frontiers in Environmental Science

OPINION
published: 30 April 2015
doi: 10.3389/fenvs.2015.00033

Machine learning components in deterministic models: hybrid synergy in the age of data

Evan B. Goldstein[1]* and *Giovanni Coco*[2]

[1] *Department of Geological Sciences, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA,* [2] *School of Environment, University of Auckland, Auckland, New Zealand*

Keywords: hybrid model, machine learning, data driven modeling, hybrid modeling, model theory

OPEN ACCESS

Edited by:
Christian E. Vincenot,
Kyoto University, Japan

Reviewed by:
Mikhail Kanevski,
University of Lausanne, Switzerland
Nathaniel K. Newlands,
Science and Technology, Government of Canada, Canada

***Correspondence:**
Evan B. Goldstein,
evan.goldstein@unc.edu

Specialty section:
This article was submitted to Environmental Informatics, a section of the journal Frontiers in Environmental Science

Received: *05 March 2015*
Paper pending published: *29 March 2015*
Accepted: *16 April 2015*
Published: *30 April 2015*

Citation:
Goldstein EB and Coco G (2015) Machine learning components in deterministic models: hybrid synergy in the age of data. Front. Environ. Sci. 3:33. doi: 10.3389/fenvs.2015.00033

> "When the information changes, I change my mind. What do you do sir?"
> –paraphrased from Paul Samuelson (1970)[1].

Introduction

Physics-based numerical models designed to study processes on the surface of the Earth are commonly built with conservation laws. Yet conservations laws are based on a treatable subset of all physical and biological processes operating in the system of interest, and empirical relationships are always required to fully describe systems (and provide mathematical "closure"). In contrast to conservation laws, empirical expressions used in model construction are inductive, and based on observations of phenomena.

Any model that requires empirical expressions is also subject to revision as the empirical parameterization is refined: all empiricism is open to revision by corroborating or conflicting data. As more data becomes available, and more degrees of freedom are explored, it becomes harder to incorporate all available data into a single optimal empirical predictor. We argue in this contribution that empirical parameterizations for numerical models should be constructed using machine learning techniques because these techniques are built to operate on large, high dimensional datasets. Machine learning, in the context of this paper, defines a suite of algorithms used to develop predictive relationships (correlations) using a set of input data. Examples of commonly used machine learning techniques in the Earth sciences are artificial neural networks (e.g., Maier and Dandy, 2000; Pape et al., 2007; van Maanen et al., 2010), regression trees (e.g., Snelder et al., 2009; Oehler et al., 2012), Bayesian networks (e.g., Aguilera et al., 2011; Yates and Le Cozannet, 2012), and evolutionary algorithms (e.g., Knaapen and Hulscher, 2002; Ruessink, 2005; Goldstein et al., 2013). Machine learning techniques offer insight and are high-performance, reproducible, and scalable. The inclusion of powerful inductive techniques like those provided by machine learning offers the opportunity to enhance the predictions obtained from deductive approaches. Additionally the use of machine learning often leads to further insight. The use of empiricisms built from machine learning in a physics-based model results in a "hybrid" model.

Combining the strengths of inductive (data-driven) and deductive (physics-based) approaches in a single hybrid model has been suggested as a valuable step forward in model development because of increases in accuracy (e.g., Babovic et al., 2001; Hall, 2004) and speed (Krasnopolsky and Fox-Rabinovitz, 2006). An additional benefit is the direct coupling of models to data,

[1] This quote (and a discussion of its history) is available as a long-lived URL through the 'Internet Archive': https://web.archive.org/web/20150305151453/http://quoteinvestigator.com/2011/07/22/keynes-change-mind/comment-page-1/

especially valuable for exploratory models (Murray, 2003). Several studies have already used this approach, embedding machine learning components directly in models (e.g., Jain and Srinivasulu, 2004; Corzo et al., 2009; Goldstein et al., 2014; Limber et al., 2014) or during model calibration (Knaapen and Hulscher, 2002, 2003; Ruessink, 2005). The focus of this article is on machine learning components that are directly embedded within physics-based models. We believe the further use of this "hybrid" approach will result in more accurate model components (i.e., individual empirical predictors) and more accurate models. We do not intend to cast machine learning solely as a data fitting procedure. There is a rich set of problems where machine learning is applicable, but a clear use is in optimizing the fit of empirical predictors in nonlinear multivariate datasets.

Five main points highlight the advantages of this hybrid approach: These benefits are not without costs, mostly in the form of extra work/time and problems associated with less-rigorous usage of machine learning. As a convention we refer to machine learning model components as "parameterizations" and the hybrid model as the "model."

Machine Learning Highlights "Theory-Less" or "Data-Less" Model Components

Building machine learning predictors is usually motivated by a lack of theory or a perceived inadequacy in theory. Parameterizations that do not have an accurate or well-developed "theory" might negatively reflect upon model predictions because they are incorrect, or poorly parameterized. As an example, Goldstein et al. (2013) built a new predictor for bedforms generated under wave action motivated by the fact that no predictors were explicitly tested in conditions of large grain size variation and strong wave conditions. This predictor was eventually used in a larger "hybrid" numerical model with success (Goldstein et al., 2014).

Machine learning predictors also highlight heuristic or theoretical elements of a numerical model that do not have sufficient data to test. Both types of problems (lacking theory and lacking data) can motivate future research, specifically theory creation and targeted data collection.

Machine Learning Can Be Used to Gain New Theoretical Insight

Machine learning techniques on a dataset may provide theoretical insight. Crutchfield (2014) has termed this process "artificial science," where theoretical insight is derived directly from data. New machine learning techniques suggest that this is possible (Schmidt and Lipson, 2009). In this way developing predictors might provide theoretical insight into model or system behavior, at the very least giving new hypotheses to test. The wave ripple predictor developed by Goldstein et al. (2013) provides a new, inductively derived predictor for ripples under various grain sizes and forcing conditions. This new mathematical relationship describes an observed relationship (between grain size, wave forcing, and ripple size) derived from observations. This new relationship (an inductive statement) provides a testable hypothesis.

The Possibility for Emulation

Beyond detecting a new theory, hypothesis, or physical relationship, machine learning may provide a more parsimonious empirical relation than existed previously (e.g., Tinoco et al., 2015). This could help to speed up model runtime, a main goal of previous hybrid model work where the use of artificial neural networks allowed for emulation of entire components of a global climate model based on physical processes with no accuracy loss (e.g., Krasnopolsky and Fox-Rabinovitz, 2006). The computational gain associated with the use of a hybrid model cascades into a series of additional advantages including the possibility of simulating more scenarios, decreasing grid size or exploring finer-scale parameterizations.

Machine Learning Outperforms Arbitrary Curve Fitting

Multidimensional empirical parameterization are often built by assembling the data, collapsing the data to a two- dimensional plane, and fitting a user–defined function through the data cloud. Several steps require "user input," and may be arbitrary. First, the collapse of the multidimensional data onto a 2D plane may require developing nondimensional groups. Though the number of nondimensional groups is mandated by the well-known Buckingham's Pi theorem, the actual makeup of each group is not, and is often guided by utility, physical reasoning or user intuition (e.g., Bridgman, 1922). Second, a user defined curve must be selected to fit to the data. This curve may not be the most optimal basis function to fit to the data.

Both of these ambiguities are avoided in machine learning because: (1) user input can be the raw data parameters, or all possible nondimensional parameter groupings (e.g., Tinoco et al., 2015) and (2) machine learning often does not require the use of a set basis function, or the basis function is sufficiently flexible to allow the approximation of any arbitrary function. These benefits suggest that machine learning is a powerful set of tools for developing parameterizations when data is high-dimensional, noisy, and nonlinear: these techniques outperform traditional curve fitting for their ability to truly provide an optimal curve to be fit to data.

Machine Learning is Reproducible and Scalable

Machine learning is inherently reproducible if the methodology is clearly described and the data is open and available. Reproducibility relies strongly on researchers to provide the exact data used as training data (to "teach" the learner) and those used to test the model. Any specific initialization is also required.

Because machine learning techniques are repeatable, they are also scalable. As new data is collected, it can be integrated into the machine learning routine to develop a new, more optimal predictor. Not all new data is equally relevant, and Bowden et al. (2012) present a technique to determine if the new data extends the range of the predictor.

Caveats and Open Problems

Several issues remain when using machine learning. Predictors can become overfit if too much data is shown to the learner, or the optimization routine is performed without bounds. Proper time to halt an optimization, and other relevant methods to avoid overfitting, are topics of active research (e.g., Maier and Dandy, 2000; Schmidt and Lipson, 2009; O'Neill et al., 2010; Tuite et al., 2011). Users should invest energy and time to mine the literature for these techniques.

Often data used to train the model is not selected optimally. We have previously advocated a deliberate sampling strategy to select data from the entire range of phase space available (e.g., Goldstein and Coco, 2014; Tinoco et al., 2015), as have others (e.g., Bowden et al., 2002; May et al., 2010; Wu et al., 2013). Predictors tend to be less overfit and more optimal when sampling is a considered process. This step adds extra work (especially to thoroughly document the process for repeatability), but we believe it is needed to develop the most optimal predictor.

Operators of machine learning algorithms should be experts in the data being examined. Nonphysical predictors can often appear as a result of regular usage of machine learning, data and/or computational errors. These erroneous results must be understood and manually discarded. We acknowledge that this adds a level of subjectivity to the analysis, but this subjectivity is also present in traditional empirical techniques (e.g., why did a researcher choose to fit the data using one function vs. another?). As a result, thorough examination of the physical correctness of the predictor should be performed, and expert knowledge should be exercised before machine learning results are accepted as correct and inserted into a hybrid model.

When combining machine learning components with physics-based model components users should be wary of the general structure of the predictor, and the potential for competing or mismatched nonlinearities in model components. We have personally encountered the mismatch between machine learning derived and theoretical components in a numerical model (Goldstein et al., 2014). This mismatch initially restricted our ability to understand sensitivity over a broad range of parameter values. We stress that it is always critical to understand and investigate how model components will interact.

Conclusion

Models constructed to study Earth surface processes are often intended to study large-scale, long-term phenomena. Little data may exist to parameterize long time-scale processes. However, ample data often exists for smaller space- and time- scale processes. Earth surface models should leverage all available data to build empirical parameterizations by adopting a hybrid approach.

Machine learning tools represent our best ability to process empirical data in a reproducible way. Best practices (explicit mentions of data selection and learner initialization) allows for reproducible results. The process of developing these predictors also explicitly highlights known gaps in knowledge. We believe these benefits should motivate the widespread adoption of hybrid models that combine machine learning approaches with physics-based models.

Acknowledgments

We thank R. O. Tinoco, A.B. Murray, P. W. Limber, K. Ells, and Z. Li for many enjoyable and insightful discussions, the editors for proposing this "research topic," and two reviewers for their constructive comments.

References

Aguilera, P. A., Fernández, A., Fernández, R., Rumí, R., and Salmerón, A. (2011). Bayesian networks in environmental modelling. *Environ. Model. Softw.* 26, 1376–1388. doi: 10.1016/j.envsoft.2011.06.004

Babovic, V., Canizares, R., Jensen, H. R., and Klinting, A. (2001). Neural networks as routine for error updating of numerical models. *J. Hydraul. Eng. ASCE* 127, 181–193. doi: 10.1061/(ASCE)0733-9429(2001)127:3(181)

Bowden, G. J., Maier, H. R., and Dandy, G. C. (2002). Optimal division of data for neural network models in water resources applications. *Water Resour. Res.* 38, 1010. doi: 10.1029/2001WR000266

Bowden, G. J., Maier, H. R., and Dandy, G. C. (2012). Real-time deployment of artificial neural network forecasting models: understanding the range of applicability. *Water Resour. Res.* 48, W10549. doi: 10.1029/2012WR011984

Bridgman, P. W. (1922). *Dimensional Analysis*. New Haven: Yale University Press.

Corzo, G. A., Solomatine, D. P., Hidayat, de Wit, M., Werner, M., Uhlenbrook, S., et al. (2009). Combining semi-distributed process-based and data-driven models in flow simulation: a case study of the Meuse river basin. *Hydrol. Earth Syst. Sci.* 13, 1619–1634. doi: 10.5194/hess-13-1619-2009

Crutchfield, J. P. (2014). The dreams of theory. *WIREs Comput. Stat.* 6, 75–79. doi: 10.1002/wics.1290

Goldstein, E. B., and Coco, G. (2014). A machine learning approach for the prediction of settling velocity. *Water Resour. Res.* 50, 3595–3601. doi: 10.1002/2013WR015116

Goldstein, E. B., Coco, G., and Murray, A. B. (2013). Prediction of wave ripple characteristics using genetic programming. *Cont. Shelf Res.* 71, 1–15, doi: 10.1016/j.csr.2013.09.020

Goldstein, E. B., Coco, G., Murray, A. B., and Green, M. O. (2014). Data driven components in a model of inner shelf sorted bedforms: a new hybrid model. *Earth Surf. Dynam.* 2, 67–82, doi: 10.5194/esurf-2-67-2014

Hall, J. W. (2004). Comment on 'Of data and models'. *J. Hydroinform.* 6, 75–77.

Jain, A., and Srinivasulu, S. (2004). Development of effective and efficient rainfall-runoff models using integration of deterministic, real-coded genetic algorithms and artificial neural network techniques. *Water Resour. Res.* 40, W04302. doi: 10.1029/2003WR002355

Knaapen, M. A. F., and Hulscher, S. J. M. H. (2002). Regeneration of sand waves after dredging *Coast. Eng.* 46, 277–289. doi: 10.1016/S0378-3839(02)00090-X

Knaapen, M. A. F., and Hulscher, S. J. M. H. (2003). Use of a genetic algorithm to improve predictions of alternate bar dynamics. *Water Resour. Res.* 39, 1231. doi:10.1029/2002WR001793

Krasnopolsky, V. M., and Fox-Rabinovitz, M. S. (2006). A new synergetic paradigm in environmental numerical modeling: hybrid models combining deterministic and machine learning components. *Ecol. Model.* 191, 5–18. doi: 10.1016/j.ecolmodel.2005.08.009

Limber, P. W., Murray, A. B., Adams, P. N., and Goldstein, E. B. (2014). Unraveling the dynamics that scale cross-shore headland amplitude on rocky coastlines, Part 1: model development. *JGR Earth Surf.* 119, 854–873. doi: 10.1002/2013JF002950

Maier, H. R., and Dandy, G. C. (2000). Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications. *Environ. Model. Softw.* 15, 101–124. doi: 10.1016/S1364-8152(99)00007-9

May, R. J., Maier, H. R., and Dandy, G. C. (2010). Data splitting for artificial neural networks using SOM-based stratified sampling. *Neural Netw.* 23, 283–294. doi: 10.1016/j.neunet.2009.11.009

Murray, A. B. (2003). "Contrasting the goals, strategies, and predictions associated with simplified numerical models and detailed simulations", in *Prediction in Geomorphology*, eds R. M. Iverson and P. R. Wilcock (Washington, DC: AGU, AGU Geophysical Monograph 135), 151–165.

O'Neill, M., Vanneschi, L., Gustafson, S., and Banzhaf, W. (2010). Open issues in genetic programming. *Genet. Program. Evol. M.* 11, 339–363. doi: 10.1007/s10710-010-9113-2

Oehler, F., Coco, G., Green, M. O., and Bryan, K. R. (2012). A data-driven approach to predict suspended-sediment reference concentration under non-breaking waves. *Cont. Shelf Res.* 46, 96–106. doi: 10.1016/j.csr.2011.01.015

Pape, L., Ruessink, B. G., Wiering, M. A., and Turner, I. L. (2007). Re- current neural network modeling of nearshore sandbar behavior, *Neural Netw.* 20, 509–518. doi: 10.1016/j.neunet.2007.04.007

Ruessink, B. G. (2005). Calibration of nearshore process models: application of a hybrid genetic algorithm. *J. Hydroinform.* 7, 135– 149.

Schmidt, M., and Lipson, H. (2009). Distilling free-form natural laws from experimental data. *Science* 324, 81–85. doi: 10.1126/science.1165893

Snelder, T. H., Lamouroux, N., Leathwick, J. R., Pella, H., Sauquet, E., and Shankar, U. (2009). Predictive mapping of the natural flow regimes of France. *J. Hydrol.* 373, 57–67. doi: 10.1016/j.jhydrol.2009.04.011

Tinoco, R. O., Goldstein, E. B., and Coco, G. (2015). A data-driven approach to develop physically sound predictors: application to depth-averaged velocities on flows through submerged arrays of rigid cylinders. *Water Resour. Res.* 51, 1247–1263. doi: 10.1002/2014WR016380

Tuite, C., Agapitos, A., O'Neill, M., and Brabazon, A. (2011). "Tackling Overfitting in Evolutionary-Driven Financial Model Induction," in *Natural Computing in Computational Finance*, eds A. Brabazon, M. O'Neill, and D. Maringer (Springer, Heidelberg), 141–161.

van Maanen, B., Coco, G., Bryan, K. R., and Ruessink, B. G. (2010). The use of artificial neural networks to analyze and predict alongshore sediment transport. *Nonlinear Process. Geophys.* 17, 395–404. doi: 10.5194/npg-17-395-2010

Wu, W., May, R. J., Maier, H. R., and Dandy, G. G. (2013). A benchmarking approach for comparing data splitting methods for modeling water resources parameters using artificial neural networks. *Water Resour. Res.* 49, 7598–7614. doi: 10.1002/2012WR012713

Yates, M. L., and Le Cozannet, G. (2012). Brief communication 'Evaluating European coastal evolution using Bayesian networks', *Nat. Hazards Earth Syst. Sci.* 12, 1173–1177. doi: 10.5194/nhess-12-1173-2012

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.