Hydrol. Earth Syst. Sci. Discuss., 10, C527–C536, 2013
www.hydrol-earth-syst-sci-discuss.net/10/C527/2013/

Hydrology and Earth System Sciences Discussions

Interactive comment on "Comparative assessment of predictions in ungauged basins – Part 1: Runoff hydrograph studies" *by* J. Parajka et al.

J. Parajka et al.

parajka@hydro.tuwien.ac.at

Received and published: 26 March 2013

Author response to review 3

We would like to thank the reviewer for her/his very thorough and helpful comments on the manuscript. We have addressed the comments as follows (our response is in italics):

General comments

This synthesis paper aims to compare studies predicting runoff hydrographs in ungauged basins. The methodology consists first on analyzing the median Nash-Sutcliffe Efficiency criteria (NSE) from 34 studies reported in the literature involving 3874 catch-

ments, and second on a more detailed analysis of individual basins. The results discuss the performance NSE of various regionalization methods depending on the climate, the type of the method, the data availability, the model complexity, etc. The topic is novel and I appreciated the large efforts undertaken by the authors to synthesize the majority of the international literature on the topic. The paper is clearly structured and I enjoyed reading it. While the title announces "runoff hydrograph studies", the whole paper is based on the interpretation of only the NSE. My main comments concern the justification of the choice of the NSE criteria in comparison to other criteria, the original data analysis, the impact of data uncertainty on the NSE values and the consequences on paper results such as the significance of the classification of methods, and some secondary comments.

Thank you for the positive and very thorough comments. We made a couple of changes and revised and extended some sections of the manuscript. For details, please see the response to specific comments.

Specific comments:

1. The choice of the Nash-Sutcliffe criteria (NSE): Since 1970, the NSE is an international well-known hydrological standard, and I totally agree to choose it as a criterion because probably this is the only available information on model performance. The NSE is useful to compare different methods or models on the same catchment or using the same set of data, but comparing NSE among various basins is not so evident, and other criteria can be analyzed : i) The NSE is one among other performance criteria such as the error on the total volume, the error on runoff coefficient, the NSE calculated on the root square of the discharge or on the Log of the discharge etc. (please cite the large literature on the criteria functions used in hydrology; e.g. a synthesis in Dawson et al., 2007). Conventional objective functions such as the root mean square error, the NSE, or the index of agreement were largely discussed in the literature because they tend to emphasize the high flows, and consequently, are oversensitive to

extreme values and outliers (Legates and McCabe,1999). On the opposite, the mean absolute percent error tends to emphasize the low flows. It is not evident that the paper obtains similar results when using other criteria. The choice of the NSE must be discussed and the results discussed if other criteria were chosen. The NSE is also very sensitive to high discharge data (especially values and frequencies of peak-flows in comparison to the mean discharge value). If there is a high heterogeneity in time discharge series, a low NSE value may result from bad simulations on a very few number of data corresponding to high peak-flows. Hence the comparison of NSE among basins is not trivial. For all these reasons, I suggest that the authors discuss the large international literature on the significance of the NSE criteria especially when used in different basins (see for example Schaefli and Gupta, 2007), and justify the choice and the use of only the NSE and why not other complementary criteria? If other efficiency criteria than NSE are available in some of the literature used in Levels 1 or 2, it will be very interesting to see if we obtain similar (or different) results than those obtained with the NSE.

We fully agree with the reviewer that comparing NSE across different studies/regions/regimes has some advantages, but also weaknesses and that it is important to discuss it in the manuscript. The criterion for selecting NSE in this assessment was very simple, it is the only one measure which consistently appears in the surveyed literature. There are some studies which report the runoff prediction accuracy by some additional performance criteria (i.e. volume error-Zhang and Chiew, 2009; snow model efficiency - Parajka et al., 2005) or modified NSE (Oudin et al., 2008), but the number of such studies is rather small for a consistent comparison. Hence we would like to stress that future studies should apply and present some additional information and performance measures that will enable to evaluate also different parts of runoff hydrographs, i.e. peak errors, time to peak or event recession.

In response to this comment, we have added following section in the Discussion (same as for the comment of reviewer 1):

"The predictive accuracy of different regionalisation methods was quantified in terms of Nash-Sutcliffe efficiency (NSE). Since it is a traditional performance measure used in hydrology, it has an advantage that almost all reviewed studies evaluate the predictive accuracy by using NSE (an exception is the study Vogel, 2005 that uses R2). On the other hand, NSE is a normalized skill score that measures runoff model performance relative to a baseline model, which is in this case mean of observed runoff values. This can lead to overestimation of NSE in catchments with strong seasonal runoff regime (see e.g. discussion in Schaefli and Gupta, 2007). As pointed out in Gupta et al. (2009), a comparison of NSE across basins with different seasonality should therefore be interpreted with caution. For future comparative evaluations, we would hence suggest to use additional information and performance measures that will also enable evaluation of different parts of runoff hydrographs, i.e. peaks, times to peak (Nester et al., 2011) or event recessions. This will help shed more light on the ability of different regionalisation methods to predict different hydrograph signatures across different runoff regimes."

2. Data analysis: A very important data based was analyzed in this paper. However, it is not clear how the NSE values were identified in Table 1. It will be pedagogic to explain on one study case (one line from Table 1), how the NSE values were extracted from literature and then used in this study; a short explanation can be added in an appendix. What can we learn from the whols set of original papers, and from the median, minimum and maximum values of NSE? Do the authors of the original paper use other criteria functions? This explanation will be helpful to discuss the significance and the uncertainty on the value considered of the NSE.

The identification of NSE values from the literature was rather obvious - we were looking through existing papers and since most of the results were published only in some aggregated ways (e.g as median or range of NSE), we decided to synthesize them consistently in the format as it is presented. In some cases, the results were presented as Figures, so we tried to contact the authors and asked for NSE values (or their sum-

mary). We believe that the level of detail of presented methodology is clear and hence we prefer to retain this part of the manuscript as it is. Please see also our response to the specific comment 1 (above), which discusses the use of other criteria in model evaluations. As it is already discussed in the manuscript, the use of different performance measures and/or development of a universal protocol on reporting scientific results will significantly improve the future comparative assessments in hydrology.

3. Uncertainty on data: All discharge data, especially during peak-flows, are measured with high uncertainties. Please discuss the impact of data uncertainty on the NSE: What will be the impact on the NSE of uncertainties especially on high discharges, and does the uncertainty on data can impact the main results and classifications presented in this paper? In order to reduce the impact on the NSE value of uncertainty on peak flows especially during floods and inundation, does the use of the NSE applied on the root square of the discharge or the Log of the discharge will modify the results?

Yes, we agree with the reviewer, that measuring peak flows is one of the potential sources of uncertainty and that it will be interesting to evaluate the impact of this uncertainty on NSE estimation. Unfortunately, dataset available for the assessment does not include original measured and simulated runoff data, so we are not able to quantify such impact at this stage. However such research questions will be very attractive for future investigations.

4. It will be also interesting in the discussions to comment the highest and lowest values of the NSE for each category analyzed (Figures 2 to 6): please indicate the reference and if possible comment why some studies gave very good values of the NSE (close to 0.9) while others gave low values of NSE (0.4 – 0.5).

In response to this comment, we have added following text (section 4.1): "...runoff predictions tends to be lower in arid than in cold and humid regions. The range of NSE varies between less than 0.4 (Goswami et al., 2007, McIntyre et al., 2005) to 0.87 (Hundecha et al., 2008). The median NSE is ...". We prefer not to repeat the NSE

ranges in other sections (Figures).

5. The number of data used to compare climate regions (x-axis in Fig 2), regionalization methods (Fig 3 and 5), number of catchments (Fig 4), number of model parameters (Fig 6) can vary drastically among regions, methods, etc. When comparing methods, the conclusions depend on the number of available data. The authors must indicate clearly the number of points used for each interpretation, and some details can be added in order to improve the clarity of the paper:

i) for example P 384, L 17-26 and P 385, L 1-2: The paper indicates the number of results for “spatial proximity” (33 results), “parameter regression” (17 results), “model averaging” (11 results) and “regional calibration” (4 results) but didn’t give the number for the “similarity group”. I guess that the total number of results will give 75, but it is not evident. Moreover, the number of data used per class is not given for the other applications (Figures 2 to 6)! I suggest that the authors add on the x-axis of Figures 2, 3 and 4 and for each type of climate (on Fig 2), regionalization method (on Fig 3) and number of catchments (on Fig 4) the number of results (or the number of points) used in each column.

In response to this comment, we have extended the results section and provided more detailed information about the number of studies in particular groups of data. We have added following text in:

section 4.1: "... (Figure 1 and Table 1). In total, there are 11, 5, 16 and 43 studies in arid, tropical, cold and humid climate, respectively."

section 4.2: "The similarity group (9 results) uses parameters from those ..."

section 4.3: "As would be expected, the 21 studies with less than 20 catchments have...".

"For 12 studies with more than 250 catchments the performance however tends to increase."

" Figure 5 summarizes 33, 9, 12, 17 and 4 results for spatial proximity, similarity, model averaging, parameter regression and regional calibration methods, respectively."

Ii) I count 73 points (and not 75) on both Figures 5 and 6; please check. Please indicate also the number of points used for each regionalization method (Fig 5) and for each class of models (Fig 6).

Thank you for this very thorough comment. Yes, the number of points differs. In response to this comment, we have revised the Table 1, in order to make more clear indication which studies are presented in the Figures. E.g. the study of McIntyre et al, 2005 is not shown, as the median of their results is lower than 0.3. We have also revised the Figure 6, in order to more clearly show the number of assessments in each group.

Iii) For the same reasons, please also indicate on Figures 7, 8 and 9 the number of points used or each class of the x-axis.

In response to this comment, we have added the total number of studies for Figure 8 and 9 in the text (for more details, please see the response to other comments). For the clarity of presentation, we would prefer not to change the figures. The Figure 7 is specific as it includes a large number different studies and categories. We thus prefer not to put the exact number of studies for each category and characteristics in the text. We believe that this will has no effect on the message of the paper.

6. Data characteristics: It will be very helpful for the reader if additional characteristics of the data used are given in complementary to Table 1. For example and if available, for each study in Table 1: i) the total number of basins per study; ii) the range of variation of area, rainfall, discharge, runoff coefficients, aridity index, elevation, etc. (probably available for Level 2); iii) the range of variation of the NSE and other error criteria if available.

The more detailed information about some additional characteristics is included and will

be published in Blöschl et al., 2013 (Runoff Prediction in Ungauged Basins: Synthesis across Processes, Places and Scales).

Other comments:

P 380, L 24-25: The paper states that 4 characteristics are analyzed, while only three are given.

Corrected.

P 380, L 18: It is stated that there is 34 studies in Table 1 which results in a total of 75 assessments. However it is not clear how many results are derived from each study. Please indicate on Table 1 the number of results from each study. Table 1: I didn't understand the significance of the various values of the runoff model efficiency for a given study:

i) only one value vs a list of values separated by a comma;

ii) a range of values (e.g. 0.62-0.71). I find approximately 70 values (unique value, or range of values) in the column "Runoff model efficiency"; does these values related to the 75 assessments cited above?

In response to this comment, we have added some explanations in the text and revised the Table 1. Table 1 lists the median or range (depending on the way, how they are published in literature) for all 75 results (separating each method and/or hydrological model applied). We believe that now it is from Table 1 more clear which and how many results are presented in the assessment. We have also added following text:

section 3: "The consistency of results differs between the studies. In some papers, the results are presented only as figures, in others these are summarized by median or range of runoff model performance. Several studies compare ..."

Table 2: i) the first line of the Table indicates that there is 33 studies for Level 1 while it is indicated in the abstract (P 376, L 6) and in section 3 (P 380, L 15) that there is

34 studies; please clarify. ii) Please indicate the number of results (in brackets) for the three options of Level 2 even if the number of results is equal to the number of studies.

We have corrected the number of studies for level 1 in Table 2 and revised the caption as follows: " Table 2. Number of studies (in brackets number of results used in Level 1 assessment) and ...".

Figure 2: i) It is not evident to check that the total number of points is 75; please indicate on the x-axis the number of points used (same remark for Fig 3, 4, 5 and 6). Ii) Figure 2 shows only one line, while the legend indicates “Lines” and not “Line”. Figure 6: On the x-axis, please put “Number of model parameters” instead of “No of model ”

We would prefer not to change the figures, but in response to the comment, we have revised text in the results section. For more details, please see the responses above (particularly response to comment 5).

Figures 7 and 9: Please indicate the number of studies (and/or results) used for each class (represented on the x-axis).

For the clarity of presentation, we would prefer not change the figures. In response to this comment, we have added following textin the section 4.5: " Figure 9 summarizes the performance for different regionalisation approaches, stratified by the aridity index. The total number of catchments is 1570, 1466, 1507, 1241 and 329 for spatial proximity, similarity, model averaging, parameter regression and regional calibration methods, respectively."

Figure 9: In the title of the third figure, please replace (aridity index 3 1) by (aridity index > 1).

Corrected.

References:

Dawson, C. W., Abrahart, R. J. and See, L. M. (2007) HydroTest: A web-based tool-

box of evaluation metrics for the standardized assessment of hydrological forecasts. Environmental Modeling Software 22, 1034–1052.

Legates, D. and McCabe, G. (1999) Evaluating the use of “goodness-of fit” measures in hydrologic and hydroclimatic model validation, Water Resources Research, 35(1),233–241.

Schaefli, B. and Gupta, H. V. (2007) Do Nash values have value? Hydrological Processes 21, 2075–2080.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 10, 375, 2013.