Hydrol. Earth Syst. Sci. Discuss., 3, S206–S208, 2006
www.copernicus.org/EGU/hess/hessd/3/S206/
European Geosciences Union

Hydrology and
Earth System
Sciences
Discussions

Interactive comment on “Optimising training data for ANNs with Genetic Algorithms” *by* R. G. Kamp and H. H. G. Savenije

D. Solomatine (Editor)

d.solomatine@unesco-ihe.org

Received and published: 9 May 2006

The paper covers an important topic and I agree with the reviewers supporting this. I am in a good position when the reviewers submitted their reviews, so my comments add to theirs. It is recommended to address the following.

The problem of proper composition of the training dataset is not new and was addressed by most of the researchers dealing with ANN and other data-driven models. The authors are also invited, for example, to read and provide the references to the paper by Bowden et al (2002) Optimal division of data for neural network models in water resources applications (WRR 38(2)), where the problem under consideration was considered as well. Other authors that were not posing this problem as an optimization problem, were, however addressing it as well trying to ensure that the training set is "representative". It can be said that now it is widely recognised that the train-

Full Screen / Esc

Print Version

Interactive Discussion

Discussion Paper

ing, cross-validation and test sets have to be "statistically similar" - this would ensure proper modelling. This paper presents one of the ways ensuring higher accuracy of the model. It would be recommended to compare the method suggested with other methods, e.g. with a widely used random selection of the data vectors for the three sets used in machine learning. Presentation of PDFs of the resulting sets may help.

Abstract could have been formulated better. For example, the concluding sentence "The optimised training set resulted in significant better training data. " needs reformulation.

P. 1: The statement "The training data should be representative for the simulation period, otherwise extrapolation of model simulation is difficult" is not fully correct. Extrapolation is always difficult is not impossible. Representative data helps interpolation.

"Using existing data sets may seem a good alternative.": please specify alternative to what.

P.2: "From the available dataset usually a subset for training is selected without a predefined selection procedure." This is not really true since most researchers are careful enough to build representative training data sets (often however being suboptimal).

Sentence "The question is whether it is possible to optimise the training data using a GA" does not really reflect the essence of the paper since the answer to this question is obvious: yes it is. The question is in fact how to encode the data, set up the experiment, etc. These aspects, unfortunately, are not covered in detail. (Note that the formulation could be improved: it is not the training set that is optimized, but the model that is trained on it.)

I would agree with the reviewer 2 that the description of the main procedure is not really clear. It is also not clear if cross-validation sets are used. If not, is there a danger that the resulting model will overfit?

It is recommended to provide some justification why emulation of the Duflow model was needed.

The paper is too short to cover all the aspects of the experiments, and the clarity of the narration suffers. It is recommended to extend the paper and to provide more accurate explanations, formulations and clear justifications.

English could be improved. In a number of places, for example, plural is formed by adding "'s" instead of just "s". "in a certain extent" –> to a certain extent" (P. 2), etc.

It is recommended to address all the comments of the reviewers, and other readers, very valuable indeed.

Interactive comment on Hydrology and Earth System Sciences Discussions, 3, 285, 2006.