Geophysical Research Abstracts
Vol. 13, EGU2011-9736, 2011
EGU General Assembly 2011

The pros and cons of different strategies to avoid overfitting: revisiting neural network suspended sediment modelling of two small watersheds in Puerto Rico

Robert J. Abrahart
School of Geography, University of Nottingham, Nottingham, United Kingdom (bob.abrahart@nottingham.ac.uk)

In a perfect world, hydrological modelling datasets would contain equal representations of all possible situations and posses a well balanced spread. No major gaps or outliers would exist. No substantial interpolations or extrapolations would be required. This, however, is not the case — especially for suspended sediment datasets of tropical upland river systems such as ones found on the Caribbean island of Puerto Rico. Natural conditions on that island result in sediment discharges which are predominantly episodic and intense; delivering an observed record with: (i) outliers which must be retained; and (ii) voids that will need to be bridged. Neural network enthusiasts have engaged in the hydrological challenge to model such systems, without rainfall information, and the critical issue with regard to developing a sound solution is one of 'generalisation': how well will the preferred solution deliver predictions for patterns which were not in the training dataset? The art and science of building a finished product in such instances is to a large part based on rules of thumb, trial and error exploration, and out-of-sample assessment using an 'unseen dataset'. Of critical importance is the specific nature of each subset involved in the modelling process, since poor, unequal or unrepresentative content: (i) could deliver biased statistics with regard to appraisal metrics computed on an idiosyncratic test dataset and (ii) is not conducive to delivering a set of universal findings, perhaps leaving unanswered important questions, such as to what extent are published results unique or exceptional, or could be relied upon, as definitive statements of "best method". This issue is further complicated by that fact that each solution is in most cases not presented in published papers as anything other than a set of (i) tabulated test statistics and/or (ii) observed-vs-predicted output scatterplots — related to a black-box model. The need to avoid under- and over-fitting is nevertheless a recognised neural network hydrological modelling issue (e.g. Maier & Dandy, 2000; Maier et al., 2010): albeit that the explicit nature of potential impacts on each final model is perhaps not fully appreciated or understood. Techniques are available to support the development of an optimum trade-off but few hydrological intercomparison studies have been conducted e.g. Giustolisi & Laucelli (2005). Two main issues are involved: (i) quantity and quality of records to be modelled; (ii) number and size of weights requiring adjustment. The power of (i) Model Selection, (ii) Jittering and (iii) Bayesian Regularization to overcome substantial modelling issues are compared and contrasted at two previously studied gauging stations and some practical guidelines suggested.