Geophysical Research Abstracts
Vol. 12, EGU2010-2543-1, 2010
EGU General Assembly 2010

Optimum Input Selection For Data Driven Modeling. Mirage or Reality?

Amin Elshorbagy (1) and Zohreh Izadifar (2)

(1) University of Saskatchewan, Civil & Geological Engineering, Saskatoon, Canada (amin.elshorbagy@usask.ca, 1-306-966-5427), (2) University of Saskatchewan, Civil & Geological Engineering, Saskatoon, Canada

The implementation and evaluation of data driven modeling have been facing multiple challenges over the past two decades. Some of these challenges (e.g., the choice of the modeling technique, predictive uncertainty, assessment of the model performance) are shared with physical and conceptual modeling. But other challenges, such as selection of optimum inputs and lack of conceptual/physical justification, are unique to data driven modeling. In this study, the research question of “is it possible to select the optimum inputs of data driven models a priori?” is addressed. This question, in various forms, received considerable attention in data driven modeling literature, and ANN literature in particular. The case study of estimating the hourly actual evapotranspiration (AET) using multiple meteorological variables (air temperature, net radiation, ground temperature, relative humidity, and wind speed) was used. The correlation between the various inputs and the output, and the input contribution to the output predictability were investigated using simple correlation matrix, step-wise regression, mutual and partial mutual information, Gamma test, and wavelet analysis in an attempt to identify the optimum inputs a priori. Neural networks, genetic programming, and multiple linear regression techniques were also used to develop the AET models with trial and error to select the best inputs. The performances of the models were assessed based on their accuracy, identifiability, uncertainty, and parsimony. It was found that, even though input selection methods might provide partial help and information regarding the most relevant inputs, it is impossible or impractical to fully identify the optimum inputs without trial and error with the modeling techniques themselves. The complexity of the interrelationships among the hydrological inputs makes it impossible to identify the relative merits of the individual inputs versus the compound effects of endless possible combinations of inputs. Even though the partial possibility of input selection a priori cannot be denied, the mirage cannot be refuted.