Review of H.E.S.S. manuscript "Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology – Part 2: Application"

Authors: A. Elshorbagy, G. Corzo, S. Srinivasulu, and D. P. Solomatine

Comments: The authors made a comparison of several data driven modeling techniques over several different hydrological datasets. Although much effort was put into this endeavour, the final results were not very satisfactory, and I would recommend major revision.

1. The division of the work into two papers is not well justified. While it is not uncommon for authors to publish "concepts and methodology" as a separate paper from the results/applications, their "concepts and methodology" part usually presents a new theory or model, followed by the application of that theory/model in a second paper. Here the "concepts and methodology" only reviews existing models/methods, hence it is difficult to justify a separate paper.

2. The goal of this paper is to be able to recommend to readers that for a given type of hydrological dataset, modeling method A outperforms method B, etc. Unfortunately this type of recommendation is not really convincing since some of the complicated nonlinear models have many parameters, solution approaches etc., so that scientist X using a different setup of method A than scientist Y may end up concluding method B is better than A.
   In particular, the single ANN (neural network) method as used in this paper is certainly suboptimal. The nonlinear ANN is known to suffer from multiple minima, so one needs to use an ensemble average of say 30 ANN models (trained using random initial weights) to overcome this problem. See Bishop's 1995 book (Neural Networks for Pattern Recognition, Sect.9.6). For all the datasets I have worked with, the ensemble model forecasts have always outperformed forecasts from a single ANN, which gives the erratic behavior mentioned on p.7109, line 3.
   Even for the SVM method (which does not have to perform nonlinear optimization over a large number of weights as in ANN), how one searches for the 3 hyperparameters can lead to different skills, e.g. using a finer grid search for the hyperparameters often improves the SVM model. The poor performance of SVM mentioned on p.7103, line 24 suggests that suboptimal hyperparameters were found. Hence my point is that it is difficult to have confidence in the recommendation of "method A outperforms method B" when it is not clear how well the methods have been set up.

3. The paper uses an excessive number of tables (16 tables in total), many of them poorly designed, and many with trivial information content. Take Table 7 for example: column 3 (RMSE ave) has one significant figure, so the numbers are either 0.04 and 0.05. Maybe the range of values is really 0.0449 - 0.0450 instead of 0.4-0.5, but how can one tell with only one significant figure? Column 9 (MB ave) is even more strange, some numbers are given with 3 decimals while others are with 2 decimals. Look at the bottom two entries – is 0.00 larger or smaller than 0.001? How can one tell when "0.00" could be 0.004 or 0.0001? In Table 9, column 6 (MARE ave) all models tied with exactly the same value "0.04", thanks to the single significant figure used. Then

there are tables with trivial information content – see Table 12. Tables 6, 8, 10, 14 also have minimal information content.

4. I am not familiar with the Kolmogorov-Smirnov (KS) test, but the fact that it finds highly significant results while two probability distributions are visually indistinguishable (p.7107, first paragraph) leads me to suspect it was not done correctly. When a statistical test is applied and the results show incredibly highly level of significance, it is usually the result of overestimating the independent degrees of freedom in the data. The authors did not discuss how the independent degrees of freedom were estimated in their KS test given that the data must have autocorrelation. Also the 12 cases were from resampling from the same dataset.

5. In short, I think the authors tried to cover too many methods, too many datasets, too many statistical tests, tables etc., and ended up spreading themselves too thin (e.g. important details not given, lack of in-depth discussion of tables, figures etc.) The revision should aim at reducing quantity and improving quality.

Minor comments

1. p.7097, line 12: Spell out the acronym AMI and give reference.

2. p.7103, line 7: MARE is "standardized" by observed values, hence not affected by the narrow range of the dataset.
   p.7103, line 8: The argument that the R statistic is the most important indicator is not convincing.

3. p.7103, line 21: SVM with radial basis function kernel can also handle highly nonlinear data.

4. p.7104, line 21. I don't follow the sentence "... no correlations found among the probability distributions...". How does one calculate the correlation of probability distributions?

5. Table 4. The table caption should define the acronyms MARE, MB and R, etc.

6. Many figures have tiny fonts.

7. Fig.3. Define Tr, Va, Te in legend. Why is the magnitude of y_1 greater than that of y_2? And that of y_3 greater than that of y_4? I expect greater difference between Te and Tr than between Tr and Va.

8. Fig.5. Curves have wrong labels? K-nn and M5 are supposed to have overlapping curves, not SVM and M5.