Hydrol. Earth Syst. Sci. Discuss., 6, C764–C770, 2009
www.hydrol-earth-syst-sci-discuss.net/6/C764/2009/

Hydrology and Earth System Sciences Discussions

Interactive comment on "Evaluation of a probabilistic hydrometeorological forecast system" *by* S. Jaun and B. Ahrens

S. Jaun

simon.jaun@wsl.ch

Received and published: 19 May 2009

We would like to thank the anonymous Referee #2 for his positive and constructive comments. Here we want to address the issues raised by referee #2 and comment on possible modifications to the paper. To ease the overview on comments/requests and the associated responses, we split them into several points.

(1) The method adopted is simple but not much used so far; the choice could be motivated in respect to other methods that are more largely documented (e.g. Scherrer et al. 2004 or Houtekamer 1993).

The spread-skill evaluations conducted e.g. by Scherrer et al. (2004) are based on the use of a skill score, e.g. the ensemble RMSE or the BSS, which is compared to a

measure of spread. The resulting relationship can finally be interpreted with respect to the relationship which would result from a "perfect" forecast (e.g. by using a toy model). In difference to this, the method used by Lalaurette et al. (2005) deliberately chooses spread and skill measures/definitions which result, when averaged over spread categories, in a diagonal relationship for a theoretical perfect forecast. Therefore we directly get the information, whether a forecast complies to the given theoretical relationship of a perfect forecast. We will add some sentences explaining this difference to the methods.

(2) A perfect probabilistic forecast could be defined and the qualities addressed with the spread-skill relationship could be explained.

We do not fully understand this comment. The spread-skill relation as used by Lalaurette et al. (2005) is based on the assumption of a (theoretical) perfect probabilistic forecast and its associated properties (cf. preceding answer and p. 1850). But you are right, the qualities addressed with this evaluation are only stated in the results (p. 1856), but not in the methods (p. 1850). To better explain the qualities addressed (whether the ensemble shows enough spread to cover the associated uncertainties), we will add some sentences to the methods.

(3) The assumption that the second and third quartiles are symmetrical (Lalaurette et al, 2005) could be commented in regard of Fig. 6.

Thank you for bringing this aspect up. The assumption that the error-quantiles are symmetrical is certainly not met in this application. We would therefore like to replace Fig. 7 and Fig. 8 with the extended version attached to this reply in order to avoid oversimplified results. The discussion of the two figures will be extended, as more details/aspects need to be covered. We need to add that large negative errors are not met by a sufficiently wide ensemble spread for longer lead times. This underestimation of spread disappears with shorter lead times and results from a single event (August 2005). Excluding the period of this event also removes the underestimation of spread.

(4) The authors are appropriately replying to one possible drawback cited by Lalaurette et al (2005) by testing also their artificial ensemble (HART) which is described p. 1853. The results are mentioned p. 1859 but unfortunately not shown. (...) Wouldn't the spread-skill relationship of HART be more informative than Fig. 7 (HEPS compared with observed runoff)?

While we do not show a figure for the spread-skill relation, a chained plot for HART is shown in Fig. 5. We think that the important messages regarding HART are a) constant spread as HART cannot identify uncertainties associated with a specific weather situation and fully relies on the given runoff value and b) HREF performs better in terms of the RPSS. Therefore we think that additional figures for HART would overemphasize this part. We would prefer not to replace the figure showing the evaluation against observed runoff, as the comparison between the results with HREF and OBS allows the identification of the hydrological model error. Nevertheless we agree that HART should be discussed in section 3.3 and not only mentioned in section 3.4. We will therefore shift and slightly extend this part.

(5) In Fig. 7 and 8, the results are presented together for all the catchments. This includes catchments of the same river at different gauge locations (Table 1). Doesn't such a mixture contribute to the good statistical relationship obtained? A more detailed analysis could take a better profit of this large dataset. (...) The comment about merging the results of all catchments applies also to the RPSS.

We merged the results of all catchments mainly to evaluate the respective forecast systems in a general way, without restrictions to climatological/topographical conditions on the catchment scale (e.g. positioned in the lee of a specific mountain range) and specific catchments/catchment sizes. Also we tried to achieve a condensed and clear presentation of the results, which is only possible at the cost of details. Your concern regarding a possible benefit of scores through the consideration of downstream gauges is generally valid (the bigger the catchment grows, the better the scores should get through compensation/averaging of errors). While we totally agree to this point of

view with regard to e.g. precipitation evaluation, this general rule can only be applied with restrictions to runoff evaluation, as the regimes of the rivers changes (from alpine, nival to more rain-controlled) and lakes are retarding/flattening out runoff hydrographs. Consequently, the RPSS for the different tributaries does not rise constantly along the course of the river (e.g. for river "Aare", catchments numbers C10, C11, C12, C13, C15 and C22 show the following RPSS HEPS/HREF for the leadtime of 4 days: 0.903, 0.871, 0.823, 0.918, 0.877, 0.857). An additional test, ruling out the 6 biggest catchments, shows only a minor decrease in RPSS values. We do not consider the influence of the larger catchments on the spread-skill evaluation to be of major importance, as all runoff values are normalized with catchment sizes and the bigger catchments therefore mainly influence the smaller spread-skill relations (cf. Fig. 6). We actually thought about estimating the contribution of each of the downstream subcatchments separately, but decided against it, as we would possibly need an additional routing model and measurement errors show a higher impact.

(6) Given the large dispersion of the blue circles in such graphs, wouldn't error bars help in assessing the relationship?

An interesting suggestion, but, in our opinion, there are several reasons not to replace the circles by error bars: a) Given the large number of categories, error bars would be messy and would not improve the figure in terms of clarity b) People are well accustomed to point-clouds in dot-plots c) While the blue circles are not interpreted individually, the large dispersion shows the high daily variability, an aspect we would like to show.

(7) ... the RPSS. This latter score has been preferred to the BSS (p. 1851). However, the 0.95 quantile has been added to the quartiles to define the categories. Does this added category resolve runoff peaks as intended?

Checking the 75% quantiles, it appeared to us that higher runoff occurrences should be better resolved. But you are right, the chosen formulation "... to better resolve runoff

peaks" could be misleading (especially with regard to extreme runoff peaks) and was therefore replaced by "... to better resolve higher runoff occurrences".

Caption Fig. 7 (to long to be accepted by the upload form): The HEPS median error from observed runoff is compared to the half interquartile HEPS range for daily runoff [mm] (72–96h hindcasts) for catchments C1 to C23 in 2005. The empty blue circles represent the daily values, while the filled red circles show the means of the spread categories, averaged over 100 daily values. Positive and negative errors are considered separately.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 6, 1843, 2009.

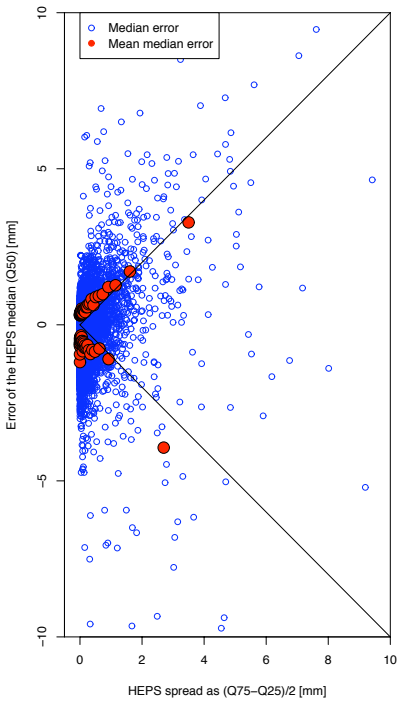

Fig. 1. Caption Fig. 7: The HEPS median error from observed runoff is compared to the half interquartile HEPS range for daily runoff [mm] (72–96h hindcasts) for catchments C1 to C23 in 2005. The ...

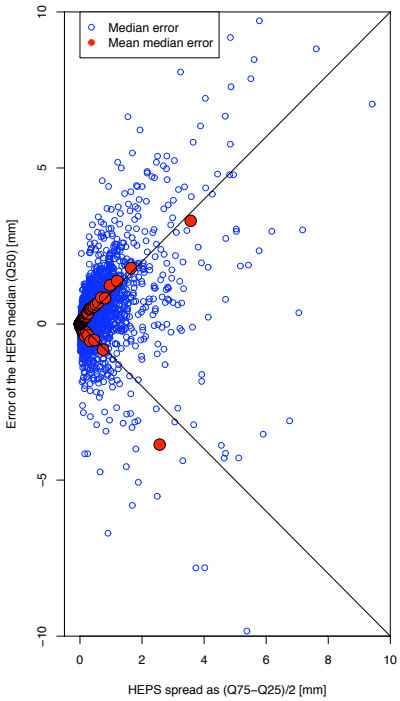

Fig. 2. Caption Fig. 8: Same as Fig. 7 but for HREF instead of observed runoff.