Response to reviewers
Manuscript for Hydrology and Earth System Sciences
Manuscript number: HESS-2018-164
Title: The potential of global re-analysis datasets in identifying flood events in Southern Africa
Authors: Gründemann, G.J., Werner, M., Veldkamp, T.I.E.

Referee #2:

This study aims at evaluating the performances of several hydrological models in identifying flood events in South Africa. The models considered are members of the Water Resource Reanalysis (WRR) developed in the European Earth2Observe project. Models performances are evaluated using a frequency analysis and several skill scores related to flood detection. The authors also provide an interesting comparison with damaging events reported in three different disaster databases. Results convincingly show the ability of such models to capture the majority of flood events, despite their coarse resolution. Performances vary from one model to another, due to differences in model structure. The authors also pointed out the improvement due to the increase in spatial resolution (from 0.5_ in WRR1 to 0.25_ in WRR2). Before concluding, the authors discuss the main limitations of the method. The conclusions are consistent with results presented all along the manuscript. The paper is well written and organized. This is a really interesting study and I think that the manuscript would be ready for publication provided that the authors address the following comments.
We thank the anonymous referee for his/her time in carefully reviewing our manuscript. The referee provided us with helpful comments, which will improve our manuscript. We are glad that the referee finds our work of importance for the scientific community to consider publication in HESS after revisions. We have carefully studied each of the remarks and will address them here in detail. We have copied the referee comments (in bold), and respond to each them below.

Major comment:
My major comment concerns the spatial resolution of the models. First, the authors often attribute the improvement in flood detection to the increases in spatial resolution. But in this case, the improvement can be due to three factors: (1) the improvement of models forcings: WFDEI (used in WRR1) and MSWEP (used in WRR2) rely on different methodologies (2) new model developments: each modeller involved in E2O included new developments in the models (e.g. multi-layer snow scheme in HTESSELCaMa, groundwater abstraction in LISFLOOD, reservoirs and water withdrawals in PCR-GLOBWB, aquifers and floodplains in SURFEX-CTRIP, etc.) (3) meteorological and hydrological processes better represented at higher resolution Many studies showed that simulated discharges are highly sensitive to meteorological forcing, especially precipitations, which are supposed to be of better quality in WRR2. Although points (1) and (3) are briefly mentioned in the discussion section (P13L2-5), I think they should be mentioned in the section presenting the models (section 2.2). Also all the conclusions on the differences in WRR1 and WRR2 performances should be put in this context. To have an idea of the impact of points (1) and (2) (without point (3)), the authors could consider the WRR2 version of the SURFEX-TRIP model which used the improved forcing (MSWEP) and new model developments but a spatial resolution of 0.5_ for the routing scheme.
We thank the referee for raising this interesting comment. We were actually considering to include SURFEX-TRIP in our initial analysis, but decided against it due to possible inconsistency reasons. All model outputs in WRR2 are available at 0.25 degrees, except for the river discharge of the SURFEX-TRIP model as the same routing scheme used in WRR1 as applied. We therefore decided to focus solely on the models that did meet the standards agreed upon by EartH2Observe.

However, as this remark has been raised, we took a look at the differences in the error statistics of SURFEX-TRIP in WRR1 and WRR2, as can be seen in the table below that is part of Table 4 in our manuscript. As can be seen, there is an improvement of WRR2 compared to WRR1, and we thus have

decided to include SURFEX-TRIP WRR2 at 0.5 degrees in our revised manuscript. We will perform further analysis and based on this new information revise our manuscript, put the differences between WRR1 and WRR2 in the context proposed by the referee, and update Figures 2, 4, 5 and 6 specifically.

	Stations with upstream catchment area [km^2]	SURFEX-TRIP	
		wrr1	wrr2
NSE	≥ 4	-2,938.32	-1,147.33
	≥ 520	-31.33	-21.73
	≥ 2,500	-0.54	-1.21
PBIAS	≥ 4	-2,415.21	-1,176.01
	≥ 520	-243.26	-167.95
	≥ 2,500	-57.74	-74.57
r	≥ 4	0.45	0.50
	≥ 520	0.50	0.56
	≥ 2,500	0.52	0.60

Another problem related to the spatial resolution is the selection of gauge stations (section 2.3.1). The authors mention that "the stations have upstream catchment areas that vary between 4 and 342,000 km2)". Is it realistic to compare observed and simulated discharges at stations with drainage area that small? Given that a model pixel has an area of approximately 2500 km2 for WRR1 and 650 km2 for WRR2, rivers with drainage area smaller than these thresholds are generally not represented in the models and the correspondence between stations and model pixels is necessary wrong. This is consistent with authors results (P8L25-27, P12l30-32). This remark is mentioned P9L3-5, and in my opinion, stations with small drainage area should be excluded, even though there would remain a small number of stations. Also, could the authors give some details on the method used to select the model grid cell corresponding to each station? Was the river network of each model used to associate each station to a model grid cell?

We thank the referee for this comment, which was also mentioned by the first referee. Our answer is similar and is copied below. The reason we have decided to include the smallest sub-catchments as well, is because part of the scope of our paper is to determine the area up to which the models are still able to represent the hydrological behaviour. As shown in the results presented in Section 3.1 (page 8 line 25), the models were shown to capture the hydrology of the river for sub-catchments that are larger on the order of 520 km^2 for WRR2, which is on the order of the 625 km^2 of the cells size of the 0.25 degree models . Including the smaller catchments provides information about the scale up to which we can use such models that are included in the WRR dataset. This has, to our knowledge, not been studied before it is valuable for the scientific research community. Particularly since there are many regions that have to rely on such global data as the observational data is insufficient for localized models. Furthermore, as is for instance shown by Figure 3 where we analyse the Spookspruit river, with a sub-catchment area of only 252 km^2, global-scale models do actually have skill in capturing the variability (indicated by MM1 and MM2), even though the actual values are indeed overestimated. We will, however, rephrase this in our revised manuscript as this was not entirely clear. Since both referees raised this point, we are nevertheless willing to disregard the smallest stations from our analysis. We are therefore interested in the opinion of the editor regarding this issue. We also consider of scientific interest that for both the coarser and the finer resolution models, the threshold is on the order of the cell size. This holds promise for the continuing effort of modelling research groups in developing increased resolution (global models).

Concerning the method we used to select the model-grid cell for each station, we looked at the location of the model and selected the exact model-grid cell that it was located in. Unfortunately we did not have any access to the river network of each model, so we couldn't use this to associate each station to the exact model-grid cell. We did, however, check the eight surrounding cells and select the cell with the highest discharge as the river cell corresponding to the discharge gauge. This did work for larger basins, but it did not for the smaller ones if there was a large river in a neighbouring cell.

Minor comments:

P3L31: "one of the largest basins"

Thank you, we will modify this sentence.

P4L15-16: Have the impacts of such modifications on floods been studied already?

Thank you, there has not yet, to our knowledge, been a study into the impacts of these types of human modifications in the Limpopo Basin in terms of floods. Previous studies have, however, looked into the impacts of human modifications such as dams in other rivers, such as the study by Fitzhugh and Vogel (2011) for the United States specifically.

P5L11: Please provide a reference for "the Kolmogorov Smirnoff statistic for the Gumbel Extreme Value Distribution".

The reference will be provided.

P5L21: Please add a few words to explain what the criteria from NatCatSERVICE is.

Thank you for your comment. In order to determine the severity of a disaster event NatCatSERVICE uses the number of fatalities and overall losses as criteria. We will add this in our paper.

P6L19: In my understanding, hydrological extremes include floods but also droughts. This study only focuses on floods. The authors should carefully revise the use of "extremes" throughout the manuscript (including the following subtitle).

We thank the referee for pointing this out. Indeed we have used the term "extremes" mainly regarding floods in our paper. We will revise our use of this term throughout the entire paper.

P7L12: "…for both the observed and the modelled discharges..."

We will change this in the final version of our paper.

P8L14: The areas mentioned in L13 are mainly related to the models spatial resolution. This should be specified.

Thank you for pointing out that this was not clear, we will clarify this further.

P8L25-27: My guess is that the difference between WRR1 and WRR2 performances mostly comes from the forcings (rather than from the resolution).

We thank the referee for this interesting comment. We have taken a closer look at the causes of the difference in WRR1 and WRR2. Indeed the different forcing used in WRR2 could be one of the reasons. However, model improvements and the improved resolution are likely to contribute to the improved performance. That model structure and resolution does influence the results can be seen by how improved forcing leads to differing changes to model performance, depending on the model. To unravel the influence improved forcing and model improvements, the higher resolution models could be re-run with the (resampled) low resolution forcing data. While this will reveal no doubt interesting results we consider this as outside the scope of the current paper.

P8L27-29: This result is hardly visible from Figure 2. It is more evident from Table 4a.

Thank you for pointing this out, we will include explicit reference to the table.

P9L8-11: I think this kind of conclusion should be built on larger basins only. Also, a fair comparison would only consider WRR1 so that the influence of forcings are excluded.
Thank you for raising this issue. We shall base this conclusion solely on the largest stations. Additionally, we actually did only consider the models in WRR1 for this conclusion, but we apologise if that was not clear. We clarify this in the revised version of our paper.

P10L9: "…were established using the observed climatology."
Thank you, we will modify this sentence.

P10L26-27: These differences in performance are also (mainly?) the result of the improved forcings.
Same as for the first major comment, we will revise this.

P12L5: "…and was evaluated by means of commonly used error statistics..."
Thank you , we will alter this.

P12L17-24: Would it be possible to get the dates of construction of major dams or reservoirs? It could be interesting to look at models performances before and after these dates.
We agree with the referee that this would indeed be interesting. In order to analyse the hydrologic impacts of dams discharge statistics from periods before and after the construction of the dam need to be analysed. It is recommended to have 20 years of data both pre- as well as post construction of the dam (Huh et al., 2005). As the WRR dataset is available between 1980 and 2012, this would mean that at maximum 16 years of data would be available, if the dam was constructed in the middle of the period of record (1996). We have analysed data on dams in the basin and the dates of commissioning. From the available information we found that o major dam was completed in 1996 or the three years before and after. The closest major dam completions are 1987 (Flag Boshielo Dam in South Africa) and 2000 (Letsibogo Dam in Botswana). We therefore think that there is insufficient data to do a robust trend-analysis. Furthermore, not all models include reservoirs, and those that do, do not do so in the same manner. Some models have "static" reservoirs, whereas others include the reservoirs over time. Therefore, doing an integrated analysis for all models equally is beyond the scope of our paper, and would be a separate study.

P12L30: Improvements are also due to forcings improvements and new model development.
Same as for the first major comment, we will revise this.

Figure 1: Please remove the marks within the inlet map. Also, please explain what the square and the circle represent (something like "stations used to illustrate the flood frequency analysis in section 3.2.1").
Thank you, we will remove the marks and add a further explanation regarding the two stations.

Figure 2: Would the NSE results be better presented using a log scale?
Thank you for your comment, which was also raised by the other referee. We will alter the figure and make the NSE results more clear.

Figure 3, in the caption: "The index value refer to models with 0.5 degree resolution (MM1 and MO1) and 0.25 degree resolution (MM2 and MO2)."
Thank you, we will change the caption in Figure 3.

Figure 6: The figure is not clear (small points/lines, close colours). Would results be better presented using box plots with mean, median, 1st and 4th quartiles?
Thank you for pointing this out. As the first referee also pointed this out, we feel like our research will benefit from this and we will hence modify this figure.

Table 2: What are the impacts of changing theses values on the results?
Changing these values is not significantly impacting the results of this research. The only instance where these thresholds are used is for constructing Figure 3. This figure is merely an illustration of two example gauges to illustrate the difference between the model climatology versus the observed climatology. If higher thresholds would be chosen, they would not be exceeded as much, whereas if lower thresholds would be chosen, they would be exceeded more often.

Table 4, in the caption: Change "Figure 3, upper/lower panel" to "Table 4, upper/lower panel". Also, it seems that the selection of stations in each line of the table relies on a lower threshold, so the "lower or equal" sign should be "upper or equal".
Thank you for pointing this out, the caption and "lower or equal" signs were incorrect. We will modify the table and caption accordingly.

References used in this revision response:
Fitzhugh, T.W. and Vogel, R.M.: The impact of dams on flood flows in the United States, River Res. Appl., 310, 1192-1215, doi:10.1002/rra, 2011.