Geophysical Research Abstracts
Vol. 19, EGU2017-11803, 2017
EGU General Assembly 2017

A generalised approach for identifying influential data in hydrological modelling

David Wright (1), Mark Thyer (1), Seth Westra (1), Benjamin Renard (2), and David McInerney (1)

(1) The University of Adelaide, Adelaide, Australia (david.p.wright@adelaide.edu.au), (2) Irstea, UR HHLY Hydrology-Hydraulics, Villeurbanne Cedex, France

Influence diagnostics identify data points that have a disproportionate impact on model calibration, and are therefore useful to identify possible erroneous data points or scrutinise the sensitivity of the model results to a small portion of the overall calibration dataset. Case-deletion Cook's distance calculates influence; however, it has a large computational demand due to the requirement for recalibration of the model parameters for every data point in the calibration data. Regression based Cook's distance provides an approximation of case-deletion Cook's distance by combining two regression components for each observed data point: 1) the leverage which is used to assess the potential importance of individual observations, and 2) the standardised residuals. By combining these two components the regression based Cook's distance requires only a relatively small number of additional runs and is therefore an attractive alternative to the computationally demanding case-deletion Cook's distance.

The objective of this study is to develop generalised regression based influence diagnostics that can be applied across a wide range models and objective functions in a computationally efficient manner. This overcomes the limitations of the current suite of influence diagnostics. For example, the regression based Cook's distance has two assumptions that are not satisfied in hydrological modelling: 1) the hydrological model is linear; 2) the objective function applied is limited to standard least squares. In addition, although the case-deletion diagnostics overcome these assumptions, they require high performance computing and therefore are not computationally feasible for most hydrological model applications.

In this study we generalise regression based Cook's distance to be applied beyond linear models and to the vast majority of objective functions currently applied in hydrological model calibration. The improvements from the new formulation are then examined by comparing generalised Cook's distance to the computationally demanding case-deletion influence metric (used as a baseline measure of the performance of each influence diagnostic). We consider three case studies: (1) a series of synthetic regression models with varying nonlinear model response and heteroscedasticity in residual error; 2) a conceptual hydrological model (GR4J), and 3) a rating curve incorporating discharge uncertainty and Bayesian parameter priors.

The generalisation of the regression based Cook's distance allows for computationally cheap influence analysis across the vast majority of hydrological model structures and objective functions. The inclusion of highly influential data in model calibration can have a substantial impact on the predicted maximum and mean flows. Due to the large amount of insight that influence diagnostics provide combined with their computationally cheap nature we recommend that influence analysis is undertaken by hydrological practitioners as a step towards more robust model calibration.