Object : Submission of a revised version of the manuscript (hess-2012-434) by L. Alfieri et al.

We thank Reviewer 1 for the useful comments and the constructive discussion he raised within the review of the article. We have carefully considered the reviewer's comments and worked to include in the revised version of the manuscript the proposed suggestions.
Please find below our point by point responses to the reviewers' comments. The original comments from the reviewer have been reproduced below. They are interspersed with our responses.

1. Structure of the paper: the chapter naming and subsequently the structure of the paper need a bit of work. Chapter 2 is called 'Data and methods' while Chapter 3 is called 'Methods'. Please make sure that all the data and methods descriptions are in one chapter. Possibly you could call chapter 3 'Performance evaluation' to make a clear distinction between what you describe in Chapter 2 and 3.

Reply: We agree with the Reviewer's suggestion and renamed section 3 to "Performance evaluation". In addition, section 2.3 was renamed to "Operational forecasting" to better reflect its content.

2. Reasons for performance: it is not clear what component of the forecast chain provides good performance at which location. In most cases it seems that good performance occurs in relatively large basins, which gives rise to the idea that most performance is gained by the initial states of the system, rather than the meteorological forcing. In fact, Fig. 8 looks like a blue-print of the travel times of the considered rivers, suggesting that most skill comes from the initial conditions. In other words: is it really necessary to run a full ESP prediction? This needs to be discussed and supported by the analysis.

Reply: We have carried out a number of changes throughout the article with substantial additions, which we hope help the reader understand more the system setup and its performance. We have clarified in Section 3 that the performance analysis has been split in two, so that the model performance can be separated by the skills of weather predictions. Indeed, from Sect. 3.1 "The aim of this analysis is to assess how the adopted model is capable to reproduce observed river discharge". On the other hand, Sect. 3.2 reads "Differently from the analysis in the previous section this approach enables the performance evaluation at each grid point of the simulated river network. Furthermore, as the datasets of streamflow predictions and proxy simulations are generated by the same hydrological model, this type of analysis focuses more on the skills of the ensemble weather predictions. Indeed, it allows one to draw indications on the maximum forecast horizon (or potential skill) for which the system yields valuable information". The aim of the paper is to test the average system performance on a global scale and to detect the main problems, so to address the next development efforts. At this level, it is difficult to detect specific errors without looking at case studies, which is what we are performing in our workgroup in parallel works. Results show that initial conditions play a prominent role for the overall system performance, particularly for large river basins (see Fig. 9). However they also strongly suggest that the range of predictability would be much shorter without including weather predictions into the system. A clear example is shown in the case study for the Pakistan flood of 2010 (Fig. 10 to 12). The ESP in Fig. 10 shows very good quantitative skills for the flow peak before the start of the severe rainfall event. An additional figure (Fig. 11) was added showing the persistence diagram for the same forecast, indicating probabilities larger than 50% of exceeding the high alert level (i.e., 5 year return period) on the 10th, 11th and 12th August as early as the 24th July, thus 17 to 19 days before the flood peak. These comments have been included in the text. Although this is just an example, we believe that this shows the strong potential of using ensemble NWPs in a flood forecasting and early warning system.

3. The term ‘value’ is misused in the manuscript and should be replaced by ‘skill’. The value of a forecast is not determined by PSS or any other skill score. This is determined by the user decision and whether the forecast results in a better decision (i.e. cost-loss ratio of the mitigative action based on the forecast should be positive). For instance, a flow forecasting system for the Sahara will have perfect skill but has no value at all for any user.

Reply: The authors agree with the Reviewer about the importance to distinguish ‘value’ from ‘skill’. We have corrected those occurrences in the text where these terms were used inappropriately. Only in one case the term value has been kept in the text (Reviewer’s comment in Sect. 5), as here we mean added value (i.e. in relative terms), and in addition it is referred to the subset of the world’s largest river basins.

4. The length of the retro-active forecast (2 years) is very short! Since the ECMWF forecast system is used, a much longer retro-active forecast could have been used. I am not requesting a new analysis based on a longer retro-active forecast, but please discuss the possible effects of the length of this short retro-active forecast on the results more clearly!

Reply: The idea of this work is to test the overall behavior of the system after the initial setup, so evaluating the whole simulation domain, even though for a relatively short time window. Daily forecasts involve heavy computation, being about 45 times heavier than EFAS simulations, due to larger modeled domain and longer forecast horizon (45 days versus 10 of EFAS). Overall, 2 years of daily 51-member ensemble forecasts spanning 45 days need roughly 85 TByte of disk space, so it requires considerable IT infrastructures. A 2-year simulation window is considered enough to evaluate the overall quantitative behavior of the system and identify the main areas which need improvements. Longer simulations are being tested by our working group for specific case studies, particularly where observed discharge measurements are available for comparison. These will enable further insight on the system performance, particularly in the detection of extreme events. This kind of simulations is performed only on selected river basins, so they are faster and require less disk space for storing the model output, as there is no need to run the whole global domain. As the simulation strategy is substantially different for these two options, we opted for separating them into different research works. In addition, one should consider that operational weather forecasts are subject to model updates roughly twice per year. As we look at overall statistics, rather than monitor the performance over time, we reckon that two years of weather forecasts have quite similar model features and relatively homogeneous model performance. This would not be true for 10 or more years of continuous forecasts, particularly if we consider that the reference climatology is calculated with the same NWP model version. In the revised manuscript, some of these concepts have been added in the first paragraph of Sect. 5.

5. Discussion: some points brought forward in the paper are not properly discussed. I missed two things: first, what are the possible consequences of drift in the initial states (mentioned on p. 12302, l. 13). The second point refers to p. 12310, l. 23. Apparently warnings may be given in heavily regulated areas based on simulations that assume natural flow conditions. Could it be that end-user are misinformed by forecasts from such locations? This needs to be discussed as well.

Reply: Regarding the first point, a drift in the model states could lead to biased initial conditions and consequently to under- or over-estimating streamflow values, even where weather forecasts are accurate. This point has been added in Sect. 2.3 for clarification.
Regarding the second point, as written in the text, "floods and high flows, and particularly their percentile rank, are less influenced by reservoirs which often have limited storage for flood mitigation", therefore the estimation error is reduced in high flow conditions, that is, for the range of interest of the proposed system.

Particular comments to parts of the manuscript
1. Section 2.2.2. The description of the Lisflood component is very limited. It is not clear why sub-surface runoff from HTESSEL needed further routing through lisflood linear reservoirs, how it is subdivided over the different linear reservoirs in Lisflood and how consequently the residence times of these linear reservoirs are estimated over the whole globe. This needs clarification.

Reply: Upon the Reviewer's comment, parts of this section have been rephrased and complemented with additional information, to clarify the adopted approach. HTESSEL accounts for vertical water fluxes and water/snow storage on a pixel basis. However, HTESSEL is not capable of simulating horizontal water fluxes along the river network. To this purpose, Lisflood global is set up to simulate the groundwater and routing processes, using surface runoff and sub-surface runoff from HTESSEL as input fluxes on a resolution of 0.1°.

2. Section 2.3 It would help the reader if you can draw a timeline (i.e. an additional conceptual figure) indicating how initial states are produced and how a forecast is prepared in a forecast batch.

Reply: After trying with some versions of a new figure we concluded that an additional figure of this type would not be very informative. However we have added some text to Sect. 2.3 to clarify the simulation steps. In particular, we clarified that from day 16 to day 45, input maps of surface and subsurface runoff are set to zero, therefore the hydrological model (i.e., Lisflood) will simply convey towards the outlet water already within each river basin.

3. p. 12307. l. 13-14. 'this type of analysis. . ..weather predictions'. I think this is not true. In many places, the skill is to a large degree a function of the memory of the hydrological system, in particular storage of soil moisture, groundwater, surface water and snow pack (where and when occurring). In l. 17-19, this is also hinted at but it is not very clear, since only the relation between the upstream area is mentioned. Please make very clear which components could impact on skill.

Reply: Here we are not sure to understand the Reviewer's point. Operational forecasts are started from (roughly) the same initial conditions that are then used as reference for comparison. Therefore if VarEPS forecasts are correct estimates of simulated values (i.e., ERA-Interim/Land runoff), results will be the same. That is why this analysis focuses more on the accuracy of NWP and it is little affected by errors in the model parameterization or incorrect representation of initial conditions.

4. Section 4.3. The case study is interesting, but could be strengthened very easily. Now only one forecast is shown. What would make the case much more interesting is too see how far ahead a warning level could have been detected using the system and discuss the established warning lead time, accuracy of peak and timing of the peak as the forecasts progress in time. So you should include a number of earlier forecasts to show this. Finally, discuss again what the

cause of performance is. Is it the quality of the meteorological forcing? Or is it the dominance of the initial states that result in a good performance?

Reply: Upon the Reviewer's suggestion an additional figure has been added (now Fig. 11), showing the persistence diagram of 5 consecutive forecasts from 24th to the 28th July. Additional description and discussion has been added too in Sect. 4.3.

Reply to comments in the paper (as supplement)
P12297, L10: well-integrated was replaced by integrated. It means the whole system (IFS, HTESSEL, Lisflood, postprocessing and publishing results on the web) is managed globally through a single control application.

P12297, L17-19: We replaced "To set up a forecasting system…" with "To set up a forecasting and warning system…". As the Reviewer correctly commented, the discharge climatology is not strictly necessary to produce a forecast. It is necessary to derive forecast anomalies. On the other hand a warning system requires warning thresholds to compare with real-time forecasts.

P12298, L7-8: It means that weather forecasts are an intermediate product of the IFS, but they are not used by GloFAS. We extracted surface and subsurface runoff directly and then use them as input to Lisflood.

P12298, L22-27: As suggested by the Reviewer this part has been removed and merged to the description in Sect. 2.2.1.

P12301, L11-19: The text has been substantially modified to indicate why sub-surface runoff from Htessel needs further processing and to clarify the function of the two reservoirs.
In the current state no calibration took place, so we are using fixed values:
Within EFAS the parameters to control percolation to the lower groundwater zone, the residence time of the upper and lower zone and the routing parameter (a multiplier to Manning's roughness) are some of the parameters used for calibration (see Feyen et al 2007). In this setup of GloFAS these parameters are set to reference values. Parameter estimation through calibration is subject of future work.

P12309, L25: The text has been complemented with the suggestion from the Reviewer.

P12310, L6-8: This part was enhanced with additional information, as suggested. However, it is not straightforward to relate such discrepancies to specific model parts without deeper analysis which are out of the scope of the paper. Therefore we preferred to avoid speculation on the causes of the error.

P12310, L11-12: The sentence was adapted according to the Reviewer's suggestion.

P12310, L23-24: Reviewer: Could it be that end-users are misinformed by forecasts from such locations?
Reply: No, actually this is the opposite case. As stated in the text, floods and high flows, and particularly their percentile rank, are less influenced by reservoirs which often have limited storage for flood mitigation. That's why scores are higher for high flows (i.e., PSS) compared to average conditions (i.e., CV). Also, we do not claim that the system is ready to be implemented for operational warning. This study should serve as a push for further developments and improvements to those parts recognized as critical and leading to larger errors. This is also clarified in the conclusions.

P12312, L8-9: We gave additional details in this sentence as we realized it mislead the Reviewer. In this sentence, we refer to the comparison between maximum lead times of 10 days with AROC>0.7 (from Fig. 8), while for the same river reaches the CRPSS is skillful (>0) even for a lead time of 25 days (from Fig. 7). The motivation is explained in the following part of the same paragraph.

P12314, L4-6: It's 900000 cubic feet per second (cusecs), which is about 25500 m^3/s, therefore less than 39600 m^3/s. We have kept such figures in cusecs for coherence with the BBC article and then converted it in m^3/s. A link to the webpage of the article has been added as requested.

P12316, L6: Modified as suggested by the Reviewer.

Other suggested changes and reference updates have all been amended according to the Reviewer's suggestions.