Hydrol. Earth Syst. Sci. Discuss., 2, S164–S178, 2005
www.copernicus.org/EGU/hess/hessd/2/S164/
European Geosciences Union

Hydrology and Earth System Sciences Discussions

Interactive comment on "Constraints of artificial neural networks for rainfall-runoff modelling: trade-offs in hydrological state representation and modelevaluation" *by* N. J. de Vos and T. H. M. Rientjes

N. J. de Vos and T. H. M. Rientjes

Received and published: 17 May 2005

The authors are most grateful to the referees for their thorough and insightful reviews. We acknowledge their comments and believe that the resulting modifications to the paper will greatly enhance its quality.

Our paper reports on a case study on the application of ANNs for rainfall-runoff modelling. The research goals were (1) to investigate some generic aspects of ANN modelling of rainfall-runoff relationships and (2) to gain insights in the optimal choice of hydrological state indicator(s). We encountered the problem of lagged predictions when using previous discharge values as ANN input; an issue that we were unaware of since it has rarely been addressed in literature. In our article we present two approaches to this problem: (1) representation of hydrological state, and (2) model performance evaluation. Both approaches point to possible solutions: the first led to the tests with alternatives for hydrological state indication, the second led to the suggestion of using more than one measure of performance during ANN training. In this paper the latter approach is not further investigated since it is beyond the scope of our original research question, and, to our opinion, it is a considerable research topic that requires much more attention than would have been possible if it was added to the current paper. We feel however that the paper's discussion of multi-objective training of ANNs is of additional value since it allows for a more complete overview of the problem under investigation and its possible solutions.

The most important revisions made to the paper largely reflect the referees' comments and are briefly listed below. This is followed by a number of sections, in which all remarks and comments by the referees are addressed in detail.

- A number of simulations have been re-done, since the number of evapotranspiration input signals to the ANNs was a bit high (as pointed out by the anonymous referee). The new ANN models for daily rainfall-runoff simulations are more parsimonious, but there are no significant changes in model results.
- The abstract has been rewritten to be more concise.
- The background information and description on the workings of ANNs in Section 2 have been condensed and some parts of it have been reformulated.
- The topic order of Section 3 has changed and some parts of the section have been reformulated.
- Our statement on the lack of previous research on prediction lags in ANN rainfall-runoff modelling has been reformulated.
- The concept of overtraining of ANNs has been formulated.
- A short description of the GR4J soil moisture model component has been added.
- A number of explanations on model choices and interpretations have been added.
- We have toned down the discussion of multi-objective training and now address it

more as a topic of future research.
- The presentation and description of some of our results in the text (Section 4) has been modified making them more intuitive and clear. Also, there are more references to the figures added where appropriate.
- All figures have been critically reviewed with regard to the referees' comments. This resulted in updates to the axes labels and captions of a large number of figures.
- All suggested technical corrections have been implemented.
- Some relevant literature references as mentioned by the referees have been added.
- Apart from the significant changes mentioned above, most other detailed comments have been addressed, either by directly rewriting parts of the text according to the referees' suggestions, or by reformulating our opinions/perceptions in a more correct, clear or precise manner.

The remainder of this document describes all revisions in detail.

Abstract rewritten and shortened
The application of Artificial Neural Networks (ANNs) in rainfall-runoff modelling needs to be researched more extensively in order to appreciate and fulfil the potential of this modelling approach. This paper reports on the application of multi-layer feedforward ANNs for rainfall-runoff modelling in the Geer catchment (Belgium) using both daily and hourly data. The daily forecast results indicate that ANNs can be considered good alternatives for traditional rainfall-runoff modelling approaches, but the forecasts based on hourly data reveal timing errors as a result of a dominating autoregressive component. This component is introduced in model simulations by using observed runoff values as ANN model input, which is a popular method for indirectly representing the hydrological state of the catchment. Two possible solutions to this problem of lagged predictions are presented. First, several alternatives for representation of the hydrological state are tested as ANN inputs: moving averages over time of observed discharges and rainfall, and the output of the simple GR4J model component for soil moisture. A combination of these hydrological state representers produces good results in terms

HESSD
2, S164–S178, 2005

Interactive Comment

EGU

of timing, but the overall goodness of fit is not as good as the simulations with previous runoff data. Secondly, the possibility of combining multiple measures of model performance during ANN training is mentioned, because not all differences between modelled and observed hydrograph characteristics (e.g. timing, volume, and absolute values) can be adequately expressed by a single performance measure.

Changes in Tables and Figures

Time periods added to all figures (where relevant, of course)
Table 2 and 3: added to caption: “(results over 10 training trials)”
Fig. 7: minor changes to names in legend; resolution x-as ticks increased; caption changed to: “Correlation between the daily runoff time series and various other time series (rainfall, evapotranspiration) for various time lags.” “(Auto-)correlation coefficients of daily runoff time series for various time lags with rainfall and evapotranspiration time series.”
Fig. 8: minor change to names in legend; caption changed to: “Correlation between the hourly runoff time series and the rainfall time series for various time lags.”
Fig. 9: added to caption: (m3/s)
Fig. 10: X-axis titles changed to: Time (days)
Fig. 11: added to caption: (m3/s)
Fig. 12: added to caption: (m3/s)
Fig. 13: graph now depicts discrete values; caption changed to: “ANN performance in terms of the Nash-Sutcliffe coefficient R2 (with 95Fig. 14: X-axis changed to: “Time (hours)”
Fig. 15: X-axis changed to: “Time shift”; added to caption: “Ĕof the forecasted versus the observed time series.”
Fig. 16a+b: added P+Q simulation to figures; X-axis changed to: “Time (hours)”
Fig. 17a+b: added P+Q simulation to figures; X-axis changed to: “Time (hours)”
Fig. 18a+b: added performance criteria in graph; added to caption: (m3/s)

Changes in text directly based on referees’ comments

Tony Minns' comments

Section 1 and Sections 2.1, 2.2 and 2.3 could probably be shortened and combined to give a more succinct introduction and background rather than covering all of the standard ANN issues that are now common knowledge. For readers unfamiliar with this technique, you can refer to one of the many excellent textbooks on this subject.

We have combined subsections 2.1, 2.2 and 2.4 into one ("Introduction"), and subsections 2.3 and 2.5 into one ("Model training and evaluation"). Moreover, both subsections have been condensed, and references to textbooks have been added (Haykin, Hecht-Nielsen).

Figures 6, 9, 10, 11, 12, 13, et seq. show results from the ANN model. Are these results all using the testing data set? I assume that this is so but it is not explicitly stated in the main text.

The (new) subsection 2.2 contains a statement about the fact that all presented results are indeed results over the test set.

Section 3.3 mentions a standardisation range of -0.8 to 0.7 in order to allow extrapolation. Unfortunately, this modification of the standardisation range only provides a very limited extrapolation on the results (see Minns, 1996). Have the authors examined how much of a problem they have with extrapolation in their particular case? That is, do the maxima in cross-validation and validation data sets significantly exceed the maximum value in the training data set?

A statement has been added about the highest peaks occurring in the training data, and that for this reason we did not consider extrapolation issues.

Is Section 3.4 really relevant to the current paper?

We think so, but we have reformulated the section to make our point more clear. Moreover, the order of the subjects in section 3 has been changed (see comments by anonymous referee).

Section 4.1, page 381, line 17: The problem with prediction lags was identified already by Minns (1998), which is also summarised in Minns Hall (2004). Varoonchotikul (2003) addressed this problem by using recurrent neural networks to feed back the modelled phase error. It is therefore not entirely correct to say that "no previous researchers have appreciated prediction lags in ANN model forecasts". It would, however, be fair to say that no-one has yet been successful in developing a robust methodology for handling this problem.

We have rephrased this part of the text significantly based on this comment and the one by Gaume. Also, some relevant references mentioned by the referees have been added.

Section 4.1, page 382, line 3: The authors have mentioned that the importance of the prediction lag is not always significant. It would be useful to place this in a little more perspective. The authors need to address the issue of whether it is more important to have an accurate estimation of the peak flow or an accurate estimation of the timing of the peak. Even though we would prefer to have both, it is sometimes necessary to make comprises for the sake of expediency. For catchments with a long time-to-peak of several days or more, it is really not at all important if the timing of the peak flow is wrong by a few hours. On the other hand, for catchments subject to flash floods, the timing is far more important.

In the revised version of our manuscript, we have briefly elaborated on this issue for our specific catchment. Nevertheless, we are of the opinion that it is not our task to evaluate performance in terms of magnitude versus timing in a general context. In the explicit cases mentioned by the referee such an evaluation is quite trivial, but this is often not the case.

Section 4.3: Having determined in the rest of the paper that one of your prime objectives is to reduce the prediction lag, it is a shame that this section does not build significantly further on your conclusions. Just stating that the error measures are prob-

ably not sufficient is a rather disappointing conclusion to this otherwise well structured research paper. Problems with the error measures have already been identified by many other authors and, in particular, Hall (2001) demonstrates the effect of timing errors upon the values of traditional error measures like the coefficient of efficiency. The authors of the current paper have identified a major problem without suggesting any possible solutions or tips on how to address this.

In fact, reducing the prediction lag is an additional objective that follows naturally from the combination of our primary objective (the investigation of hydrological state representation in ANNs) with the general aspect of training ANNs. Therefore, we consider our suggestion of multi-objective training (possibly in combination with alternative state indicators) sufficient, considering the scope of this paper.

Anonymous referee's comments

p. 376, ll. 13 to 22: I would posticipate the description of the selection of the number of hidden nodes in section 3.4 (here it is not, in addition, specified to which application -daily or hourly data - nor to which training algorithm the selection refers to); p. 377, ll. 23-25: the description of the selected delays should be moved here from p. 379. p. 378-379: I would suggest to merge the three paragraphs from l. 8 to l. 11 of p. 378, from l. 15 to l. 19 of page 378 and from l. 6 to l. 13 of p. 389.

The subject order of section 3 has been changed significantly and the section is rewritten to get a better flow and to be more concise (see also the comments by Minns). All the text referred to in the above comments has been carefully re-evaluated.

p. 377, l. 7: "a = net"?

This equation was considered redundant and has been removed.

p. 377, l. 18-20: it does not seem appropriate to justify such an important modelling choice on the base of personal experience, not referring to published works (or, better, providing the results of a comparison with other input sets).

We have omitted our personal opinion on this modeling aspect and now mention the fact that it is the most often used method. An exemplary nonlinear method (Mutual Information) is mentioned in the revised manuscript.

p. 379, l. 19: please justify the choice of a time memory of 5 steps for the evaporation when the correlation seems almost flat in Fig. 7.

Indeed, the ANN models are be more parsimonious if only 1 ETP value had been used. And the flat correlation suggests redundancy in the inputs. This fact has been initially overlooked by the authors, and the simulations that used ETP inputs (i.e., the daily simulations), have therefore been redone. The changes in the outcome due to this small improvement, however, are minimal.

l. 24: please specify how the catchment mean lag time is obtained.

A short explanation on this has been added to the text.

ll. 22 to 24: actually, the fact that the mean lag time is so high in comparison to the lead time would advocate a more than reasonable forecasting accuracy

Acknowledged. We have replaced “reasonable” with “good”.

p. 380, ll. 18-19: the multistep ahead forecasts are “disappointing” or, as said at l. 11, “reasonably good”?

This discrepancy was removed. The results are in fact are not satisfying.

p.381, ll. 3-4: actually, in Fig. 14, the major disagreement between the 6-h lead-time forecasts and the observed series does not seem the timing error

True, what was meant is that the timing error is still apparent. We have reformulated this.

ll. 8-9: I would suggest putting off the introduction of Fig. 15 after the description of the procedure of time shifting (ll. 11-13). An interpretation of the results shown in Fig. 15 *is also recommended.*

We realise that the meaning of this figure may indeed be somewhat difficult to grasp. We have followed the referee's instructions to improve on this.

p. 383, l. 15: please justify the choice of 192 h for the memory length of the moving average runoff.

A trial-and-error approach using various memory lengths was followed. This information has been added to the text.

ll. 19-20: a (synthetic) description of the soil moisture component of the GR4J model is here needed. Such description may also help to understand the comment (p. 384, ll. 21-22) regarding the possible reasons for the abrupt changes in the recession curves obtained when using the SM series as input.

A description of the GR4J soil moisture model component has been added to the text.

p. 385, ll. 3-4: please justify the deduction that the information from the input signals is equally weighted.

We have elaborated on this deduction in the text.

p. 373: the PI index is introduced at the end of the section, seeming a subsequent addition, whereas the other two indexes are indicated at the beginning of the section as "the most important measures": I would give the same importance to the indexes.

The PI performance measure is now valued equal to the R2 and RMSE. The section has been rewritten to accommodate for this change.

p. 373, 377 and 378: I suggest removing the references to Zijderveld (2003) and De Vos (2003) because a reference to not easily accessible papers does not seem necessary when describing general properties of ANN often cited in the international literature.

HESSD

2, S164–S178, 2005

Interactive Comment

Acknowledged. The references have been removed.

p. 380, l. 18: the indexes are described more completely (that is along with the mean and standard deviation values) in Table 2 than in Fig. 11 (where, by the way the PI index is 0.13 and not 0.121 as indicated in Table 2).

The table refers to 10 training trials, the figure to the median of this ensemble (as mentioned in the text). The table caption now states this explicitly to avoid confusion.

Eric Gaume's comments

Page 369, line 10 : ANN R-R modelling is presented as relatively new. The references show that ANN have been tested in hydrology for at least 10 years.

The authors believe that since traditional rainfall-runoff modelling dates back many decades, ANN applications can be considered relative new. However, we have rewritten the text and replaced "the relatively young field" with "the still rapidly developing field".

Page 371 lines 10-25, the authors reproduce, without any distance, the justification for the split-sampling method which appears in many papers and text books on neural networks. The concept of overtraining, used essentially in the field of neural networks, is highly questionable. The modelling conventional approach consists in limiting the number of parameters of a model to be calibrated (Beck, 1987; Jorgensen, 1988; Perrin et al 2003) rather than finding 'tips' to handle over-parameterized models. 'The network will start learning the noise in the training data and lose its generalisation capability'. Has this been demonstrated ? To my knowledge no. Moreover, it has been shown that 'overtraining' or over-parametrization occurs even while calibrating ANN on signals with no noise (Gaume Gosset, 2004). There is no miracle ! ANN are not able to distinguish noise and the deterministic part of a signal: the calibration does not begin to fit the deterministic signal and than the noise. Everything is mixed. But the split-sampling method aims at selecting the calibration trials which seem to have cached the greater *part of the deterministic signal: it is in a way a trial and error method. I wonder if the efficiency of the split-sampling method combined with under-calibration has been compared to that of a standard calibration parameter parsimonious method.*

These comments are appreciated. Our presentation of the concept of overtraining of ANNs has been scrutinized and a more conservative formulation of the effectiveness of our approach was chosen. A reference to the work of Gaume and Gosset (2004) was also added.

Page 375 : line 19, it would be interesting to summarise the characteristics of each period in a table: mean annual rainfall amount, maximum daily and hourly rainfall intensities, mean annual discharge, maximum peak and daily discharge...

The authors are of the opinion that such a table has too little added value to the manuscript. We have, however, included a statement about the similarity between the distributions of the separate periods.

Page 376 : lines 20 to 22, I am not convinced by this square root relation between hidden and input neurons. Moreover, the number of parameters of a x-y-z architecture of a neural network is y(z+1). The ANN in tables 2 and 3 have between 96 and 140 parameters ! I am not sure that the term parsimonious' (line 20) is well suited to these models.*

The square-root relation that we mention is not to be taken as an absolute truth; it is just an observation based on extensive testing (Figure 6 shows some proof of this). Interestingly, it is in accordance with most rules of thumb for network structure (e.g., Zijderveld, 2003; Swingler, 1996). Our statement on the parsimony of the models has been reformulated.

Page 380 : line 11, 'reasonably good forecasts'. Nothing supports this statement. The performances of other more 'conventional' models (linear models, lumped R-R models coupled with an AR model on the errors ...) should be given as an element *of comparison. The absence of this reference is a real lack of the paper and should absolutely be included. The efficiency of ANN can only be evaluated in reference to alternative forecasting models. The PI criterion (line 17) is not so good indicating, contrary to what is stated, poor forecasting performances. Nash and R2 criteria are clearly not suited for the evaluation of short term forecasting models.*

The comparison with other model approaches is considered beyond the scope of this paper. This can also be understood from the formulation of our research objectives, which clearly state that we want to investigate some generic aspects of ANNs for rainfall-runoff modelling. Based on our results, we want to raise some issues about ANN modelling without attempting to judge the value of ANNs versus other model techniques. (Note that our introduction contains some literature references to papers that deal with this interesting issue.)

Page 381 : line 20 'this issue has been remarkably overlooked'. This is excessive: 1) many authors have mentioned the lag in the ANN forecasts, Gaume Gosset (2004) can by the way be added to the list, 2) no efficient solution is proposed in this paper. This lag is an important drawback of neural networks as well as of linear models which are both not able to make an efficient use of the rainfall data. Therefore the 'auto- regressive' component dominates in both models. P. 387, Line 8 'has hitherto been neglected by hydrologists': I understand that the authors want to put forward the originality of their paper, nevertheless this statement is excessive and therefore untrue.

These statements have been reformulated (also, see reply to comment by Minns). No solution to the lagged prediction problem was found, but this is partly due to the fact that there may not be an adequate solution. Moreover, this was considered beyond the scope of the paper. Nevertheless, we try to add to the knowledge of the problem by looking at it from two viewpoints that we consider to offer possible solutions: (1) the use of multiple other state indicators, and (2) improvement in the training (evaluation) of ANNs.

Page 382 : this part of the paper is really disappointing, none of the tested approaches prove to be efficient. The conclusions drawn are not really convincing. It appears clear that what the authors call the 'auto-regressive' component is the explanation for the 'reasonably good forecasts' of the ANN. But they partly fail in representing the R-R relation: this is to my point of view the main conclusion which can be drawn from this part of the paper.

Runoff production in a catchment is highly dynamic and non-linear and a large number of flow processes contribute to this. The fact that none of the traditional model approaches is able to simulate all catchment response modes accurately as reflected in the channel flow hydrograph stresses the fact that rainfall-runoff modeling is very complex. ANNs too fail in this respect, particularly when auto-correlative components are not used as ANN model input. Moreover, the use of only a single ANN model for simulating all processes and hydrograph components might limit the capabilities of ANNs significantly. All in all, since these techniques are still rapidly evolving, drawing a definite conclusion at this stage might be a bit rash.

Page 383 : The results reported in Table 2 corresponding to the model (P,Qma) seem to be incorrect. It is strange that this model has much better performances than the (P,Qma,Pma,SM) model including the same variables plus Pma and SM. Figure 18 is not referred to in the text. As in figures 11 and 12, it would be interesting to mention the values of the criterions (R2, Nash and PI). The figures 16 and 17 are not really commented in the text. Are all these figures necessary? They seem to show the results obtained on one of the most important floods. Are they representative of the whole series?

The results are correct and, in fact, summarise the main point of our findings. The (P, Qma)-model performs better in terms of R2, but worse in term of timing, and the (P, Qma, Pma, SM)-model worse in terms of R2, but much better in terms of timing. Performance measures have now been included in Figure 18 (which was referred to on page 385), as recommended. The authors consider figures 16 and 17 valuable

Full Screen / Esc

Print Version

Interactive Discussion

Discussion Paper

additions to the paper. We have included a statement on the representativeness of this short period for the whole series.

Page 385 : what is the aim of this section here in the paper, based only on a literature review ? The authors should test some of the proposed methods or remove this section.

The authors do not agree with this comment. As we argued before (see the reply to the last comments by Minns), a detailed investigation is considered beyond the scope of our original research question. However, we firmly believe that this discussion of multi-objective training is of additional value since it allows for a more complete overview of the problem under investigation and, of course, its possible solutions. Moreover, the topic is considered future research, which has been explicitly stated in the revised manuscript.

Page 387 : line 6 'ANNs are alternatives for traditional R-R modelling'. Such a conclusion can not be drawn without having compared the performances of both approaches.

This ANNs are alternatives, since they offer the possibility of simulating rainfall-runoff processes. By no means do we draw the conclusion that they are substitutes for other methods.

Page 387, Line 25 'complementary conceptual models can be valuable additions to ANN' : this may be true but the results shown in the paper are not really encouraging.

We have reformulated this statement.

Page 388 : last sentence. This conclusion is interesting but not particularly original as mentioned before.

We have rephrased this sentence to be more specific. However, since the majority of applications use low-performance traditional learning techniques such as the back-propagation method (see Dawson and Wilby, 2001), we think that this is an important recommendation to other researchers in the field.

Other changes

- In complete text: replaced “representator(s)” with “representer(s)”
- In complete text: replaced “evaporation” by “evapotranspiration”
- All technical corrections mentioned by the referees are implemented in the revised manuscript.
- A lot of other minor grammatical/textual changes throughout the manuscript.

Added to references

Hall, M.J., 2001, How well does your model fit the data?, Journal of Hydroinformatics, 3, 1, pp. 49-55.

Hecht-Nielsen, R.: Neurocomputing. Addison-Wesley, Reading, MA, 1990.

Minns, A.W.: Artificial neural networks as subsymbolic process descriptors, Ph.D. thesis, Delft University of Technology, Delft, The Netherlands, 1998.

Varoonchotikul, P.: Flood Forecasting using Artificial Neural Networks, Zwets Zeitlinger, Lisse, 2003.

References used in this reply

Dawson, C.W., Wilby, R.L.: Hydrological modelling using artificial neural networks, Progress in Physical Geography, 25, 1, pp. 80-108, 2001.

Swingler, K.: Applying neural networks - a practical guide, Morgan Kaufman, San Francisco, CA, 1996.

Zijderveld, A.: Neural network design strategies and modelling in hydroinformatics, Ph.D. thesis, Delft University of Technology, Delft, The Netherlands, 2003.

Interactive comment on Hydrology and Earth System Sciences Discussions, 2, 365, 2005.