Dear Editor,

Thank you for your comments.

Please find in this document our responses and adjustments to the manuscript explained in detail.

We hope that after considering each comment of the three reviewers and yours that the manuscript has final reach maturity and will be considered for publication.

Sincerely Yours,

Vera Thiemig

1) I think it would be good, based on available papers on global hydrology, to say something on the expected) source of skill in different parts of Africa (see van Dijk et al. 2013, Candogan et al. 2013 and others) although you did not do this work yourself (maybe in the introduction)

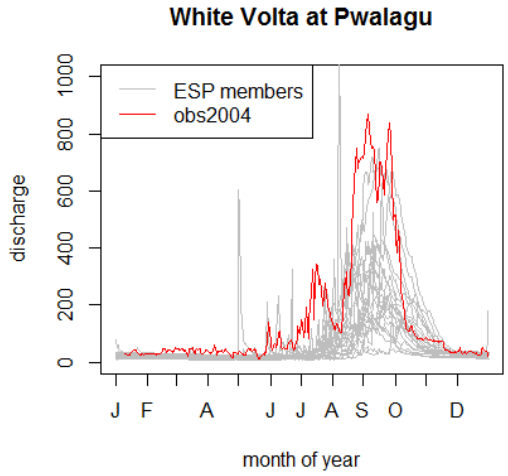
We added a paragraph in Section 3.2.1 (of new manuscript structure), in which we are speculating on the forecasting skill

As the skill of the conventional ESP (not shown here) decreases similar to the skill of the AFFS with increasing lead time, the decrease in forecasting performance cannot be affiliated to possible inaccuracies of the ENS only, but there must be other additional influencing factors. However, establishing the sources of predictability is beyond the scope of this paper, but subject of future research. However, cross-comparing the CRPSS and the limit of predictability with the KGE' received during calibration (Fig. 4) suggests that the skill of AFFS to predict streamflow is strongly dependent on the optimization of the hydrological model. For locations where LISFLOOD seems to be well fitted, expressed by a good hydrological performance (KGE' > 0.6), the forecasts were mostly skilful (positive CRPSS); while they were without skill (CRPSS negative and limit of predictability equal zero) exclusively at locations where the KGE' was less than 0.6 during calibration. Studies on global, seasonal streamflow prediction (Yossef et al. (2013), Albert et al. (2013)) show that the source of forecast skill varies from basin to basin. Their results suggest that the forecast skill in monsoonal and semi-arid basins is mainly dependent on the skill of the meteorological predictions, while in large basins they were found to be more dependent on the skill of the initial conditions. […]

2) I agree with reviewer 2 that using persistence makes much more sense than a seasonal mean (especially because you focus on flood forecasting application) and requires little extra work
The choise of the verification metrics is unclear. I would also like to see reliability plots and maybe others. Besides the choice of the verification metrics, there is a problem with the one-year hindcast period which makes your statistics in essence meaningless and no claims in performance can be justified. Boostrapping procedures might give insight into the effect of limited sample sizes (see for instance Lopez Lopez etal., 2014 HESS)

Reply to: using persistency → conventional ESP

A conventional ESP has been derived from the long-term simulation (running LISFLOOD with the GPCP-corrected ERA-Interim over the time period 1989-2010). The Limit of predictability and the CRPSS was calculated over the 36 stations for which we have measurements. Unfortunately, only at one station out of the 36 stations a flood event occurred, hence, if this analysis intends to evaluate AFFS capability to forecast flood events we would have to limit our focus on this one station. For this station (see Figure below) the conventional ESP would have not been able to capture/forecast this flood event as it is outside of the range of the climatology build based on the 20-year long-term run. This was also confirmed by the limit of predictability and the CRPSS which are – for this particular location – much better for AFFS than for the ESP.



AFFS capability to reproduce the general streamflow is of minor interest as AFFS is not designed as a seasonal or long-term streamflow prediction system. Of course it is to expect that in this case the Limit of predictability and the CRPSS are slightly better using the conventional ESP. This however, is not surprising, but a direct consequence of the nature of the ESP being based on 20 years and not on only one year as the AFFS and hence, it reproduces the climatology better. Additional, five of the 20 years were used during calibration to optimize the performance of LISFLOOD, calculating a skill over the same time period and locations contribute to a skewed perspective on the results. However, regardless of those two concerns, the main reason for not including the conventional ESP is that showing these figures would not add any value to the evaluation of AFFS as flood forecast system, which is the main objective of this study.

In any case, the skill (CRPSS) of the conventional ESP decreases similar to the skill of the AFFS with increasing lead time. Hence, the decrease in forecasting performance cannot only be associated solely due to possible inaccuracies of the ENS, but there must be other additional influencing factors – therefore we modified the part of the manuscript to:

~~"Whether the decrease in forecasting performance is caused by possibly inaccurate ENS cannot be assessed here, as the influence of the ENS cannot be filtered out.~~ As the skill of the conventional ESP (not shown here) decreases similar to the skill of the AFFS with increasing lead time, the decrease in forecasting performance cannot be affiliated to possible inaccuracies of the ENS only, but there must be other additional influencing factors. However, establishing the sources of predictability is beyond the scope of this paper, but subject of future research. However, […] "

Reply to: Reliability plots

Even though we would like to keep our main focus of the evaluation on AFFS' capability and limitations to predict flood events and not that much on the skill of how well it reproduces streamflow (reason: AFFS is not a seasonal streamflow prediction system; see reply to "choice of verification metrics" below), we complemented the analysis with a reliability plot (Figure 9). We chose to show the distribution of the median reliability (of the different lead times) for six different frequency of occurrences (0, 0.1, 0.3, 0.5, 0.7 and 0.9). This should provide further insight into the uncertainty of the model to reproduce streamflow.

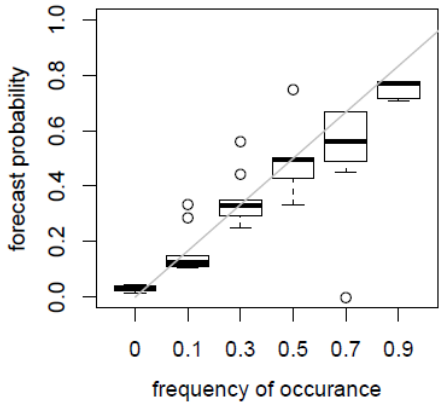


Figure 9: Reliability diagram

Also we added a paragraph on page 5572 explaining the use of the reliability diagram:
How closely the forecast probabilities correspond to the actual chance of observing the event is assessed using a reliability diagram. The reliability diagram plots for various sub-groups of forecasting probabilities the frequency with which the event was observed to occur. A forecast system has perfect reliability if the forecast probability and the frequency of occurrence are equal, and the plotted points are lying on the identity line. As the CRPSS, the reliability was calculated for each lead time at the 36 key locations.

And the result on page 5575
Figure 9 illustrates the average reliability of AFFS. Each boxplot summarizes the median reliability of the 36 key stations, considering each lead time. The diagram shows that the forecasting probability increases

together with the frequency of occurrence, following closely the identity line (grey line). This indicates a good overall reliability of the forecasts. However, notable is a slight underestimation of frequently occurring events. A possible explanation is that flood events with short durations and/or small affected areas are more frequent than large-scale and long-lasting events, but at the same time more difficult to capture due to various constrains set by the resolution of the input data and model

Reply to: issues around 1-year forecast period

It would indeed be preferably to have a longerA forecasting period than one year, however there are IT capacity and data post-processing issues which made that not feasible for this study. It has to be emphasized here that unlike in seasonal forecasting we are targeting the prediction of effects of relative short durations, in the magnitude of days to weeks. Also, unlike in some seasonal forecasting studies our forecasting periods overlap each day about 9 days, with a fresh set of ensemble weather forecast each day, which is not the case in seasonal forecasting, where the update is processed on a weekly basis or longer. So in fact we would have the same number of data points as if we were to calculate 10 years of forecast without overlap, which would have surely been considered as a sufficient number of data points for calculating statistics. Lastly, it has to be recognized that our model is running for the entire African continent on a relatively high resolution of 0.1 degree, 25 times more computational effort than a 0.5 degree model. This is – speaking again in data points – comparable to a global system running on a courser resolution (e.g. 1 degree). To sum up, considering the temporal magnitude of the phenomena to predict, the extend and spatial resolution of the system, as well as the careful selection of the evaluation period (we did not randomly decide for a year, but the decision for that specific year is the result of prior analysis on the flood situation in Africa between 2001 and 2010), we reckoning a continuous 1-year forecast as a representative period for drawing valid conclusions. However, we do acknowledge that a longer period of study should be considered in follow up research.

Reply to: choice of verification metrics

The main target of this study it the evaluation of the system's capability to predict flood events, hence the main focus on evaluation is on this aspect. As unfortunately we don't have any/ sufficient hydrological measurements of flood events for evaluating AFFS flood forecasting capabilities we had to revert to reported information. For this reason we merged all-accessible flood-related information from various trustful disaster databases to build a reference that we used for evaluation of flood forecasts. We used a contingency table and related verification metrics (FAR, POD and CSI) to quantify the performance. This is the best possible, given the limitations.

As already mentioned above, AFFS capability to predict streamflow is of minor importance as its capability to predict flood events is not depending on that. However, to give the reader a brief indication about AFFS capability to reproduce the general streamflow, we included the CRPSS and the limit of predictability, which are both standard skill scores for the evaluation of probabilistic flood forecasting and have been used widely in previous studies of similar nature.

Upon repeated requests we are also including reliability plots as it will provide the reader with further information on the system's capability to predict the general streamflow (not main target of the system).

To settle the issue of further evaluation of the general streamflow we modified the introduction to emphasize the main target of the evaluation:

The aim of this study is to investigate the ~~predictive~~ capability of AFFS to predict flood events in order ~~and~~ to ~~estimate~~ derive its potential as operational flood forecasting system that could in future contribute to the reduction of flood-related losses by providing national and international aid organizations timely with crucial flood forecast information. The predictive capability is assessed in a hindcast mode. For every day of the flood-intense year of 2003, 50 hydrological forecasts are calculated over a lead time of 10 days. Applying hydrological thresholds on the resulting ensemble of hydrological predictions, flood signals can be derived spatially. The forecasting capacity of AFFS is assessed from two perspectives: ~~its overall performance to predict streamflow, and~~ its particular ability to detect and predict flood events and its overall performance to predict streamflow. The first is of paramount importance for the assessment of AFFS as flood forecasting system as it focuses on the detection and prediction of flood events. This is done on an event-based analysis, comparing the AFFS flood signals against information collected from various disaster databases such as Dartmouth Flood Observatory, the Emergency Event Database (EM-DAT), the NASA Earth Observatory and Reliefweb to determine the number hits, false alerts and missed alerts as well as the Probability of Detection (POD), False Alarm Rate (FAR) and Critical Success Index (CSI). ~~Lastly~~ Further, to illustrate the flood forecast performance of AFFS and also to give an example of its potential output, the hindcast for the March 2003 flood event in the Save Basin is presented in detail. ~~The two analyses are complementary in disclosing the strength and shortcomings of AFFS.~~ The second is of minor importance for the assessment of AFFS as it is not focused on the prediction of flood events in particular. However, for the sake of completeness, a basic insight into the prediction of the general streamflow is given. This is done by calculating the Continuous Rank Probability Skill Score (CRPSS), a statistical indicator for probabilistic forecasts, in combination with the limit of predictability and reliability plots, for 36 key locations across Africa to gain an understanding of the general accuracy and the reliable time span of the streamflow forecasts. The two analyses are complementary in disclosing the strength and shortcomings of AFFS.

3) The setup of the experiments are unclear. It is unclear what data is used when and for what. This has also to do with the structure and I advise to restructure the paper more inline with other articles Intro-Material & Methods (including data, model, experiments both calib/hind, verification both calib and hindcast), Results and Discussion and Conclusions) this will help to avoid restating which data set is used for what etc (this could be done in 1 table) because this is rather unclear and it will help to separate results from the experimental approach etc.

We agree. The paper was restructured as following:

1. Introduction
2. Material and Methods
   - 2.1 Study area
   - 2.2 Data
     - 2.2.1 Hydrological reference data
     - 2.2.2 Meteorological data
     - 2.2.3 Other data
   - 2.3 African Flood Forecasting System (AFFS)
     - 2.3.1 Structure and functionality
     - 2.3.2 Hydrological modelling framework
       - 2.3.2.1 LISFLOOD
       - 2.3.2.2 Calibration
     - 2.3.3 Test: pan-African hindcast
   - 2.4 Verification
     - 2.4.1 calibration
     - 2.4.2 hindcast
       - 2.4.2.1 General streamflow
       - 2.4.2.2 Flood events
3. Results
   - 3.1 Model calibration
   - 3.2 Hindcast verification
     - 3.2.1 General streamflow
     - 3.2.2 Flood events
4. Discussion and Conclusion

4) For calibration it is ERAInterim +GPCP (page 5568):
For the hindcast in 2003 it is EARInterim+GPCP (page 5570) with some strange sentences: "The initial hydrological conditions, i.e. all state variables, were computed for each forecasting date between 1 January and 31 December 2003 by running LISFLOOD with near real-time meteorological observations. During the hindcasting period these needed to be approximated by using the daily GPCP-corrected ERA-Interim; however, during real-time forecasting, the first day of each ECMWF deterministic forecast could be used."

However on page 5566 it is stated:
State variables are calculated for each 0.1 pixel by forcing LISFLOOD with the near real-time meteorological observations over the forecasting period (here: 1 January–31 December 2003

This is also present in the conclusions where suddenly SRFE is being used (5578 line 17)

Thank you for this very valid objection; as written it is in fact not clear / coherent. With the restructuration of the article there is a whole Section (Section 2.2.2) dedicated on the description of the meteorological input data.

2.2.2 Meteorological data

Two meteorological data sources were used: ERA-Interim GPCP-corrected and ECMWF-ENS. Technical specifications are given in Table~2. The first were used as historical meteorological data during the model calibration as well as near real-time meteorological data for the calculation of the initial conditions. The second, the ensemble meteorological forecasts, were used for the calculation of the hydrological forecast, i.e. hindcast.

To use the ERA-Interim GPCP-corrected as a proxy for near real-time meteorological data is only possible in a hindcast mode; however, during real-time forecasting, the first day of each ECMWF deterministic forecast could be used.

Table 2: Meteorological input data

	ERA-Interim	ECMWF-ENS
provider	European Centre for Medium-Range Weather Forecasts	
spatial coverage	Global	Global
temporal coverage	Since 01.01.1989	Since 01.01.1990
spatial resolution	T255 (~80 km)	T639 (~28km day 1-10), T319 (~50km, day 11-15)
temporal resolution	6 h	1-12 h (variable temporal resolution)
brief description	Precipitation is estimated by a numerical model based on temperature and humidity information derived from assimilated observations originating	Precipitation is estimated by a numerical model, For a detailed description of the current model see http://old.ecmwf.int/research/ifsdocs/CY40r1/

	from PMV data and in situ measurements	
Reference	Dee et al.~(2011)	The ECMWF-ENS is continuously upgraded. Details as well as a description can be found on http://old.ecmwf.int/products/changes/

Further, corrections were made on page 5570:

The initial hydrological conditions, i.e. all state variables, were computed for each forecasting date between 1 January and 31 December 2003 by running LISFLOOD with ~~near real-time meteorological observations. During the hindcasting period these needed to be approximated by using~~ the daily GPCP-corrected ERA-Interim~~; however, during real-time forecasting, the first day of each ECMWF deterministic forecast could be used~~.

and on page 5566:

State variables are calculated for each 0.1 pixel by forcing LISFLOOD with the near real-time meteorological observations (for this study ERA-Interim GPCP-corrected is used as proxy; see Sect.~2.2.2) over the forecasting period (here: 1 January–31 December 2003).

and page 5578:

[…] a) the limited precision given by ~~SRFE~~ the meteorological input data to capture small-scale meteorological events accurately in the correct time and place, and b) […]

5) In the manuscript you refer to 5 data sources (which are not explained in 4.1):
Page 5565: The five main data sources on which AFFS relies are: historical hydrological observations, historical as well as near real-time meteorological observations, real-time meteorological forecasts and an African GIS dataset. Specifications of these are given in Sect. 4.1.

In the end only ERAInterim/GPCP is used and it is suggested to use EMCWF first day as proxy for this …????

If the same flow of data is not used operationally, it is needed to state in the conclusions and abstract that the operational results are probably less good than what is presented here in this paper

Corrected. A separate section was included for the meteorological input data (Section 2.2.2; see above reply to 4)) and also for the African GIS dataset (Section 2.2.3; see below "Other data"). The hydrological input data were already described (now Section 2.2.1). Further, we linked to the respective Sections on page 5565.

2.2.3 Other data

Information on topography, river channel geometry, land use, soil and vegetation properties were extracted from different data sources such as the Harmonized World Soil Database 1.0, the VGT4AFRICA project or the SRTM. A list of all the required input maps is given in Burek et al. (2013) and a more detailed description of the source of the input maps for Africa is specified by Bodis (2009). In the following we refer to this collection of thematic layers as the African GIS dataset.

Adjustments on page 5565:

The five main data sources on which AFFS relies are: historical hydrological observations (see Sect. 2.2.1), historical as well as near real-time meteorological observations (see Sect. 2.2.2), real-time meteorological forecasts (see Sect. 2.2.2) and an African GIS dataset (see Sect. 2.2.3). ~~Specifications of these are given in Sect. 4.1.~~

In the updated manuscript the issue related to the choice of meteorological input data as proxy for the near real-time meteorological input data is clearer described (see Section about the meteorological input data):

"To use the ERA-Interim GPCP-corrected as a proxy for near real-time meteorological data is only possible in a hindcast mode; however, during real-time forecasting, the first day of each ECMWF deterministic forecast could be used."

We also included a statement in the discussion (page 5578) about this issue:

It has to be noted here, that the performance of AFFS in an operational mode might differ from the one evaluated here. This is due to the meteorological input data used for the calculation of the initial conditions which are different during hindcasting and operational forecasting (see Sect. 2.2.2). Along the same lines, one might raise concern about the FAR, …

6) Page 5572: "For the calculation of the CRPSS, discharges were normalized to remove possible systematic biases, while the seasonal mean was used as benchmark." This is stated without any explanation, how it impacts the results and it is unclear for me and many other readers what the impact of this choice is.
I think it is good that some other simpler metrics like reliability plots or simple obs versus ens forecasts for various leadtimes at various locations this will give more insights for many folks instead of something that is rescaled.

Thank you for this comment, it is actually a mistake, we did not normalize the discharge. However, we normalized the CRPS (→ which is the CRPSS) as suggested by Trinh et al. (2013) to make the resulting values comparable across different catchments. Seasonal mean might also be a vague term, we used a moving average of 30 days before and after the respective day. We added some clarification on page 5571, l. 15 about the normalization, removed the wrong statement that the discharge was normalized during the calculation of the CRPS (p.5572, l. 2), and also added a definition of the seasonal mean (p.5572, l. 3)

… using the Heaviside Function (Hersbach, 2000). It is necessary to compute the CRPSS rather than the CRPS, as the latter one is depending on the magnitude of discharge and as such does not allow spatial comparison across different catchments. To circumvent this issue a normalized version of the CRPS is necessary (Trinh et al., 2013), for which reason the CRPSS was computed. Values of the CRPSS range from …

For the calculation of the CRPS~~, discharges were normalized to remove possible systematic biases, while~~ the seasonal mean (here: moving average considering 30 days before and after the respective observation) was used as benchmark. The CRPSS for the lead-times of …

We are conservative about including plots showing obs versus ens forecast for various reasons:

1. *At the locations where we have ground observation, no flood event occurred, hence these plots would only show general streamflow without any flood event. AFFS is a flood forecasting system not a seasonal or long-term streamflow prediction system, hence showing these figures would not add any value to the evaluation of AFFS as flood forecast system.*
2. *Also, even if we would include such a plot and it would show that there is an offset between obs and sim this would give no indication of AFFS performance to predict flood events, as the system is working with the exceedance of critical thresholds.*
3. *The choice on which plot to show would be very prone to be subjective. We are calculating every day a 10-day forecast, hence 365 10-day forecasts. We have 36 ground observations, hence we would have to make a decision which of the 365*36 plots to show.*

(As already mentioned in the response to concern 2)

Yes, we would like to give the reader a brief indication about the general performance system for which reason we have included the CRPSS and the limit of predictability and we are also including as suggested

the reliability (see reply to issue 2) above). However, we would like to keep the main focus of the analysis on the evaluation of the system's capability to predict flood events, which is the main target of the article.

7) I find the first part of the discussion and conclusion not up to HESS standards (5577/5578) which is a scientific journal and not a venue to do promotion please rewrite and keep it to the science (and I think the statements on page 5568 line 8-12 are not related to the FAR).

Reply to: venue of promotion

The first two paragraphes of the discussion and conclusions reads as:

"The predictive capability of the African Flood Forecasting System (AFFS) was investigated in a hindcast mode to estimate its potential as an operational flood forecasting system for the whole of Africa.

AFFS detected correctly the majority of reported flood events. The system showed particular strength in predicting riverine flood events of long duration (>1 week) and large affected areas (>10 000km2). This type of flood has the capacity to impact the socio-economic structures of a country to the extent that it might cause setbacks in the country's development (UNCSD Secretariat, 2012; United Nations (UN), 2005). The example of the flood forecast for the Save River demonstrated the precision of AFFS, gave an example of the output products that could provide the end-user with clear and concise information about the possible future hydrological situation, and showed that AFFS is capable of producing flood warnings even in ungauged river basins, i.e. in river basins where no observations are in the public domain. Hence, AFFS demonstrated a good potential to predict large-scale and long duration flood events well in advance."

We don't agree on this point. This is not promotion, the first paragraph is a statement of fact giving the motivation of this work, while the second paragraph provides a summary of the main findings and highlights the strength of AFFS, hence the results/conclusions of this study.

Reply to: statement on page 5568 line 8-12

We presume that it refers to page 5578 line 8-12, which read as: "Second, AFFS is a probabilistic flood forecasting system and as such it gives the probability with which a flood event might happen; i.e. a flood that is predicted with a probability of 70% should (ideally) also occur in only 70% of cases and not in all. Hence, the user has to keep in mind the difference between a deterministic and probabilistic forecast while interpreting the results."

Yes, we agree, this statement cannot be directly linked to FAR. It might explain the FAR, but we cannot conclude that out of the analysis we did, therefore we removed it.

Minor:
I find the phrase "The predictive capability of the African Flood Forecasting System (AFFS) " reoccurring several times please adjust this it is not necessary to repeat this every time, once should be enough etc.

We went through the manuscript and sorted this issue out.