Geophysical Research Abstracts
Vol. 17, EGU2015-2084-2, 2015
EGU General Assembly 2015

On the Various (Good and Bad) Ways to Evaluate Bayesian Model Weights

Anneli Schöniger (1), Thomas Wöhling (2,3), Luis Samaniego (4), and Wolfgang Nowak (5)

(1) University of Tübingen, Center for Applied Geoscience, Tübingen, Germany (anneli.schoeniger@uni-tuebingen.de), (2) Water & Earth System Science Competence Cluster (WESS), University of Tübingen, Tübingen, Germany, (3) Lincoln Environmental Research, Lincoln Agritech, Hamilton, New Zealand, (4) Helmholtz Centre for Environmental Research UFZ Leipzig, Leipzig, Germany, (5) Institute for Modelling Hydraulic and Environmental Systems, LS3/SimTech, University of Stuttgart, Stuttgart, Germany

Bayesian model averaging (BMA) is a rigorous statistical framework to rank a set of plausible, competing models according to their plausibility, i.e. according to their fit to available data and their complexity. With the resulting posterior model weights, model predictions are averaged to obtain a robust estimate along with uncertainty intervals. The evaluation of these model weights requires determining Bayesian model evidence (BME), which is the average likelihood of a model.

Technically, BME is an integral of likelihood over each model's parameter space. The evaluation of this integral is highly challenging, because its dimensionality increases with the number of model parameters. In general, the following three classes of techniques are available to evaluate BME, each with its own challenges and limitations:

1) Exact analytical solutions are fast, but rarely available due to their strongly restricting assumptions.
2) Numerical integration is accurate, but quickly becomes computationally unfeasible.
3) Approximations known as information criteria (ICs, e.g. AIC, BIC, KIC) potentially yield misleading results that do not reflect the true Bayesian ranking.

We conduct a systematic comparison of numerical integration and IC approximation with regard to theoretical differences, computational effort, and accuracy of the estimated BME value. For the latter, we use (1) a simplified synthetic example with an exact analytical solution as a first-time validation against a true solution, and (2) a real-world application to hydrological model selection between two variants of the mHM (distributed mesoscale hydrologic model, Samaniego et al. 2010), where we use a brute-force Monte-Carlo method as benchmark solution.

Our results reveal that the error in BME approximation varies substantially over the available techniques. Consequently, the choice of approximation method severely impacts the final model ranking and model-averaged predictions. In specific, the ICs show a huge range of potential errors. In the hydrological test case, they are not able to identify the most adequate model according to BMA theory. The KIC evaluated at the maximum a posteriori parameter estimate (KIC@MAP) performs best, but in general none of the ICs is satisfying for non-linear model problems. Hence, there is still no alternative to numerical evaluation to obtain reliable model weights.