 2007; Jongeneel et al., 2007). The SMRs are based on pre-existing EU Directives and Regulations such as Nitrate Directives. The GAECs focus on soil erosion, soil organic matter, soil structure and a minimum level of maintenance; for each of these issues a number of standards are listed (Alliance Environnement, 2007).

It remains nevertheless a challenge to monitor compliance and to assess the impact of the cross-compliance legislations not only on the environment, but also on animal welfare, farmer's income, production levels, etc. In order to help with this task, the EU-project Cross-Compliance Assessment Tool (CCAT) developed a simulation platform to provide scientifically sound and regionally differentiated responses to various farming scenarios (Elbersen et al., 2010; Jongeneel et al., 2007).

CCAT integrates complementary models to assess changes in organic carbon and nitrogen fluxes from soils (De Vries et al., 2008). Carbon and nitrogen turnover are very complex processes, characterized by a high spatial variability and a strong dependence on environmental factors such as meteorological conditions and soils (Shaffer and Ma, 2001; Zhang et al., 2002). Quantification of fluxes, and specifically a meaningful quantification of the response to mitigation measures at the regional level requires the simulation of farm management and the soil/plant/atmosphere continuum at the highest possible resolution (Anderson et al., 2003; Leip et al., in press b). For the simulation of N_2O fluxes and N leaching, the process-based biogeochemistry model DNDC-EUROPE (Leip et al., 2008; Li et al., 1992; Li, 2000) was used. As DNDC-EUROPE is a complex model imposing high computational costs, the time needed to obtain simulation results in large-scale applications (such as the European scale) can be restrictive. In particular, the direct use of the deterministic model is prohibited to extract efficiently estimations of the evolution of N_2O fluxes and N leaching under changing conditions. Hence, there is a need for a second level of abstraction, modeling the DNDC-EUROPE model itself, which is called a *metamodel* (see Section 2 for a more specific definition of the concept of metamodeling). Metamodels are defined from a limited number of deterministic simulations for specific applications and/or scenario and allow to obtain fast estimations.

This issue is a topic of high interest that has previously been tackled in several papers: among others, Bouzaher et al. (1993) develop a parametric model, including spatial dependency, to model water pollution. Krysanova and Haberlandt (2002) and Haberlandt et al. (2002) describe a two-steps approach to address the issue of N leaching and water pollution: they use a process-based model followed by a location of the results with a fuzzy rule. More recently, Pineros Garcet et al. (2006) compare RBF neural networks with kriging modeling to build a metamodel for a deterministic N leaching model called WAVE (Vanclooster et al., 1996). The present article compares in detail different modeling tools in order to select the most reliable one to metamodel the DNDC-EUROPE tasks in the CCAT project (Follador and Leip, 2009). This study differs from the work of Vanclooster et al. (1996) because of the adopted European scale and of the analysis of 8 metamodeling approaches (also including a kriging and a neural network method). The comparison has been based on the evaluation of metamodel performances, in terms of accuracy and computational costs, with different sizes of the training dataset.

The rest of the paper is organized as follows: Section 2 introduces the general principles and advantages of using a metamodel; Section 3 reviews in details the different types of metamodels compared in this study; Section 4 explains the Design Of the Experiments (DOE) and show the results of the comparison, highlighting how the availability of the training data can play an important role in the selection of the best type and form of the approximation. The supplementary material of this paper can be found at: http://afoludata.jrc.ec.europa.eu/index.php/dataset/detail/232.

2. From model to meta model

A model is a simplified representation (abstraction) of reality developed for a specific goal; it may be deterministic or probabilistic. An integrated use of simulation-based models is necessary to approximate our perception of complex and non-linear interactions existing in human-natural systems by means of mathematical input–output (I/O) relationships. Despite the continuous increase of computer performance, the development of large simulation platforms remains often prohibited because of computational needs and parameterization constraints. More precisely, every model in a simulation platform such as DNDC-EUROPE, is characterized by several parameters, whose near-optimum set is defined during the calibration. A constraint applies restrictions to the kind of data that the model can use or to specific boundary conditions. The flux of I/O in the simulation platform can thus be impeded by the type of data/boundaries that constraints allow – or not allow – for the models at hand.

The use of this kind of simulation platform is therefore not recommended for all the applications which require many runs, such as sensitivity analysis or what-if studies. To overcome this limit, the process of abstraction can be applied to the model itself, obtaining a model of the model (2nd level of abstraction from reality) called metamodel (Blanning, 1975; Kleijnen, 1975; Sacks et al., 1989; van Gighc, 1991; Santner et al., 2003). A metamodel is an approximation of detailed model I/O transformations, built through a moderate number of computer experiments.

Replacing a detailed model with a metamodel generally brings some payoffs (Britz and Leip, 2009; Simpson et al., 2001):

- easier integration into other processes and simulation platforms;
- faster execution and reduced storage needs to estimate one specific output;
- easier applicability across different spatial and/or temporal scales and site-specific calibrations, as long as data corresponding to the new system parameterization are available.

As a consequence, a higher number of simulation runs become possible: using its interpolatory action makes a thorough sensitivity analysis more convenient and leads to a better understanding of I/O relationships. Also it offers usually a higher flexibility and can quickly be adapted to achieve a wide range of goals (prediction, optimization, exploration, validation). However, despite these advantages, they suffer from a few drawbacks: internal variables or outputs not originally considered cannot be inspected and the prediction for input regimes outside the training/test set is impossible. Hence, a good metamodeling methodology should be able to provide fast predictions. But, considering that limitations, it also must have a low computational cost to be able to build a new metamodel from a new data set including new variables and/or a different range for these input variables.

Let $(\mathbf{X},\boldsymbol{y})$ be the dataset consisting of N row vectors of input/output pairs $(\mathbf{X}_i,y_i)$, where $\mathbf{x}_i = (x_i^1,\ldots,x_i^d)^T \in \mathbb{R}^d (i = 1,\ldots,N)$ are the model input and $y_i \in \mathbb{R} (i = 1,\ldots,N)$ are the model responses for N experimental runs of the simulation platform. The mathematical representation of I/O relationships described by the detailed model can be written as

$$y_i = f(\mathbf{x}_i) \quad i = 1,\ldots,N \tag{1}$$