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

We thank Dr. Paiva (Reviewer 2) 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.

MAJOR COMMENTS:
1) Objective and conclusion:
The authors claim that the "aim of this study is to assess the feasibility of transferring methodologies and concepts from the EFAS system to the global scale and . . .". However, there is no discussion at the "Discussions and Conclusion" section explicitly comparing the EFAS and GloFAS systems. How the performance of the GloFAS system compares with the EFAS's? What methods and concepts that worked well at European scale (or region) were successful or unsuccessful on the global scale? Why? What should be improved in the EFAS system to useful on the global scale? This kind of questions should be answered at the discussion and conclusion section.

Reply: In the Conclusion section we have clarified that GloFAS has been set up following similar structure as in EFAS (i.e., from a methodological viewpoint), but no specific comparison has been carried out (in Europe) between the two systems, as it is not the scope of the paper. In fact the aim is not to search for an alternative for EFAS in Europe but rather having a similar system outside Europe. This for sure imply having worse skills for GloFAS (considering the same basin), due to coarser model resolution, the use of coarse scale model reanalysis (ERA-Interim) rather than observed meteorological measurements as in EFAS and the current lack of model calibration (not easy to perform at the global scale due to the scarcity of discharge measurements for validation in large areas of the Earth).

2) Performance evaluation:
- Page 12307, line 7: Why have you performed a hindcast test of only 2 year length? It seems that with the methods used by the authors, it would be possible to do the same tests for the 1990-2010 period. Also, as explained in the manuscript (Page 12311, lines 23-25), a longer test period would allow the evaluation of the system performance using warning thresholds based on discharges related to 2, 5 and 20-yr return periods, as it is actually used in the GloFAS. Please justify the option for a 2 year test period in the manuscript.

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.

- Page 12309, line 10: Why not testing the system performance against discharge observations? It would be interesting to see Figures of CRPSS and AROC (or max lead time where AROC > 0.7) for each gauging station, similar to Fi. 2 and 6.

Reply: We acknowledge that this would be a very interesting analysis to perform for assessing the model performance in operation. However, most discharge measurements were available for the 1990s and early 2000s, with only very limited data availability in the period of simulation of the operational forecasts. As mentioned in the conclusions, the aim of this work is to test the overall behavior of the uncalibrated system after its initial setup, and to help identify the main components where to address the main future development efforts. As stated in the previous point we foresee to include the suggested type of analysis in future works, focusing on specific case studies where long enough windows of measured discharge are available.

- Page 12309, lines 5-21: The approach of testing the system performance against results from a reference run considers that the only source of errors is the weather forecasts. But it is not true, important errors can arise from imperfect model structure and parameters and also wrong model initial states at the beginning of a forecast. That is why it would be important to test the system performance against discharge observations too.

Reply: We agree with the Reviewer. In this work we have separated the performance analysis of the hydrological model (Sect. 4.1) from the one of the weather forecasts in producing streamflow predictions (Sect. 4.2). While this approach is useful to separate the different proportions of the total system error, we acknowledge that a verification against measured discharge is important to determine the overall system performance. As discussed in the previous point this choice was necessary to take profit of longer windows of verification and of a larger number of stations (presented in Sect. 4.1). In addition, comparing the system forecasts with a simulated reference run has the advantage of computing scores on the whole global domain, which allows us to detect spatial variability of the system performance, both along the same river basins and for different climatic regions.

-Page 12314, line 22 to Page 12315, line 5 and Page 12315, lines 17-29 In my view, the analysis performed in the manuscript is too weak to conclude, for example, that it is possible to detect flood events with forecast horizon as long as 1 month. The methods used in this paper do not consider all source of uncertainty such as errors in model structure and parameters or errors in model initial states. For example, aiming at studying the sources of prediction uncertainty in the Amazon River basin, Paiva et al. (2012b) showed that initial conditions of model states (e.g. river discharge and water levels, groundwater storage, among others) play an important role for discharge predictability even for large lead times (↑ 1 to 3 months) on main Amazonian Rivers. These analyses indicated the feasibility of hydrological forecasts based mostly on optimal model initial states, while the weather forecasts would have a secondary importance. Also, development of data assimilation methods was encouraged to estimate the optimal initial states.

Also, figures 7, 8 and 9 can show an incorrect good performance of the GloFAS system where the model can't represent observed discharges at the climatology simulation. In my view, to perform a stronger analysis, it would be necessary to evaluate forecast results against discharge observations.

Reply: We have carried out some modifications particularly throughout the Conclusion section to clarify that this (i.e., a skillful detection of hazardous events with forecast horizon as long as 1 month in large river basins) is not the result of the proposed system, but rather of a perfectly calibrated system and provided that initial conditions are estimated correctly. Indeed there is no reference to GloFAS in this sentence. It is just a reflection coming from the analysis between the simulated climatology and the ensemble forecasts, which in a way tests the limits of predictability of current weather forecasts. The same is valid for figures 7, 8 and 9, though we have make some additions to clarify that these are not to be intended as the current system performances. "Results in **Error! Reference source not found.**-8 show the current system potential assuming that the simulated climatology corresponds to the actual river conditions, that is, for a perfect model process representation, calibration, and perfect input forcing."

-Section 4.1.: The manuscript brings some explanations for model errors at some regions of the globe, including dam regulation, water withdrawals for irrigation, incorrect modelling of snow processes or biased input temperatures. In my view, it should be added some references or analysis to support these explanations, otherwise it seems only speculation.

Reply: Following the Reviewer's suggestion we have complemented Sect. 4.1 with additional information in support of our statements. In details, regarding the low CV values in Mexico and Australia, we added some quantitative information on the average specific discharge at those stations and compared to the average of all the considered stations. Also, we have added that the reason for poor results in north-eastern Russia was found by analyzing the time series at these stations (not shown in the article). The idea is to find simple indicators to explain some model behaviour without going to a high level of detail at specific locations or focus on specific case studies.

3) Timing errors:

- Figure 3: The large timing errors between simulations and observations presented in figure 3 opens room to the following discussion. The authors claim that, since the goal of the GloFAS system is to provide warnings based the exceedance of warning thresholds (developed based on simulation results), it is much more important that the simulation model reproduce discharges in terms of percentile ranks rather than quantitative values. Consequently, errors related to bias between simulations and observations wouldn't cause problems in the flood forecasts. That is actually a very interesting idea. However, how does this method deals with timing-related errors? In the case of advanced hydrographs, during the operational monitoring of the system results, I would expect first an advanced false warning followed by a miss of a flood event. That can have important implications in the performance of this system. The timing errors between simulated and observed discharge series could be easily evaluated using the "Delay Index" presented in Paiva et al. (2012a) or using more complicated metrics as the one presented by Liu et al. (2011). It would be interesting to include a discussion on the implications of timing errors in the performance of the flood forecasts system.

Reply: In this work the main focus is detecting the main potential sources of errors affecting the system, so to address future development work. Some timing errors have been detected and reported in the paper, in particular due to incorrect modelling of snow accumulation and melting processes (at some stations in in Russia and Canada). The authors believe this is an interesting analysis to perform but which goes out of the scope of the paper, as it misses dedicated analysis to substantiate conclusive discussions. In fact, in the present conditions such discussion could only be based on speculations. However, we have included a comment in the delay of the flow peaks shown in Fig. 3, which can be partly attributed to the non-representation of floodplains and backwater effects (see reply to next point).

4) Simulation model:
-Page 12301, lines 16 - 18: Is the storage and flow of water in floodplains taken into account for streamflow rounting? There is some recent advances in regional / global modeling of surface waters flow, moving to diffusive and full hydrodynamic modeling of channel flow to consider backwater effects (e.g. Paiva et al, 2011, 2012a,b; Yamazaki et al. 2011) and floodplain store of water volumes exchanged with the river (e.g. Paiva et al, 2011, 2012a,c; Yamazaki et al. 2011; Decharme et al. 2008). The Fig. 3 shows results in the Amazon basin, where simulated discharges are very advanced in comparison to observations. On the other hand, Paiva et al. (2012a,c) and Yamazaki et al. (2011) showed that not representing backwater effects and mainly floodplains in simulations can lead to very advanced hydrographs. Possibly, the delay errors shown in Fig. 3 are due to the same reasons. It would be interesting mentioning these recent advances and relating it to the model and results from this manuscript.
-Page 12316, lines 7-9 You could also mention that there is room for improving the representation of river- floodplain flow to consider effects of backwatering and floodplain storage in flood wave traveling along the river network.

Reply: Upon the reviewer's comment, a new paragraph has been added at the end of Sect. 4.1 to mention the additional source of error due to the non-inclusion of river floodplains in the hydrological modelling. Relevant publications on the topic have been added too. This helps explaining the considerable delay of flood peaks shown in Fig. 3.

SPECIFIC COMMENTS:
-Page 12294, lines 21-24: Is the 230% increase in economic damage computed comparing the 2011 year with the 2000-2010 decade? If yes, I am not sure that such comparison is appropriate, considering the natural inter-annual variability of climate.

Reply: These are figures taken from the cited publication by Guha-Sapir et al. (2012). Unfortunately no data are included for each year of the past decade, besides the last year (i.e., 2011). We still think it is a interesting comparison to show and anyway it does not imply that the 230% increase is going to be a linear trend for the coming years. For sure there is a significant inter-annual variability, but on the other hand a decadal average is a good basis for comparison. These figures show that the last year was well above those of the last decade, which was mostly due to the flood affecting Thailand in Aug-Dec 2011.

-Page 12296, lines 17-20: Data assimilation methods can also be used to derive optimal estimates of model initial states based on observations.

Reply: This was added in the description as suggested.

-Page 12300, line 27: ". . . hydrodynamic channel routing model. . ." This term is usually related to full (or close to full) Saint Venant equations, where the flow in channels is modeled considering all inertia, gravitational, friction and pressure forces. Indeed, as described in the manuscript (Page 12301, line 18), Lisflood simulates the flow in the river network using a simpler kinematic wave approach, that considers only gravitational and friction forces. It would be better to remove the term “hydrodynamic routing model” for the Lisflood model to avoid confusion.

Reply: We agree with the Reviewer’s point and have removed the term hydrodynamic. We left in the text that it includes a (channel) routing model, which is the kinematic wave formulation. This is important to be stated as it is one of the main features which distinguish Lisflood from HTESSEL

-Page 12301, lines 11-16: How the parameters of subsurface and groundwater reservoirs were estimated? How river parameters (e.g. river width, Manning coefficient) were estimated? Is there another publication describing the global application of the Lisflood model? If not, it would be important to provide more details about model parameters.

Reply: We have added to the text that we did not perform any calibration yet and all parameters are set to literature 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 parameters (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 fixed values. Parameter estimation through calibration is subject of future works.
River parameters like channel gradient, Manning’s coefficient, river length, width and depth were estimated by taking the digital elevation model, the river network and the upstream area into account. Further details of the Lisflood model can be found in van de Knijff et al. (2010). These comments have been added to the text.

-Page 12301, lines 18 - 19: Was any upscaling method used to derive flow directions at the 0.1 degree resolution?

Reply: Upon the Reviewer’s comment, we have specified in the text that the river network is taken from the Hydrosheds project (Lehner et al., 2008) and upscaled to 0.1° by using the approach of Fekete et al. (2001). In the next developments the upscaled 0.1° dataset of Wu et al. (2012b) will be used.

-Page 12302, lines 5 - 8: What does it mean? Does it mean that the model is forced with null precipitation between 15 to 45 days lead time? At some regions, null precipitation may not be the best guess. Have you considered using an ensemble of historical precipitation data from past years, such as used in the “Ensemble Streamflow Prediction” method from Day (1985) ? Maybe you could mention this idea at the conclusion section. And what about the other meteorological forcings, such as surface temperature or solar radiation? Do you consider a persistence criterion?

Reply: Indeed 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. This was clarified in the text. Also, in the conclusion section we added that for future applications, climatological average values could represent better alternative to use to the current assumption of no input flow.

-Page 12302, lines 5 - 8: What about the initial states of the hydrological model Lisflood (River discharge, surface, subsurface and groundwater volumes) ?

Reply: As described in Sect. 2.3, the initial conditions are updated on a daily basis using weather parameters of the first day of VarEPS forecasts.

-Page 12302, line 19: “the 5 and 20-yr” = “the 2, 5 and 20 yr”?

Reply: No, indeed threshold exceedance maps are only plotted for 5 and 20 year return periods, which are those of highest interest for floods. On the other hand, the ESP at reporting points are plotted also for points exceeding the 2-year return period level, as they are useful for diagnostic.

-Page 12302, lines 4-15: Do you assimilate data from observations to update the initial states of HTESSEL or Lisflood model?

Reply: No, the only update with observation is the assimilation of weather observations to run the forecasts with the IFS model, which then produce surface and subsurface runoff through the HTESSEL component. Lisflood has no data assimilation.

-Page 12302, line 24: Have you considered using fixed reporting points located at the world most populated cities?

Reply: Not specifically, however in many large cities located along major rivers there is a station which is included in our database. In general these are only visible when the ensemble mean of streamflow forecasts exceeds the 2-year return period.

-Page 12303, line 15: “Persistence diagrams”. These diagrams were not present in the manuscript.

Reply: An additional figure has been added (now Fig. 11) which shows the persistence diagram for the forecast in Figure 10, concerning the flood in Pakistan in 2010. The corresponding description is in Sect. 4.3.

-Page 12304, line 10: How can you be sure that a visual check can remove all problems related to discharge regulation in stream flow data from gauging stations? For example, Fig. 2 show lots of gauges used in this manuscript located at Paraná River basin (southeast Brazil), where there are lots of reservoirs from hydropower plants. The discharge time series of these gauges are possibly affected by reservoir regulation and hydropower operation. Have you considered using a database of world large reservoirs?
-Page 12310, line 6: “. . . model performance is often substantially affected by dam regulation. . . ” How can you tell that the model performance is affected by dam regulation, if you claim (Page 12304, line 10) to have removed all gauging stations affected by discharge regulation using the visual check?

Reply: In our opinion the selection of points for verification is not a critical decision for the presented analysis. The aim was to keep a relatively loose criterion to have a quite large number of stations to compare. Some of them (i.e., “those with evident discharge regulation”) were manually excluded because the regulation was really clear in the time series, due to large regulation volumes in small river basins. We are aware that many other rivers are regulated, but this has a less critical effect in those rivers where the regulation volume is relatively small compared to the average runoff, as this effect becomes even less evident during flood events. It is actually useful to see the effect of a regulated river on this analysis as reservoirs can be considered as another potential feature to be included in future developments for the system. This is the reason why removing all regulated rivers was not a strict constraint in this analysis.

-Page 12306, line 16: Please provide a reference for the Peirce's skill score.

Reply: The reference has been added as suggested.

-Page 12306, line 17: Why have you choose to use a threshold equal to the 90th percentile of discharges? It would be better to use thresholds equal to the discharge values corresponding to 2, 5 and 20-yr return period, since these are the values actually used in the warning system.

Reply: There are mainly three reasons for this choice:
1 - First, such percentile is a good tradeoff between being representative of high flow values and including a sufficient number of events to draw robust statistics. This is also stated in Sect. 3.1.
2 - To be coherent with the threshold percentage used in the calculation of the AROC (see Sect. 4.2). In this case it is even more important to choose a relatively low threshold, in order to have at least a threshold exceedance in every pixel of the map (otherwise missing values are produced as results).
3 - The PSS should not be used for very rare events, as in such case the second term would tend to zero due to the large number of correct negatives, therefore the score would reduce to a simple probability of detection (POD).

-Page 12307, line 22 to Page 12308, line 5: This paragraph is not very clear. For example, how do you deal with Lisflood state variables in the bias correction?

Reply: As suggested, we have clarified in the text that the bias corrected climatology is used only for validation purpose, while model state variables are not affected by the correction.

-Page 12308, lines 17-19: Not clear, please rephrase it.

Reply: This way of describing the CRPS is taken from the book by Wilks (2006), so a reference has been added to it.

-Page 12309, line 6: What threshold value was used? I guess it was the 90th percentile. Please clarify it here.

Reply: In Sect. 3.2 a general description of the ROC is presented, while technical details on how the score was applied in the analysis are shown in Sect. 4.2. In the same section we indicated that for this analysis we used the 90th percentile as threshold value.

-Page 12310, line 11-12: You could add regressions lines (e.g. PCC = a0 + a1*Area or PCC = a0 + a1*log(Area)) for different latitude ranges at Fig. 5 to support this affirmation.

Reply: The regression seems not to have linear trend, even in the logarithmic space, so we consider it not very appropriate to fit the data with analytic functions with many (i.e., three or more) parameters.

-Section 4.1.: Please, also provide ranges, mean, median or any percentile value for Nash and Suttcliffe index values.

Reply: Some relevant figures on the Nash-Sutcliffe efficiency were added ad the beginning of Sect 4.1, as suggested.

-Page 12312, line 29: In large rivers that usually have marked seasonal regimes, a better reference forecast would be the climatological value of discharge for each Julian day.

Reply: We agree with the Reviewer's point. As the system follows a distributed approach, we decided to adopt a unique criterion (i.e. of persistence) so that it is applicable to any grid point.

-Page 12314, line 8: "MODIS Rapid Response" Please provide a reference for this dataset. Is is one of the products mentioned in Page 12295, lines 25-26?

Reply: MODIS is a Imaging Spectroradiometer installed aboard the Terra (EOS AM) and Aqua (EOS PM) satellites. A reference has been added where suggested.

-Page 12315, lines 25-29 It could be mentioned the use of data assimilation methods to update model states.

Reply: We have added in the text that recent works in data assimilation and correction techniques demonstrated large potential for improving quantitative streamflow forecasts at those stations where discharge measurements are provided in real time (e.g., Bogner and Pappenberger, 2011).

-Page 12315, lines 6-16 The authors should present some Area x Skill Scores figures to support these conclusions (maybe using regression lines too).

Reply: This sentence refers to added value of the forecast system, rather than pure skill scores, so a figure would be of difficult application. What we mean is that in these types of basins the proposed warning system has the highest potential of being useful, compared to any alternative methods of forecasting future river states. As stated in the article, in large river basins the skill of the system can be up to 1 month. However in these basins variations of the river level are slow and quite predictable, as they often follow seasonal patterns. On the contrary, the maximum added value is given where the proposed system represents the forecast option which provides the longest horizon in predicting the future trend.

TECHNICAL CORRECTIONS:

-Sections 2. and 3. : The names of sections "2. Data and Methods" and "3. Methods" are confusing.

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

-Page 12300, line 10: "Orography" = Topography?

Reply: Yes, meaning the morphology of the terrain

-Figures 2 and 6. Maybe some gauges were hidden behind the ones with large upstream area represented by large circles. An alternative would be to represent gauges with different ranges of upstream area using different symbols, such as circles, squares, triangles.

Reply: After trying some different visualization options we opted for keeping the initial version. Other options looked more confused and of different interpretation. In any case some points would overlap, due to the large number of stations and the variable density.

-Figure 9. caption: “see red markers in Fig. 7” = “see red markers in Fig. 8”

Reply: Amended.

-Figure 10: It would be easier for the reader if the location of the area presented in Figure 10 was shown as a rectangle in one of the figures showing the globe (e.g. Fig. 8).

Reply: Upon the Reviewer’s suggestion a black-contoured rectangle of the area of Figure 10 was drawn in Figure 6.

-Figure 11: It would be easier for the reader if the location of the area presented in Figure 11 was shown as a rectangle in Fig. 9.

Reply: Upon the Reviewer’s suggestion a black-contoured rectangle of the area of Figure 12 was drawn in Figure 10.