Hydrol. Earth Syst. Sci. Discuss., 9, C4385–C4392, 2012
www.hydrol-earth-syst-sci-discuss.net/9/C4385/2012/

Hydrology and Earth System Sciences Discussions

Interactive comment on "Reservoir computing as an alternative to traditional artificial neural networks in rainfall-runoff modelling" *by* N. J. de Vos

N.J. de Vos

njdevos@gmail.com

Received and published: 26 September 2012

Thank you for an insightful and thorough review of my manuscript. Most of the raised points I felt were valid, and I've consequently made several significant adjustments to the original manuscript.

Below a point-by-point response to all your comments.

Reservoir computing is a collective term that covers several variants viz. Echo State Network (ESN: Natschlager et al., 2002), Liquid State Machine (LSM: Jaeger, 2001; Lukosevicius and Jaeger, 2009), etc. However, given that only ESN modelling is com-

pared and contrasted against other related methodologies in the reported analysis, perhaps the final paper might be better entitled "Echo state networks as an alternative to traditional artificial neural networks in rainfall-runoff modelling".

I agree, and have changed the manuscript title to get rid of the implied but unwarranted generalization from ESNs to Reservoir Computing.

I am somewhat uncertain as to why the present author has opted to restrict his river forecasting analysis to a simple consideration of daily one-step-ahead predictions, since this situation represents the least challenging of all potential hydrological modelling opportunities that could have been pursued. More context and justification is necessary to explain why such a problem was selected for investigation in the first place and the potential relevance of any identified findings. The reported analysis whilst interesting may indeed serve no strong scientific or practical or operational purpose? Surely the main point of a demonstration project, such as the one which is being reported, should be to showcase the numerous strengths and weaknesses of a particular algorithm, by providing a rigorous assessment, performed against a set of increasingly more demanding requirements and/or complex numerical explorations?

Regarding the data set: I have chosen the MOPEX data set because it is a well-known, often-used test case for hydrological modeling, which is in line with the suggestions of Abrahart et at. (2012), who argue for the use of benchmark data sets in ANN river forecasting. It is composed of a range of catchments with different physical characteristics and hydrometeorological conditions. The catchments exhibit varying levels of nonlinearity and reaction time, but they are generally complex and quite tough to model. (This is proven, for example in Duan et al. [2006], see Figure 12 of that paper, where 8 well-known conceptual models were calibrated over the entire available history and 4 of them subsequently have a mean Nash-Sutcliffe value over the 12 catchments that is below 0.55.) Indeed, for operational purposes forecasts based on smaller time scales such as hourly data could shed more light on the relative performance of ESNs, but this was outside the scope of the present research.

Considering this is one the earliest investigations on ESN in river forecasting, and the fact that ESNs are in their infancy, the number of potentially useful modelling investigations is vast. I have originally attempted to do what the reviewer refers to as a "rigorous assessment" by having (1) a strong and varied set of benchmark and traditional models, (2) thorough performance evaluation through multiple objectives and (3) data for 12 catchments. In the revised manuscript, I have taken the suggestions from you and the other reviewer, and added results for the more challenging task of forecasting multiple daily time steps ahead, as discussed in my reply to another comment below. The above arguments have been woven into the revised text.

The only other known hydrological modelling paper involving reservoir computing is that of Coulibaly (2010) on forecasting monthly water levels for the Great Lakes. He also used an ESN. That paper was subsequently discussed and extended by means of a simple linear benchmarking operation in Abrahart et al. (2012b). The current author has not identified or included a consideration of the latter publication in his opening paragraphs and is accordingly directed to it for additional argument. The principal concern in that initial study was a need for more accurate longer term forecasts i.e. greater than one-step-ahead. ESN modelling was found to be substantially superior over longer lead times and this appeared to be its greatest potential offering. The current paper is clearly not fully testing or highlighting what might indeed prove to be its best advantages: although pointers to further research are provided in the closing paragraph. Stronger engagement with published material is called for.

I was unaware of the existence of this highly relevant discussion paper, and I have added it to the introduction.

The revised manuscript includes a graph with model comparisons over multiple lead times. Nash-Sutcliffe values deteriorate rapidly for larger lead times with the the current data set, because of the time scale of the time series (daily) in combination with the residence time of the catchments (often in the order of 12 to 24 hours). Still, the results broadly confirm the earlier findings.

The discussion section on recurrent and partial recurrent neural networks could be improved by a more detailed clarification of terminology regarding the different architectural arrangements, perhaps supported by a hierarchical schematic. The basic structure of a fully recurrent model is a network of neuron units, each with a directed connection to every other unit. Any other neural network variant should be classified as either a partial recurrent network, or a feedforward network, according to the permitted direction(s) of information flow and/or which particular components are allowed to be connected.

Both the text in Section 2 and Figure 1 have been updated to, I believe, more clearly represent the (classes of) network architectures and training approaches.

In data-driven modelling the data is all-important. I would have expected to see a set of tabulated statistical descriptions covering all datasets and subsets that are used in a reported investigation. Simply referring the reader to an earlier paper, published by a different author in a different journal, is not good practice since each individual paper should contain sufficient information within its pages to support the production of a full peer-reviewed publication as a stand-alone entity in its own right.

I believe that including what would be a rather large table with descriptive statistics of training, cross-validation and test data would reduce readability and brevity of the manuscript. It might add some insight into the data, but it is not essential in reproduction of the research. Also note that the MOPEX data set is a benchmark data set that is freely and openly available for download from an FTP site, in case the reader is really interested. (I have included a link to the MOPEX data website in the acknowledgements to facilitate the reader in this.)

Perhaps we could leave the decision whether or not to include such a table to the editor.

In many of the reported instances it is apparent that persistence and/or linear benchmarking models do reasonably well in comparison to some of their more complicated neural network counterparts, suggesting that the matter under examination is in sev-

eral cases perhaps being seen as either a near-linear or perhaps marginally non-linear problem (Abrahart and See, 2007). Full particulars on the linear correlation analysis and average mutual information testing, conducted between each input and output series, must as a result be provided since such mechanistic selection procedures could be a significant controlling factor. This is particularly important in cases where a high degree of near-linear modelling is apparent since the input selection process could perhaps be introducing a bias effect. The data was first converted into a normalised format but thereafter apparently pre-processed using principal component analysis. I do not understand exactly what has happened in the latter process or why it was necessary. Further clarification is required.

The input selection was relatively straightforward, since there was high linear correlation at t0 and a sharp drop-off after that (for both P and Q). I have opted to quantitatively summarize the correlation analysis in the revised manuscript, and I mention that all catchments showed strong similarities in this respect. By presenting this information I make clear that the risk of bias is not that significant. Regarding the second point: after testing, I found that there doesn't seem to be any added value to the use of the PCA. Neither is there strong support in the literature for it. I have therefore no longer used it for input pre-processing.

I am slightly confused about the reported use of training and cross-validation datasets in the modelling process. Most models were calibrated on the training dataset, with the cross-validation dataset being used to perform early stopping. This is standard practice in the field. It means that two independent datasets were included in the model development process and the second dataset may in fact have actually handicapped the production of superior solutions, as opposed to providing a set of clear improvements related to enhanced model generalisation/ prevention of overfitting. Early stopping was not used in the linear regression modelling or reservoir computing operations so were these particular models calibrated on just the training dataset, or a combination of training and cross-validation datasets? If different development datasets are used, how is

inter-modelling fairness achieved, given that particular datasets will offer different modelling advantages and shortfalls. How is everything balanced out in such cases?

The models that do not employ early stopping were trained on the training data solely. This way, all the models largely (N.B. of course the early-stopping models have a peek at the cross-validation data through the early-stopping) base their training on the same data. This has been elucidated in the revised text.

In situations where the last known discharge record is included as a predictor in the modelling process, a neural network model will tend to become a "prisoner of that measurement". This issue is perhaps best exemplified in recent attempts to identify and specifically fix such problems by Abrahart et al. (2007). Their paper should be cited and included in the list of referenced material.

The paper is indeed relevant and a reference has been added to the discussion of this problem.

I wonder if it would have been more logical to include a hyperbolic tangent transfer function in the output neurons of the standard neural network models, so as to match the fully recurrent method?

In hindsight, I believe it is. The revised manuscript is based on simulations with models that use hyperbolic tangent functions in their output, so as to allow for a fair comparison. The methods section has been changed accordingly.

The model comparison section is rather limited. The author states that the overall performance of all models is quite good in comparison to various conceptual models presented in a different paper applied to the same dataset(s). Surely some sort of analytical comparison should be provided? I am also concerned about the fact that the final modelling comparison is primarily restricted to a consideration of statistical metrics and no hydrographs or scatter plots are depicted or inspected. This would of course enable a more detailed analysis of modelling outputs to be performed from which a

deeper understanding of matters might be obtained.

The comparison with the referenced papers has been quantified in the revised manuscript by summarizing the results presented in those papers.

I agree hydrographs and scatter plots are generally of interest, but the relatively small differences between models do not allow easy and insightful comparison in this case. They would have been indispensable in the case of detailed analysis of the inner workings of the models involved, but such analysis was deemed outside the scope of the present manuscript. For these reasons, and the sake of brevity, I have chosen to omit such figures.

The author has included two additional variants of reservoir computing in the final stages of his paper which appear to be an afterthought. It would be better if these items were considered as individual stand-alone models and included in the opening sections as alternative solutions, under the guise of some larger overall predetermined analytical operation. If not, it raises the question, as to whether or not some of the other models under test could also have been improved following a detailed inspection of their respective difficulties and failings?

Thank you for this suggestion. I have followed your advice and moved the theoretical/methodological information regarding the variants to the relevant sections.

Table 2: please explain the difference between trained and untrained weights.

The table caption now explains this.

Figure 1: more detailed explanation required for error loop components.

Figure 1 has been completely revised.

Figures 4, 5 and 6: the plots are deceptive since each graphic is drawn to a different scale and so one cannot compare the different basins in a meaningful manner. Please ensure that all plots are drawn to the same vertical scale to support improved reader

interpretation and prevent misunderstandings. There is no legend. I can only assume that multiple runs were performed on each different type of model and that the red and blue represent some sort of mean and standard deviation values? If so, how many model runs were performed? The main text must be amended to include an explanation for this missing aspect of your overall modelling methodology.

The differences in scaling in Figure 4 has been largely resolved. There is a trade-off between uniformity of the scaling on the one hand, and precision due to the large differences between the catchments on the other hand. I therefore have opted for a trade-off by having only 2 different scales (CE from 0 to 1 or from 0.5 to 1), depending on the overall performance of all models on a catchment. (A note has been made in the caption to warn the reader of this.) Figure 5 now uses the Nash-Sutcliffe coefficient of log-transformed flows, which also allows better comparisons. Scaling of Figure 6 has also been made more uniform.

I have updated the text to clarify that the same number of runs (20) was used for all model variants. The figure captions have also been updated to include this information, along with a legend of the box plot.

References: Duan, Q., et al. (2006), The Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops, J. Hydrol., 320, 3-17.

Final note: A correction has been made to the original manuscript: I now present results for both the original Williams-Zipser fully recurrent ANN, and a variation on this network where there are extra no-delay connections between hidden neurons and the output neuron. (This allowed for stronger direct non-linear capabilities compared to the original form of the network.) In the original manuscript, I mistakenly presented results for the latter under the name of the traditional Williams-Zipser network.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 9, 6101, 2012.