Comment on "On the calibration and verification of two-dimensional, distributed, Hortonian, continuous watershed models" by Sharika U. S. Senarath et al.

Jan Seibert[1]

Department for Environmental Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden

1. Introduction

In a recent paper *Senarath et al*. [2000] extended an event-based runoff model, the CASC2D model, to allow continuous simulations. This is a valuable approach because it allows avoiding the estimation of initial soil moisture values. They applied the modified model and tested its performance in various ways. Senarath et al. address important topics in hydrological modeling and put more effort into model testing than is done usually. Nevertheless, further discussion is warranted for some points, where a more careful treatment of model testing would have been possible. This comment seeks to highlight some of these issues to promote an open dialog on principles in model testing. These principles are relevant also for other types of hydrologic catchment models than the physically based, Hortonian model applied by Senarath et al., despite the fact that there are significant differences between the various types of models.

The criticism can be summarized in five points: (1) the use of data from the "independent" test periods in the calibration procedure; (2) an inadequate evaluation of model performance for subcatchments, namely, neglecting that catchment runoff depends on runoff from subcatchments and introducing a bias through inappropriate selection of considered events; (3) an incomplete interpretation of large relative model errors for small events, (4) missing the opportunity to judge model performance by comparing the new model with results obtained with other models (e.g., the original model) or simple methods (e.g., the rational method), instead of using the vague term "acceptable accurate"; and, (5) drawing conclusions that are not supported by the presented study. The following text elaborates on the different points.

2. Point 1: Use of Data From Test Period for Calibration

Senarath et al. found (p. 1502) a "large number of potential parameter sets" for the calibration period and found it difficult to select the optimal parameter set. This equifinality, that is, the phenomenon that equally good model simulations might be obtained in many different ways [*Beven*, 1993], is a common problem in hydrological modeling. Instead of addressing equifinality by allowing for different parameter sets using, for instance, the generalized likelihood uncertainty estimation (GLUE) framework [*Freer et al*., 1996], Senarath et al. chose their one optimal parameter set by testing all potential parameter sets with data sets from three other years. Using this procedure, they selected a parameter set which had a cost-function value of 0.0417 for the calibration period, although a better fit (i.e., a lower cost-function value of 0.0397) for the calibration period was obtained with a different parameter set (p. 1502). In other words, the information from the "independent" periods actually influenced the selection of the optimal parameter set. It might be reasonable to incorporate additional runoff data into the search for the best parameter set, but as soon as this is done, these data no longer provide an independent test period. The only fully independent test period for the model was thus the second part of the 1982 data (five events).

With a long period of available data a model can be tested in two ways: (1) calibration to the entire period (what is the best fit that can be achieved for the entire period?) or (2) calibration to part of the data, using the remaining data as an independent test (split-sample test: how good is the calibrated model for periods not used during calibration?). The strategy chosen by Senarath et al. falls between these two methods. It deals with the problem of finding a good fit for a longer period by searching only among parameter sets that gave a very good fit for a shorter period. This is certainly a less relevant question than that tackled by the split-sample test, which allows estimating model errors for periods without any runoff observations.

3. Point 2: Evaluating Model Performance for Subcatchments

Senarath et al. also tested their calibrated model against observed runoff at several stations within the catchment. Testing simulations against runoff data at points within a catchment is an important test for a distributed model. However, Senarath et al. claim (p. 1503) that "since the data from these gauging stations were not used in any way for model calibration, this comparison provides a true test of model validity at internal locations." This is not correct. When using runoff from subcatchments, one must consider that the runoff at the catchment outlet is related to the runoff at internal locations. Therefore information from the subbasins had been used implicitly in the calibration procedure. As the size of a subcatchment increases, the quality of the runoff simulations for this subcatchment inevitably approaches the quality of the simulations for the entire catchment, yet this does not tell us anything about internal model validity. The largest subcatchment used by Senarath et al. covers 84% of the entire catchment. Runoff from this subcatchment was a large portion of the catchment runoff, which was used for calibration. Therefore it is not surprising that the fit for this subcatchment is about as good as for the entire catchment. It is more important to note that the quality of the fit decreases for the smaller subcatchments, for

[1] On leave from Department of Forest Engineering, Oregon State University, Corvallis, Oregon, USA.

Paper number 2000WR000017.
0043-1397/01/2000WR000017$09.00

which there is less dependency with runoff at the outlet. This indicates that the spatially distributed simulations may not be very accurate.

Senarath et al. used only events with a peak discharge >0.5 $m^3\ s^{-1}$ in their analysis and used the same volume-based threshold value also for all subcatchments, regardless of catchment size. This introduced a bias that impedes comparison of model performance for the different subbasins. In the smaller catchments, small events were neglected, although their relative sizes, i.e., runoff per unit area, were comparable to those of events that were included in the larger catchments, and thus fewer events were analyzed. Analyzing all events selected at the outlet for all subcatchments, or using a per-unit-area value as threshold, would be more reasonable. Results for the smaller subcatchments can be expected to be poorer without this bias because of the tendency of relative errors to increase for smaller events.

Senarath et al. say (p. 1506) that their model "is capable of simulating peak discharges at internal catchment locations with reasonable degree of accuracy." However, there are significant deviations between simulations and observations, as can be seen from Figure 13 (p. 1508) of Senarath et al. It is important to observe that the deviations appear smaller because of the log scale. This highlights a problem of assessing model performance. The term "reasonable degree of accuracy" can be quite subjective. In the case where a model with parameter values that have been calibrated for one catchment is applied to another (sub)catchment, results often will be "acceptable." To truly assess model performance, it is important to compare one's results with results obtained in some other way, that is, to choose an appropriate benchmark.

Both simulated and observed runoff series from the catchment that has been used for calibration provide suitable benchmarks when scaled for differences in catchment area. Comparison of a simulation with the first benchmark will reveal the value of taking catchment-specific characteristics, such as the percentage or pattern of different land use classes, into account. If the model fails to be better than the first benchmark, the implementation of catchment-specific characteristics has to be reconsidered. The second benchmark evaluates the usefulness of a model approach compared with a very simple method to obtain runoff series at ungauged locations [e.g., *Seibert*, 1999]. The benefit of this benchmark is that it takes the observed variation between (sub)catchments into account. It helps in assessing whether the results are good because the model is good (and takes differences between the catchments into account) or because runoff dynamics are similar.

4. Point 3: Interpretation of Model Errors for Small Events

Senarath et al. found that the relative errors were larger for smaller events. They interpret (p. 1508) this as an effect of "diminishing influence of parameter uncertainty with increasing storm intensity." These large errors could also be interpreted in a way that questions the performance of the new continuous formulation of the model. Initial soil moisture conditions are most important for the simulation of small events. The poorer fit for smaller events may indicate that the new soil moisture routine, which provides the initial values, does not work that well after all. Senarath et al. did not discuss this possible interpretation. An analysis of this indirect indication would be valuable because Senarath et al. do not have any observations of soil moisture to assess the performance of their model extension directly.

5. Point 4: What is "Acceptable Accuracy"?

The simulated peak discharge differed by more than 40% from the corresponding observed value for 10 of the 25 events, and runoff volume errors were larger than 40% for 11 events. Do these results really show "that the continuous CASC2D formulation is capable of simulating Hortonian catchment dynamics with acceptable accuracy" (p. 1508)? The question is to define what we mean by acceptable accuracy. Given the substantial model errors, it is not obvious that the model is "good," especially considering the high-quality rainfall data.

In practical model applications an engineer can define the term acceptable accuracy with regard to the requirements of the problem, which has to be solved. In many cases an economic cost is associated with model errors. In scientific model applications the meaning of acceptable accuracy is less obvious. In this case acceptable accuracy should be defined with regard to what can be achieved given the specific catchment and the observed data. Comparison with another model or a simple method (e.g., the rational method) supports a more objective rating of model performance. In the case of Senarath et al., comparison with the original model (with estimated or calibrated values for initial state variables) would have helped to assess the performance of the new model and the value of the model extension.

6. Point 5: Unsupported Conclusions

Senarath et al. draw conclusions that are not supported by their study. They state (p. 1495), "Results show that calibration on a continuous basis significantly improves model performance for periods, or subcatchments, not used in the calibration and the likelihood of obtaining realistic simulations of spatially varied catchment dynamics." The model was tested only to some degree for independent periods or subcatchments, as discussed in section 2. Furthermore, the results were not compared to those from a calibration of the event-based version of the model; thus the benefits of the calibration on a continuous basis have not been demonstrated. Finally, there is little foundation for the assertion about simulations of spatially varied catchment dynamics. Internal variables, such as soil moisture, were not compared with observations, and the analysis of runoff simulations at internal points was deficient as discussed in section 3.

Senarath et al. also state (p. 1508), "The value of [the use of continuous soil moisture accounting] has been demonstrated with an unusually rigorous testing procedure . . . to verify the optimality of the calibrated parameter set." Again, the value of the continuous formulation was not demonstrated, since results were not compared with those of the event-based formulation. Although Senarath et al. tested their model in several ways, the standard for an "unusually rigorous testing procedure" should be higher. The model could have been tested using, for instance, a differential split-sample test [*Klemeš*, 1986] or data other than runoff. The "optimality" of the calibrated parameter set has not been investigated. One parameter set was calibrated on the basis of runoff at the catchment outlet, but it was not explored whether some other parameter set would have performed better in the different tests. Furthermore, it could have been investigated whether the same parameter set might be found as optimal when calibrating in different ways such as using a different objective function, using the various simulation periods in a different way, or considering runoff from the subcatchments directly. Most probably, the result of such tests would show that the optimal parameter set is very sensitive to the way it is determined, and thus it would falsify rather than verify the optimality of one single parameter set, especially because of the flatness of the optimal parameter space noted by Senarath et al.

7. Concluding Remarks

The study presented by Senarath et al. was, like many modeling studies, limited by the lack of data other than runoff. When tested only against runoff at the catchment outlet, i.e., "lumped" data that integrate over the catchment, distributed models can seldom be demonstrated to be superior to lumped or semidistributed models. Whenever a model is aimed at simulating more than just runoff, its capability to do so should be demonstrated. Much too often powerful tests are not performed "because there was no data available" [e.g., *Stagnitti et al*., 1992; *Yao et al*., 1996; *Krysanova et al*., 1998].

Admittedly, Senarath et al. put more effort into model testing than is done in many other studies. Consideration of the points discussed above would provide more powerful testing procedures, which may support, or disprove, some of the "unsupported" conclusions. To make progress in hydrological modeling, it is crucial that researchers and reviewers are aware of the importance of careful model testing. The use of appropriate benchmarks to compare and evaluate model results is important to justify conclusions about model performance.

References

Beven, K. J., Prophecy, reality and uncertainty in distributed hydrological modelling, *Adv. Water Resour*., *16*, 41–51, 1993.

Freer, J., K. J. Beven, and B. Ambroise, Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach, *Water Resour. Res*., *32*(7), 2161–2173, 1996.

Klemeš, V., Operational testing of hydrological simulation models, *Hydrol. Sci. J*., *31*, 13–24, 1986.

Krysanova, V., D.-I. Müller-Wohlfeil, and A. Becker, Development and test of a spatially distributed hydrological/water quality model for mesoscale watersheds, *Ecol. Modell*., *106*, 261–289, 1998.

Seibert, J., Regionalisation of parameters for a conceptual rainfall-runoff model, *Agric. For. Meteorol*., *98–99*, 279–293, 1999.

Senarath, S. U. S., F. L. Ogden, C. W. Downer, and H. O. Sharif, On the calibration and verification of two-dimensional, distributed, Hortonian, continuous watershed models, *Water Resour. Res*., *36*(6), 1495–1510, 2000.

Stagnitti, F., J.-Y. Parlange, T. S. Steenhuis, M. B. Parlange, and C. W. Rose, A mathematical model of hillslope and watershed discharge, *Water Resour. Res*., *28*(8), 2111–2122, 1992.

Yao, H., M. Hashino, and H. Yoshida, Modeling energy and water cycle in a forested headwater basin, *J. Hydrol*., *174*, 221–234, 1996.

J. Seibert, Department for Environmental Assessment, Swedish University of Agricultural Sciences, Box 7050, SE-75007 Uppsala, Sweden. (jan.seibert@ma.slu.se)

(Received October 13, 2000; revised August 29, 2001;
accepted August 29, 2001.)