Environmental Modelling & Software 47 (2013) 29–41

Contents lists available at SciVerse ScienceDirect

Environmental Modelling & Software

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

A typology of different development and testing options for symbolic regression modelling of measured and calculated datasets☆

Darren J. Beriro [a,*], Robert J. Abrahart [a], C. Paul Nathanail [a], Jimmy Moreno [b], A. Salim Bawazir [b]

[a] School of Geography, University of Nottingham, Nottingham NG7 2RD, UK
[b] Department of Civil Engineering, New Mexico State University, Box 30001, MSC 3CE, Las Cruces, NM 88003-0001, USA

ARTICLE INFO

Article history:
Received 3 September 2012
Received in revised form
18 January 2013
Accepted 26 March 2013
Available online 6 June 2013

Keywords:
Simulation
Emulation
Pan evaporation
Gene expression programming
Data driven modelling
Emulation simulation typology
Symbolic regression

ABSTRACT

Data-driven modelling is used to develop two alternative types of predictive environmental model: a simulator, a model of a real-world process developed from either a conceptual understanding of physical relations and/or using measured records, and an emulator, an imitator of some other model developed on predicted outputs calculated by that source model. A simple four-way typology called Emulation Simulation Typology (EST) is proposed that distinguishes between (i) model type and (ii) different uses of model development period and model test period datasets. To address the question of to what extent simulator and emulator solutions might be considered interchangeable i.e. provide similar levels of output accuracy when tested on data different from that used in their development, a pair of counterpart pan evaporation models was created using symbolic regression. Each model type delivered similar levels of predictive skill to that other of published solutions. Input–output sensitivity analysis of the two different model types likewise confirmed two very similar underlying response functions. This study demonstrates that the type and quality of data on which a model is tested, has a greater influence on model accuracy assessment, than the type and quality of data on which a model is developed, providing that the development record is sufficiently representative of the conceptual underpinnings of the system being examined. Thus, previously reported substantial disparities occurring in goodness-of-fit statistics for pan evaporation models are most likely explained by the use of either measured or calculated data to test particular models, where lower scores do not necessarily represent major deficiencies in the solution itself.

1. Introduction

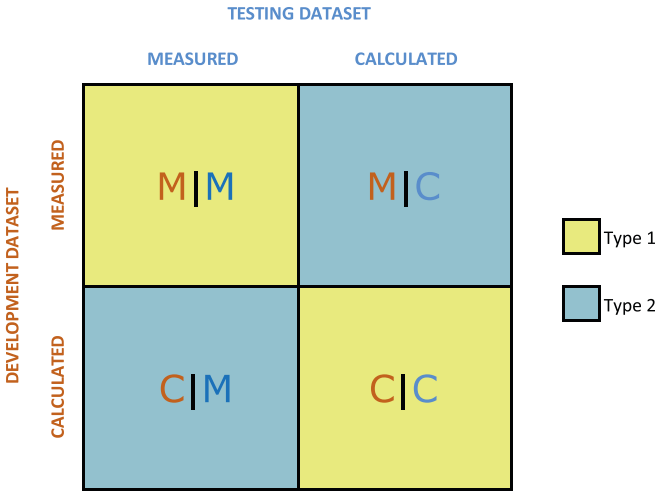
1.1. Purpose of study

For model-based approaches to remain credible tools in problem solving, a systematic and repeatable approach to iterative model development and evaluation tasks is required (Alexandrov et al., 2011; Bennett et al., 2012; Jakeman et al., 2006). Many data-driven modelling studies focus almost exclusively on goodness-of-fit metrics to determine the efficacy of solutions and little attention is paid to data provenance and/or to the conceptual underpinnings of the natural system being investigated, which consequently has a detrimental effect on scientific robustness and overall transparency of any findings. The present study has been inspired by an enthusiasm for more standardised approaches to environmental modelling explorations, including improved model evaluation (Abrahart et al., 2010; Bennett et al., 2012; Blocken and Gualtieri, 2012; Jakeman et al., 2006; Robson et al., 2008), resulting in our exposition of a data-driven modelling protocol that is able to answer our principal research question: does it really matter what type of data is modelled? A key outcome of this study is a new categorisation for research outcomes termed Emulation Simulation Typology (EST). This descriptor will enable researchers to distinguish between model type (simulator or emulator), as well as differences in model performance arising from the quality and type of data used in model development and testing.

Two different types of predictive model are recognised in our study: a simulator, a model of a real-world process developed on a conceptual understanding of physical relationships using

* Corresponding author. Tel.: +44 115 951 5428.
E-mail addresses: lgxdjb@nottingham.ac.uk, darren.beriro@gmail.com (D.J. Beriro).

1364-8152/$ – see front matter © 2013 The Authors. Published by Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.envsoft.2013.03.020



NOTES
Each EST scenario is coded for reference purposes as follows: the first letter signifies which type of dataset was used for model development; the second letter signifies which type of dataset was used to perform the subsequent model testing operation: in this case where M means use of a 'measured' dataset (EPAN) and C means use of a 'calculated' dataset (ECAL).

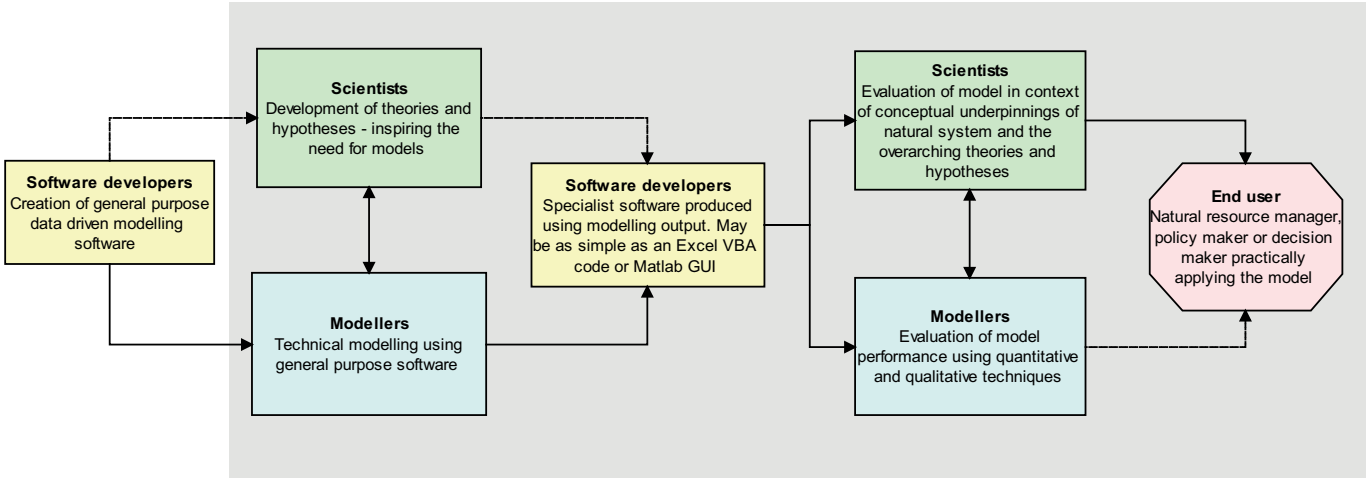
Type 1 models are developed and tested using the same sort of data i.e. measured or calculated.

Type 2 models are developed and tested using different sorts of data.

Fig. 1. Schematic of EST.

measured records (S_{MOD}), and an emulator, an imitator of some other model developed on predicted outputs calculated by that source model (E_{MOD}). This paper summarises the regulated exploration of two counterpart model types: i) a simulator, S_{MOD}, used to estimate measured pan evaporation (E_{PAN}); and ii) an emulator, E_{MOD}, used to estimate pan evaporation values originally calculated by means of the Nordenson–Fox equation (Burman and Pochop, 1994; Kohler et al., 1955) (E_{CAL}). Four independent model testing scenarios were envisaged in which each model that is developed is tested twice, once using E_{PAN} data, and then again using E_{CAL} data, as depicted by means of a 2 × 2 matrix in Fig. 1. The matrix illustrates that models may be developed using either measured E_{PAN} or calculated E_{CAL} data, and that subsequent testing could also be performed using either measured or calculated data, thus leading to four possible sets of findings. This study questions whether or not model performance is affected by which of the four development/testing combinations a given model can be assigned and examines the question of to what extent goodness-of-fit performance is related to particular scenarios. To assist with our explanation of matters, two types of model combination are defined: Type 1 models that are developed and tested using the same sort of data i.e. measured and measured or calculated and calculated; Type 2 models that are developed using one sort of data and tested on another.

It must be stressed that this analysis is not intended to establish whether emulators are better than simulators, or vice versa. Moreover, our study is not about searching for a superior predictive model, more accurate than anything published to date, and our models are not intended to represent general purpose solutions but rather specific fits to a particular dataset and study site. Finally, our study does not compare or benchmark its prediction accuracies against other types of data-driven daily E_{PAN} or E_{CAL} model such as those created by means of Multiple Linear Regression (MLR) (e.g. Abudu et al., 2011; Almedeij, 2012; Cooke et al., 2008; Tabari et al., 2010), Neural Network (NN) (e.g. Almedeij, 2012; Kim et al., 2012a, 2012b; Kisi, 2009; Kişi and Tombul, 2013; Moreno et al., 2010; Piri et al., 2009; Shiri et al., 2011; Shiri and Kisi, 2011; Shirsath and Singh, 2010; Tabari et al., 2010; Terzi and Keskin, 2010), Adaptive Neuro Fuzzy Inference System (ANFIS) (e.g. Chung et al., 2012; Dogan et al., 2010; Keskin et al., 2009; Kişi, 2006; Kişi et al., 2012; Sanikhani et al., 2012; Shiri et al., 2011) or Symbolic Regression (SR) (e.g. Guven and Kisi, 2010; Shiri et al., 2011; Shiri and Kişi, 2011; Shiri et al., 2013; Terzi, 2011, 2012). Such explorations have already been reported in past studies and need not be repeated. This study instead distinguishes between the different ideological contexts that underpin and separate simulator and emulator models in a new typology characterised by the different types of data that can be used for either model development and/or model testing purposes. This is done successfully by applying a standardised modelling protocol to examine EST and so in doing offers a mechanism for providing greater intelligibility of subsequently reported findings, for scientists, modellers, software developers and end users. We make this distinction between the different groups of people involved because the role each plays in inspiring, developing, evaluating and using outputs from data-driven modelling



Notes
Solid line indicates dominant pathway.
Dashed line indicates secondary input.
Greyed box shows stakeholders relevant to this study.

Fig. 2. Stakeholders involved in environmental modelling.