Review of HESS Opinions paper

'On forecast (in)consistency in a hydro-meteorological chain: curse or blessing?'

by F. Pappenberger, H. L. Cloke, A. Persson and D. Demeritt

Dear Authors, dear Editor,
I have reviewed the aforementioned work. My conclusions and comments are as follows:

1. Scope

The work is well within the scope of HESS

2. Summary

The proposed paper discusses (in)consistency in hydrometeorological forecasting chains as the degree of agreement between consecutive forecasts with respect to a property of interest. The subject is discussed at the example of threshold exceedence of discharge forecasts.

After a definition of (in)consistency for deterministic and probabilistic forecasts, causes for forecast (in)consistency are discussed. Further, a literature overview on consistency metrics is given together with a discussion on the use of (in)consistency information for expert and non-expert users. The relation between consistency and forecast performance and approaches for decision-making based on uncertain forecasts are presented. Finally, the authors discuss the advantages of considering the inevitable inconsistency of forecasts (rather than ignoring it) and propose a code of practice for dealing with (in) consistency.

3. Overall ranking

The work is ranked **'Minor revision'**. This is due to some aspects as explained below.

4. General evaluation

Scientific significance

The paper is significant in the sense that it points at an aspect of hydrometeorological forecast quality that is important and known among most end users which have to take decisions based on thresholds. Despite the importance of information on forecast (in)consistency for these users, it is to date somewhat neglected: It is rarely quantified in an objective way and it is not an important criterion in model improvement. Therefore I welcome that the authors raise this issue.

Scientific quality

The (in my eyes) two most important points the authors have raised in their paper deal with the relation between forecast consistency and uncertainty. One is how to describe this relation and the other is how to use consistency information for different users. Both are closely connected and I would like to add a few comments:

The relation between forecast consistency and uncertainty

Forecast inconsistency comes, as forecast uncertainty, from imperfections in the forecast chain (Sect. 2). However, their exact relation is hard to describe in a generalized way due to several points. First of all, uncertainty is not well-defined in Hydrology: As Montanari (2007) pointed out, there is currently no consistent wording about uncertainty assessment in Hydrology and no single, agreed-upon approach to quantify and apply it. Approaches encompass:

- Ensemble approaches (the spread of discharge forecasts based on the use of an ensemble weather forecast is an approximation of forecast uncertainty)
- GLUE approaches (Uncertainty due to model equifinality is accounted for by combination of model output based on many parameter sets)
- Time-averaged error statistics (error distributions derived from large sets of forecast/observation comparison are applied on new forecasts)
- Time-dynamic error statistics (Estimating the current uncertainty from forecast/observation comparison in the immediate past)
- Combinations of the above approaches
- etc.

The reason for this multitude of approaches is partly due to the fact that within the different hydrological applications and scales, the dominant sources of uncertainty strongly vary, that different users have different needs or simply the availability of data. All of these factors also determine which metric will be used to quantify uncertainty.

As long as there are many ways to quantify uncertainty in Hydrology, the relation between uncertainty and (in)consistency will remain case-specific, which makes intercomparison difficult. Also, this touches the question whether (in)consistency is a useful addition to standard/existing uncertainty information or not (be it for decision-making during an event or general comparison of model systems). I would like to illustrate this at an example:

Consider a sequence of temporally consecutive forecasts from the same model for one point of time in the future. Let us further assume that the criterion of forecast accuracy is the fraction of forecasts on the same side of a threshold as the observation (hits+correct negatives/total). This is a typical setting for the evaluation of discharge forecasts. A suitable measure for consistency with respect to amplitude could then be the fraction of threshold crossings (jumps, swings) from one forecast to the next. In this case, there exists a relation between accuracy and consistency of the form that the possible pairs of values are limited to a certain region. This region is shown as blue triangle in Fig. 1. There are two lower and one upper limit for consistency as a function of accuracy

$$cons \geq 1-(2 \cdot acc)\ \forall\ acc \leq 0.5$$

$$cons \geq -1+(2 \cdot acc)\ \forall\ acc \geq 0.5$$

$$cons \leq {}^{1}\!/_{n}\ \forall\ acc \leq \in\ ]0{,}1[$$

$$cons = 1\ \forall\ acc = 0{,}1$$

Where *cons* [0,1]is the consistency, *acc* [0,1] the accuracy and *n* the number of forecasts.

In the Figure, A marks the region with **consistently inaccurate**, B with **consistently accurate** and C with **inconsistent** forecasts. What I want to point out with this example is a) that consistency is not a value in itself (region A) and b) if a forecast is highly accurate or inaccurate, a consistency information does not add much information as regions A) and B) are of limited extend. A combination of accuracy and consistency information is useful (e.g. to for comparison of different models), though, in the region of intermediate accuracy, because here consistency may be high or low.

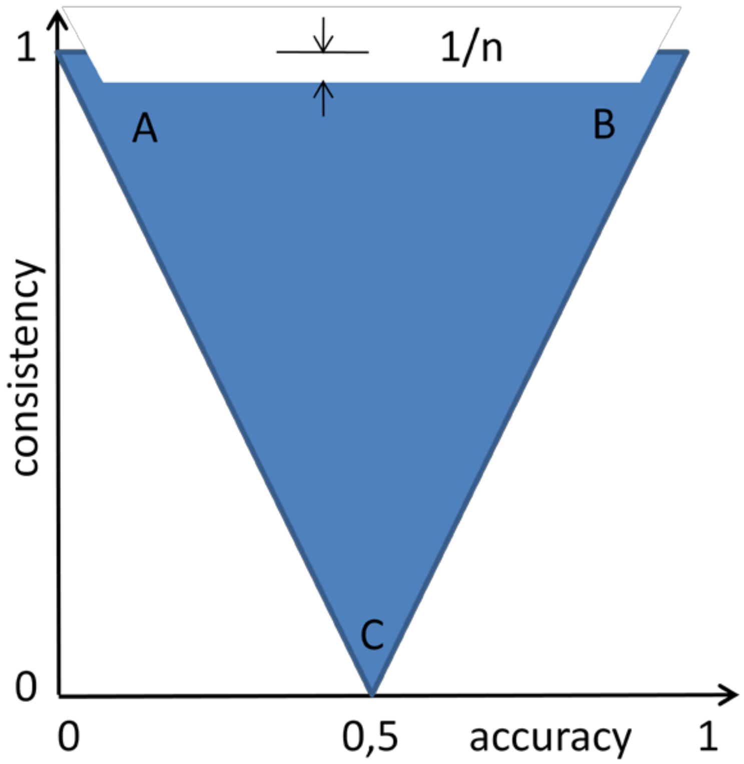

Fig. 1: Possible pairs of values accuracy/consistency for a series of forecasts evaluated for threshold exceedence.

To summarize this, it would in my eyes be worth to further work on a) developing a general framework to quantify and apply uncertainty in Hydrology and b) to investigate the relation between measures of forecast uncertainty and (in)consistency.

The use of consistency information for different users

I completely agree with the authors that a distinction should be made among users with respect to communicating measures for (in)consistency. More and more operational flood forecasts are issued with an information of uncertainty (uncertainty ranges, threshold exceedence probabilities etc.) and end users are getting used to cope with this kind of information. However, this is still a learning process both for forecasters and end users and any new information should be carefully evaluated with respect to its added value vs. the possibility for confusion. As the authors point out (Sect. 2, last paragraph), the inconsistency of individual forecasts is likely to be contained in the uncertainty range or the ensemble spread (especially if lagged ensembles are used, as they explicitly express the consistency of a forecast series). This (realistic) assumption is linked to the question about the relation between uncertainty and (in)consistency and I think we need to better answer this question first before passing information on (in)consistency to the public. This is not the case for trained users however, where (in)consistency measures may be very helpful.

Finally, I would like to add a comment on the temporal aspects of consistency: When quantifying consistency, in my eyes the lead times of the individual forecasts in a series should be considered. For example, a jump/swing/threshold crossing between two forecasts issued at day t-6 and t-5 is for a decision maker much less relevant than from, say, day t-2 to t-1. I would therefore suggest to weigh the occurrence of (in)consistency accordingly, e.g. by inverse weighting with forecast lead time (see Ehret 2010, where a convergence index was presented for meteorological forecasts, but could be adopted for hydrological forecasts as well).

Presentation quality

The work is structured in a logical and comprehensive manner and good to read. It cites relevant literature and gives a good overview on the state of the art. However, there are some points that deserve further consideration:

- Section 3: As (in)consistency is the major topic of the paper, it would be helpful for readers not familiar with the topic to give some examples on how (in)consistency is actually calculated. This could be in the form of briefly presenting selected algorithms of the cited literature and/or to apply them to the example cases in Table 1 and 2 as well as Figure 1.
- Section 5: This section mainly deals with the problem of decision-making in the face of uncertain forecasts, how it is usually handled and how (in)consistency information can help. In my opinion, the current title is a little misleading and I suggest to change it to 'The use of (in)consistency information in decision-making'.
- Section 6: In this section, the benefits of using information on forecast (in)consistency are discussed. As the authors correctly state in the conclusions, inconsistency itself is not a desireable property of a forecast, but using inconsistency information in the face of imperfect forecasts is. To make this point more clear in the text, I suggest to change the title to 'The benefits of inconsistency information' and to change the first sentence into '… the advantages of inconsistency information in the face of imperfect forecasts'.
- The example forecasts
  - 1242/table 1: The forecasts shown in the table do not agree with the ones in Fig. 1. Although they are only fictitious examples, comprehensibility of the text will increase if they agree.
  - 1227/25: The authors use the example of Fig. 1 and Table 1 to demonstrate inconsistency with respect to both amplitude (threshold exceedence) and timing. In my eyes, the forecasts shown in Fig. 1 are not a very good example for inconsistency in timing, as all forecasts show a two-peak event with constant peak timing (03/30 and 04/01). So I would see the major inconsistency of the forecast sequence in amplitude rather than timing.
- 1229/11-12: 'Thus reducing …reduce overall skill'. Can you add a reference that supports this statement?

5. <u>Minor comments</u>

- Literature
  - 1226/8: Zoster → Zsoter
  - 1226/17: Bartholmes et al. ~~2008~~ 2009
  - 1230/24: Bakshi et al., 2005 → reference is missing
  - 1230/27 and 1239/17: The new (and final) reference is Ehret, U.: Convergence index: A new performance measure for the temporal stability of operational rainfall forecasts, Meteorologische Zeitschrift, 19, 441-451, 2010.
  - 1233/5: Dedieu, ~~2009~~ 2010
  - 1240/8: ~~Murphey~~ Murphy
- Text comprehension / spelling (leading number indicates page/line)
  - 1229/18: I suggest to replace 'dimension' with 'aspect'
  - 1232/1: I suggest '… not only due to the quality of NWP- or radar-based forecasts ..'
  - 1232/23: ' … than ~~that~~ the weight that is implied …'
  - 1233/footnote 3: stems ~~from~~ more

- 1238/12: I suggest 'However a perfect system would have no issues with unreliability which complicates our decision making and communication framework'.
- 1242/first sentence: replace 'Fig. 1b' with 'Fig. 1'
- 1245/last sentence: '… ~~plot~~ the number of …'
- 1245/last sentence: what is meant by '(see next session')?

Yours sincerely,
Uwe Ehret

<u>References</u>

Montanari, A.: What do we mean by 'uncertainty"? The need for a consistent wording about uncertainty assessment in hydrology, Hydrol. Process., 21, 841-845, 10.1002/hyp.6623, 2007.

Ehret, U.: Convergence index: A new performance measure for the temporal stability of operational rainfall forecasts, Meteorologische Zeitschrift, 19, 441-451, 2010.