BEYOND CORRELATION: THE SEARCH FOR CAUSAL RELATIONSHIPS BETWEEN FLOW PERCENTILES AND WATERSHED VARIABLES

Herbert Ssegane[1], E.W. Tollner[1], Yusuf Mohamoud[2], T. C. Rasmussen[3], and John Dowd[4]

AUTHORS: [1]Department of Biological and Agricultural Engineering, University of Georgia, Athens, GA; [2]National Exposure Research Laboratory, U.S. Environmental Protection Agency, Athens, GA; [3]Warnell School of Forestry and Natural Resources, University of Georgia, Athens, GA; [4]Department of Geology, University of Georgia, Athens, GA.

REFERENCE: *Proceedings of the 2011 Georgia Water Resources Conference*, held April 11–13, 2011, at the University of Georgia.

Abstract. The study explored use of causal feature selection algorithms to select dominant watershed variables that drive high, medium, and low flows. A two step approach was implemented. The first step minimized variable redundancy by examining variable relevance, variable redundancy, and conditional relevance of variable pairs whose correlation was greater than 0.9. The second step used six algorithms that seek to reconstruct a Bayesian network structure around a target variable for each flow percentile. Nineteen (19) flow percentiles were used to characterize high, medium, and low flow conditions of 26 Piedmont watersheds in the Mid-Atlantic. The algorithms included: (1) Grow-Shrink (GS); (2) interleaved-Incremental Association Markov Boundary (interIAMB) (3) Incremental Association Markov Boundary with Peter-Clark (IAMBnPC); (4) Local Causal Discovery (LCD2); (5) HITON-PC; and (6) HITON-MB. A new method was developed to quantify the reliability of each algorithm and its performance was compared to existing reliability methods. The effect of the initial number of variables on the final variable set selected by each algorithm was tested. Fusion of the algorithms was used to determine the overall dominant features for each flow percentile.