NHESS-2020-122: Are Kenya Meteorological Department heavy rainfall advisories useful for forecast-based early action and early preparedness?

Response to review RC2

Original comments are duplicated below **in blue**, with our responses following in turn.

This short paper gives an insight into how flood warnings are generated at the Kenyan weather service and how their skill evolved over the last 5 years. Despite the relatively small number of cases and some data inhomogeneity, I find the paper useful for practitioners and generally welcome publication of such work. Overall the paper is well written and logically structured. There is, however, substantial room for improvement with respect to data and the evaluation methodology as detailed in the following major and minor comments.

We thank the reviewer for their support of our work and their useful feedback. In order to meet the request for improved data and evaluation we plan significant additional analysis for a revised version of the manuscript. In particular we will use new datasets created by the Kenya Red Cross Society (KRCS), which have been generated since submission of the manuscript.

Firstly we will use ward-level data on population exposed to flood risk instead of total population in order to improve estimates of the scale of preparedness implied by each advisory.

Secondly we will use a county-level database of all reported flooding between 2015-2019 to complement the verification of advisories against EM-DAT. This new dataset contains 184 unique days with reported flooding, which will enable us to evaluate skill statistics more robustly.

A full description of additional planned changes and our responses follow.

MAJOR COMMENTS:

1.) Evaluation procedure: Classically one would consider hits, false alarms, missed events and correct non-events. This would enable the computation of all the classical scores such as Proportion Correct, Heidke Skill Score etc. Your analysis gives a good idea of hits and false alarms but the missed events are only treated with respect to the 7 flood cases from the EM-DAT database. Can you not use CHIRPS to give some idea for missed heavy precip events that you could define to have a certain intensity and spatial reach (as pointed out in Point 2 of Reviewer 1)? After that, all days that remain would be correct negatives. This would allow a more quantitative treatment of skill.

Since preparation of the manuscript, KRCS have carried out work to identify all reported flooding events in Kenya. We have secured the use of this dataset for inclusion in a revised analysis. The conditions for inclusion of flood data are less strict than EM-DAT and so many more events are included: for our study period the dataset contains a total of 184 unique days of recorded flood events. We will use this data to carry out additional analysis of hits and misses as is suggested.

In response to the question of using CHIRPS to define misses; this faces a challenge in defining a consistent event across advisories (which vary in forecast window) and across counties (which vary considerably in size). This is discussed in the manuscript in section 2.2. We had previously investigated the possibility of requiring a certain intensity and spatial reach in observations (e.g. requiring a certain area to receive significant accumulation). However the definition is somewhat arbitrary and different choices of event definition lead to quite different results, without a clear way to objectively choose between them. For this reason, we instead have used the observed record of flooding to define hits as it is less ambiguous. Given the extended flood database we mention above, we will now be able to more robustly calculate the hits and misses as suggested above.

2.) Language: Overall the paper is nicely written and the level of language high. However, some passages are a bit wordy and redundant and I would therefore ask the authors to careful assess the potential for shortening. Given your overall low levels of statistical significance, I would also be a little more cautious with statements on skill throughout the text.

We will reduce redundant text and ensure our statements on skill are consistent with the results presented in a revised version of the manuscript.

3.) Abstract: In its current state the abstract does not really explain well what the paper is all about and in what way it is important, new and special. There should be more information on data, method, results and limitations.

We will increase the information content in the abstract in a revised version of the manuscript.

4.) Rainfall data: This is always an issue. There are many different products with strengths and weaknesses. Please provide more evidence that CHIRPS is a good one (the best?) to use and possibly repeat exercise with an alternative source of information.

We use CHIRPS as it takes advantage of satellite coverage, whilst 'ground-truthing' against station records. CHIRPS compares favourably to other satellite-based datasets over East Africa (Dinku et al 2018).

In addition, a global evaluation of 22 rainfall datasets (Beck et al 2017) recommend the use of CHIRPS in particular for tropical regions. Beck et al 2017 note difficulty in providing reliable recommendations in regions such as Africa where rain gauge data is limited. However for Kenya in particular the station density used in CHIRPS is relatively good (see for example `https://data.chc.ucsb.edu/products/CHIRPS-2.0/diagnostics/chirps-n-stations_byCountry/Kenya/Kenya.1981.01.png`).

Weaknesses with CHIRPS include spurious drizzle and an underestimation of peak magnitudes (Beck et al 2017). Our analysis is likely to be insensitive to spontaneous drizzle, although an underestimate of peak rainfall implies a conservative bias to our evaluation of the skill of advisories for predicting threshold accumulation (e.g. figure 4b, figure 6). Here the advisory area receiving threshold accumulation may be higher than CHIRPS suggests. Having said this, Beck et al 2017 find

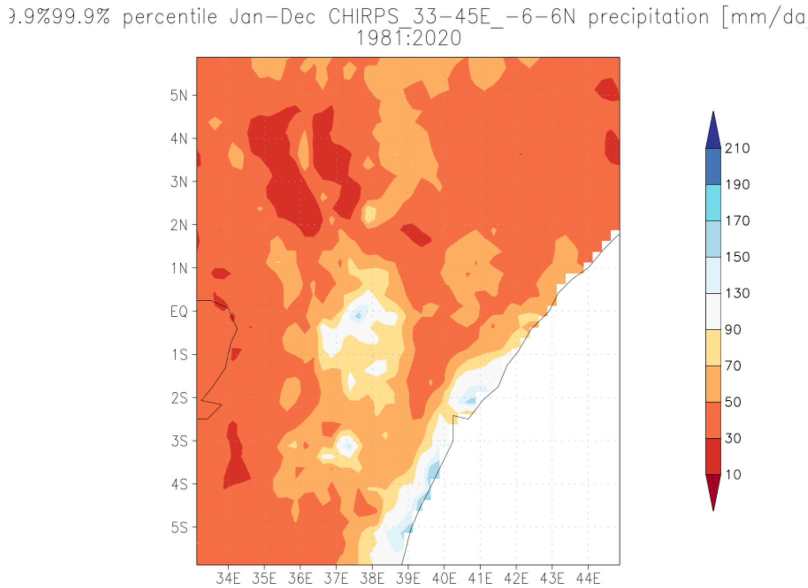

Figure 1: 99.9th percentile of CHIRPS daily rainfall over Kenya, from Climate Explorer

that the most extreme rainfall (e.g. 99.9 percentile of daily rainfall) shows most underestimation. Analysis shows that a 99.9 event over Kenya in CHIRPS ranges from 30-130mm per day (figure above), whilst our focus is on multi-day accumulations of 25, 50, 75 and 100mm, suggesting that CHIRPS estimates of totals close to our thresholds of interest are less affected by underestimation compared to the highest magnitudes.

Overall we will add a justification of the use of CHIRPS in the manuscript and following the above. However do not feel that the use of an additional precipitation dataset for verification would bring more robust results. In particular, our analysis ultimately moves beyond uncertainty in rainfall observations by focusing directly on specific flood events (EM-DAT, and the additional flood record from KRCS, see point one above).

Beck, H.E., Vergopolan, N., Pan, M., Levizzani, V., Van Dijk, A.I., Weedon, G.P., Brocca, L., Pappenberger, F., Huffman, G.J. and Wood, E.F., 2017. Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. Hydrology and Earth System Sciences, 21(12), pp.6201-6217.

Dinku, T., Funk, C., Peterson, P., Maidment, R., Tadesse, T., Gadain, H. and Ceccato, P., 2018. Validation of the CHIRPS satellite rainfall estimates over eastern Africa. Quarterly Journal of the Royal Meteorological Society, 144, pp.292-312.

5.) Section 2.2: I think that the approach you are taking is largely well conceived (but note my reservations under Point 1) given all the restrictions at hand but the section as written is quite long and your quantitative metrics are only described and nowhere cast into formulas. I suggest giving this section a clearer structure and a more "recipe like" description of how you compute metrics. If you give names or abbreviations to your metrics, you would not need to repeat the description again in Section 3.

We will review this section in a revised version of the manuscript and attempt to reduce unnecessary detail. However we do feel that given the particular challenges to verification, some space is needed in the manuscript in order to motivate and justify why we are unable to follow a standard approach. We are not convinced it will aid readability to introduce equations to abbreviate the metrics we use.

6.) EM-DAT: I find the thresholds of 10 deaths too high and would feel that even one death would justify a weather warning. Given that you have authors from Kenya that may have access to government documents, is there no alternative source of information that would give you a list of flood events of smaller magnitude, too? This would much improve your statistics relative to the few events in EM-DAT!!

See point one above: we plan to evaluate the advisories against a larger record of flood events provided by KRCS.

7.) Population numbers: I agree with Reviewer 1 that a distinction between all population of a county and the fraction likely affected by floods (in particular riverine) would be desirable. However, I can imagine that such fractions are not easily available and feel that the paper would be of value without it. In this case the authors could raise this point more clearly in the text and give at least some orders of magnitude from literature.

Given this comment and the request from Reviewer 1, we plan to improve this part of the analysis by using a dataset provided by KRCS, who have recently carried out analysis of exposure to riverine flooding at ward level as part of the IARP project. In particular we will use the data for population exposed to a 5 year return period flood (a 5 year event is the focus of the development of flood preparedness triggers in IARP). The data itself has been created by KRCS by integrating inundation areas estimated by ECMWF using GLoFAS with ward level population data.

Using flood exposure data will provide a much more realistic estimate of the scale of potential intervention implied by each advisory. We note that it will only consider preparedness actions aimed at the population exposed to riverine flood and not those exposed to flash flooding or landslides. We will add a discussion on this point.

MINOR COMMENTS: 1.) Punctuation: There are a lot of places with inconsistent or suboptimal use of commas. Please check carefully throughout the entire manuscript. 2.) L2: remove "a" as in plural 3.) L5-6: What are you trying to say with this sentence. Please reword! 4.) L12: no comma 5.) L19: is it really a "movement"? In L31 it is called a "society"? 6.) L30: IFRC? 7.) Section 1: this gives a nice introduction to the topic but some bits are a little redundant and could be streamlined. 8.) L75-76: avoid repetition of "improve" 9.) L120: remove period after figure 2 10.) L125: better turn this into a proper sentence 11.) L167: this question? 12.) L194: requires? 13.) Table 1: Why don't you merge the first two entrances? 14.) L234: fell during . . . 15.) L241: "quite a reasonable chance" is very fuzzy, reword! 16.) L245-248: What result or figure does this paragraph refer to? 17.) Figure 4 could be discussed in a little more detail. 18.) Figure 5 I would rather include in the Methods section 2. You can then also discuss there the difference between all people and those affected by a given flood (see above). 19.) L255: remove "extreme" as upper bound is already an extreme 20.)

L286: highest number? 21.) L319: on 18th November? 22.) L385: I would maybe not use the word "all" here, as it remains a probabilistic problem, where some missed events are unavoidable. 23.) L441: double period 24.) L443: comma instead of period 25.) L456: 2x would 26.) L458-59: not a proper sentence 27.) Section 4.2.2: Too much detail to my taste. This is a scientific paper and not a government technical document. 28.) Figure 2 caption: include that these statistics are done for the cases listed in Table 2. 29.) Figure 4 caption: these should be 5kmx5km gridpoints 30.) Figure 5 caption: two brackets at end 31.) Figs.: I would generally not start a caption with a question.

These will all be addressed in a revised version of the manuscript. Though regarding too much detail in section 4.2.2 (point 27), we feel that provides the broader institutional context of developing hazard early warning systems in Kenya and will be of interest to (at least some) readers of the work, and fits within the scope of NHESS.