Hydrol. Earth Syst. Sci. Discuss., 6, C1082–C1086, 2009
www.hydrol-earth-syst-sci-discuss.net/6/C1082/2009/

Hydrology and Earth System Sciences Discussions

Interactive comment on "*HESS Opinions* "Crash tests for a standardized evaluation of hydrological models"" *by* V. Andréassian et al.

R. T. Clarke

clarke@iph.ufrgs.br

Received and published: 5 June 2009

The authors do well to remind us that model evaluation involves comparisons (lines 14-16 on page 3678: "model evaluation is only meaningful in a comparative framework (a model can only be ranked good in comparison with alternative models"). So if we are comparing models, what we are doing in effect is an experiment: an experiment in which different models are tested on data-sets (of precipitation, runoff, and possibly other variables) from different catchments. In each such test (i.e., in each combination of model and catchment data-set), measures of model performance are recorded, such as the Nash-Sutcliffe criterion shown in the authors' Fig.1. Based on these measures, conclusions must be drawn about relative model performance – which model,

if any, performs better overall than the others, or if no one model always performs better, which models perform better under different conditions. However before the "comparison with alternative models" (i.e., the comparative experiment) is initiated, it is important at the planning stage to ensure that the design of the experiment allows valid conclusions to be drawn. I have tried to address the question of experimental design in the comparison of hydrological models in two recent papers (Clarke, 2008a, b) and I am grateful to Vit Klemeš for mentioning my interest.

The design of comparative experiments has been developed over many decades since the pioneering work by R A Fisher and F Yates, in the context of agricultural field trials, in the early part of the twentieth century, but the possibility of applying this vast body of knowledge and experience to the comparison of hydrological models appears to have been largely overlooked. In their paper, the authors use the analogy of crash tests on motor vehicles; I should like to offer another analogy, which I believe brings out the similarity between procedures for the comparison of hydrological models on the one hand and, on the other hand, the procedures used in agriculture to compare crop varieties, in medicine to compare treatments, and in industry to compare methods of production. Indeed, the design of "crash tests" is just the kind of application that might be found in many well-known texts on industrial experimentation.

Suppose a plant breeder has the task of selecting varieties of wheat to be recommended for widespread planting according to normal cultivation practices. She has seeds from a number of potentially useful varieties, and very much likes the look of one of them which yielded well when grown in a small pot of very fertile soil under controlled greenhouse conditions. When seeds from the different varieties were grown in greenhouse pots, the plant breeder recorded, for each variety, variables such as weight of seeds, number of seeds per ear of wheat, the number of stems (tillers) bearing ears, and so on; these variables are analogous, in the context of hydrological modeling, to the Nash-Sutcliffe criterion, the mean of absolute differences between observed and modeled discharges, and the many other criteria that are used. Also, the particular variety

which the plant breeder likes, and the greenhouse conditions under which she found that it performed so well, are analogous to the hydrologist's "pet" model, developed using "good" data, perhaps from a small experimental catchment that is densely instrumented with rain-gauges, with well-established rating curve and carefully monitored gauging structure – conditions that may be quite unlike the conditions in catchments encountered in hydrological practice, just as greenhouse conditions bear no relation to conditions on farms where the plant-breeder's preferred wheat variety would be grown, if it were to be recommended for field cultivation.

So the next stage in the plant-breeder's selection procedure would be to plan some comparative experiments in fields of farmers who are known to use good practices for weed control, fertilizer application, and other agronomic practices. In these experiments, the preferred variety would be compared (in terms of yield, and its components) with varieties that are currently grown. Each such variety would be replicated: that is, allocated at random to a number of experimental plots, according to some structured design. And since weather conditions and associated pest infestations vary from year to year, the experiments in which the varieties are compared would need to be repeated in each of several years. This would again be followed by another set of comparative experiments, on a random selection of farms, in some of which agronomic practices may be far from what is recommended.

Without flogging the analogy to death, it can be seen that the essential characteristics of the procedure for comparing hydrological models, like those of varietal selection, are (i) replication, by which each model is tested on several data sets (the fact that each model is tested on a number of data sets, from different catchments, ensures that the uncertainty in the measure(s) of model performance can be calculated) ; (ii) randomization, in which models are allocated at random to the data-sets on which they are tested (to eliminate the possibility of bias, whether conscious or unconscious, in allocating a "pet" model to more reliable data-sets). Specification of the degree of replication, and the random allocation procedure, jointly constitute the experimental design

for the comparison. Similarly, comparisons between the models should be repeated under different conditions of climate, geology and topography (in the jargon of experimental design, there may be an "interaction" between models and climate conditions, with some models performing better, or worse, than others under different climates, or in different geologies). As in agronomic experiments, the degree of replication (i.e., the number of catchments whose data are to be used) will depend on (i) the magnitude of differences in model performance that must be detected in the experiment, and (ii) the variability between catchments (the greater this variability, the larger the number of catchments whose records will need to be used in the comparison.)

My interpretation of the authors' arguments in their paper – particularly Sections 2.1 ("Arguments in favour of catchment monographs"), 2.2 ("Arguments in favour of large data sets") and 2.3 ("Any arguments in favour of a hybrid approach?") is that they are really discussing the degree of replication to be used in the comparative experiment (randomization is not mentioned in the paper, although in a vehicle crash test, it would be essential to ensure that the vehicle(s) to be crashed are selected at random from the production line). Their Section 3 ("Large data sets and data quality") resonates, in the analogy of the plant-breeder, with the stage in which the comparison between varieties is extended to "ordinary" farms, (that is, a stage at which hydrological models are compared on the questionable data-sets commonly found in hydrological practice.)

To conclude, I welcome the paper for drawing attention to the need for careful thought about how hydrological models should be compared, and about how their performance should be evaluated. I believe that there is much benefit to be found through increased awareness of the enormous body of knowledge, already widely used in so many other fields of applied science, on the design and analysis of comparative experiments.

References

Clarke R T (2008a) A critique of present procedures used ot compare performance of rainfall-runoff models, J. Hydrology, Vol. 352, Issue 3-4, 379-387.

Clarke R T (2008b) Issues of experimental design for comparing the performance of hydrologic models, Water Resources Research, Vol. 44, Issue 1, Number W01409.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 6, 3669, 2009.