EGU24-11880, updated on 20 May 2024
https://doi.org/10.5194/egusphere-egu24-11880
EGU General Assembly 2024

Understanding geoscientific system behaviour from machine learning surrogates

Oriol Pomarol Moya[1], Derek Karssenberg[1], Walter Immerzeel[1], Madlene Nussbaum[1], and Siamak Mehrkanoon[2]

[1]Department of Physical Geography, Faculty of Geosciences, Utrecht University, Utrecht, The Netherlands
[2]Department of Information and Computing Sciences, Faculty of Science, Utrecht University, Utrecht, The Netherlands

Machine learning (ML) models have become popular in the Earth Sciences for improving predictions based on observations. Beyond pure prediction, though, ML has a large potential to create surrogates that emulate complex numerical simulation models, considerably reducing run time, hence facilitating their analysis.

The behaviour of eco-geomorphological systems is often examined using minimal models, simple equation-based expressions derived from expert knowledge. From them, one can identify complex system characteristics such as equilibria, tipping points, and transients. However, model formulation is largely subjective, thus disputable. Here, we propose an alternative approach where a ML surrogate of a high-fidelity numerical model is used instead, conserving suitability for analysis while incorporating the higher-order physics of its parent model. The complexities of developing such an ML surrogate for understanding the co-evolution of vegetation, hydrology, and geomorphology on a geological time scale are presented, highlighting the potential of this approach to capture novel, data-driven scientific insights.

To obtain the surrogate, the ML models were trained on a data set simulating a coupled hydrological-vegetation-soil system. The rate of change of the two variables describing the system, soil depth and biomass, was used as output, taking their value at the previous time step and the pre-defined grazing pressure as inputs. Two popular ML methods, random forest (RF) and fully connected neural network (NN), were used. As proof of concept and to configure the model setup, we first trained the ML models on the output of the minimal model described in [1], comparing the ML responses at gridded inputs with the derivative values predicted by the minimal model. While RF required less tuning to achieve competitive results, a relative root mean squared error (rRMSE) of 5.8% and 0.04% for biomass and soil depth respectively, NN produced better-behaved outcome, reaching a rRMSE of 2.2% and 0.01%. Using the same setup, the ML surrogates were trained on a high-resolution numerical model describing the same system. The study of the response from this surrogate provided a more accurate description of the dynamics and equilibria of the hillslope ecosystem, depicting, for example, a much more complex process of hillslope desertification than captured by the minimal model.

It is thus concluded that the use of ML models instead of expert-based minimal models may lead

to considerably different findings, where ML models have the advantage that they directly rely on system functioning embedded in their parent numerical simulation model.