Geophysical Research Abstracts
Vol. 17, EGU2015-2340, 2015
EGU General Assembly 2015

An entropy-based input variable selection approach to identify equally informative subsets for data-driven hydrological models

Gulsah Karakaya (1), Riccardo Taormina (1,2), Stefano Galelli (1), and Selin Damla Ahipasaoglu (1)

(1) Pillar of Engineering Systems Design, Singapore University of Technology and Design, Singapore, (2) Department of Civil and Environmental Engineering, Hong Kong Polytechnic University, Hong Kong

Input Variable Selection (IVS) is an essential step in hydrological modelling problems, since it allows determining the optimal subset of input variables from a large set of candidates to characterize a preselected output. Interestingly, most of the existing IVS algorithms select a single subset, or, at most, one subset of input variables for each cardinality level, thus overlooking the fact that, for a given cardinality, there can be several subsets with similar information content. In this study, we develop a novel IVS approach specifically conceived to account for this issue. The approach is based on the formulation of a four-objective optimization problem that aims at minimizing the number of selected variables and maximizing the prediction accuracy of a data-driven model, while optimizing two entropy-based measures of relevance and redundancy. The redundancy measure ensures that the cross-dependence between the variables in a subset is minimized, while the relevance measure guarantees that the information content of each subset is maximized. In addition to the capability of selecting equally informative subsets, the approach is characterized by two other properties, namely 1) the capability of handling nonlinear interactions between the candidate input variables and preselected output, and 2) computational efficiency. These properties are guaranteed by the adoption of Extreme Learning Machine and Borg MOEA as data-driven model and heuristic optimization procedure, respectively. The approach is demonstrated on a long-term streamflow prediction problem, with the input dataset including both hydro-meteorological variables and climate indices representing dominant modes of climate variability. Results show that the availability of several equally informative subsets allows 1) determining the relative importance of each candidate input, thus supporting the understanding of the underlying physical processes, and 2) finding a better trade-off between multiple measures of prediction accuracy (e.g., RMSE, NSE, MAE) and the desired number of input variables.