Geophysical Research Abstracts
Vol. 16, EGU2014-2211, 2014
EGU General Assembly 2014

How Reliable is Bayesian Model Averaging Under Noisy Data? Statistical Assessment and Implications for Robust Model Selection

Anneli Schöniger (1), Thomas Wöhling (2), and Wolfgang Nowak (3)

(1) Center for Applied Geoscience, Universität Tübingen, Tübingen, Germany (anneli.schoeniger@uni-tuebingen.de), (2) Water & Earth System Science Competence Cluster (WESS), Universität Tübingen, Tübingen, Germany (thomas.woehling@uni-tuebingen.de), (3) Institute for Modelling Hydraulic and Environmental Systems, LH2/SimTech, Universität Stuttgart, Stuttgart, Germany (wolfgang.nowak@iws.uni-stuttgart.de)

Bayesian model averaging ranks the predictive capabilities of alternative conceptual models based on Bayes' theorem. The individual models are weighted with their posterior probability to be the best one in the considered set of models. Finally, their predictions are combined into a robust weighted average and the predictive uncertainty can be quantified. This rigorous procedure does, however, not yet account for possible instabilities due to measurement noise in the calibration data set. This is a major drawback, since posterior model weights may suffer a lack of robustness related to the uncertainty in noisy data, which may compromise the reliability of model ranking.

We present a new statistical concept to account for measurement noise as source of uncertainty for the weights in Bayesian model averaging. Our suggested upgrade reflects the limited information content of data for the purpose of model selection. It allows us to assess the significance of the determined posterior model weights, the confidence in model selection, and the accuracy of the quantified predictive uncertainty.

Our approach rests on a brute-force Monte Carlo framework. We determine the robustness of model weights against measurement noise by repeatedly perturbing the observed data with random realizations of measurement error. Then, we analyze the induced variability in posterior model weights and introduce this “weighting variance” as an additional term into the overall prediction uncertainty analysis scheme. We further determine the theoretical upper limit in performance of the model set which is imposed by measurement noise. As an extension to the merely relative model ranking, this analysis provides a measure of absolute model performance. To finally decide, whether better data or longer time series are needed to ensure a robust basis for model selection, we re-sample the measurement time series and assess the convergence of model weights for increasing time series length.

We illustrate our suggested approach with an application to model selection between different soil-plant models following up on a study by Wöhling et al. (2013). Results show that measurement noise compromises the reliability of model ranking and causes a significant amount of weighting uncertainty, if the calibration data time series is not long enough to compensate for its noisiness. This additional contribution to the overall predictive uncertainty is neglected without our approach. Thus, we strongly advertise to include our suggested upgrade in the Bayesian model averaging routine.