The potential of global re-analysis datasets in identifying flood events in Southern Africa

Gaby J. Gründemann[1,2], Micha Werner[1,3], and Ted I.E. Veldkamp[4,5]

[1]IHE Delft Institute for Water Education, 2601 DA, Delft, the Netherlands
[2]Delft University of Technology, 2628 CN, Delft, the Netherlands
[3]Deltares, 2629 HV, Delft, the Netherlands
[4]Institute for Environmental Studies (IVM), VU University Amsterdam, 1081 HV, Amsterdam, the Netherlands
[5]International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria

Correspondence: Gaby J. Gründemann (g.j.gruendemann@tudelft.nl)

Abstract. Sufficient and accurate hydro-meteorological data are essential to manage water resources. Recently developed global re-analysis datasets have significant potential in providing these data, especially in regions such as Southern Africa that are both vulnerable and data poor. These global re-analysis datasets have, however, not yet been exhaustively validated and it is thus unclear to what extent these are able to adequately capture the climatic variability of water resources, in particular for extreme events such as floods. This article critically assesses the potential of a recently developed global Water Resource Re-analysis (WRR) dataset developed in the EU FP7 eartH2Observe project for identifying floods, focussing on the occurrence of floods in the Limpopo River basin in Southern Africa. The discharge outputs of seven global models and ensemble mean of those models as available in the WRR dataset are analysed and compared against two benchmarks of flood events in the Limpopo River basin. The first benchmark is based on observations from the available stations, while the second is developed based on flood events that have led to damages as reported in global databases of damaging flood events. Results show that while the WRR dataset provides useful data for detecting the occurrence of flood events in the Limpopo River basin, variation exists amongst the global models regarding their capability to identify the magnitude of those events. The study also reveals that the models are better able to capture flood events at stations with a large upstream catchment area. Improved performance for most models is found for the 0.25 degrees resolution global model, when compared to the lower resolution 0.5 degrees models, thus underlining the added value of increased resolution global models. The skill of the global hydrological models in identifying the severity of flood events in poorly gauged basins such as the Limpopo can be used to estimate the impacts of those events using the benchmark of reported damaging flood events developed at the basin level, though could be improved if further detail on location and impacts are included in disaster databases. Large-scale models such as those included in the WRR dataset are used by both global and continental forecasting systems, and this study sheds light on the potential these have in providing information useful for local scale flood risk management. In conclusion, this study offers valuable insights in the applicability of global re-analysis data for identifying impacting flood events in data sparse regions.

1 Introduction

Floods are among the most common and destructive natural hazards globally (Jongman et al., 2015). Approximately 90% of disasters worldwide in the last decades were caused by weather-related events. Among them, floods are the most frequent, and affected 2.3 billion people between 1995 and 2015 (UNISDR & CRED, 2015). It is generally acknowledged that due to projected climate and socio-economic changes, extreme events such as floods may further increase in frequency, magnitude and intensity (IPCC, 2012, 2014; UNISDR, 2015, 2016). In order to minimise the negative effects of floods, disaster risk reduction is increasingly important (Trigg et al., 2016). The urgency of mitigating flood risks is also recognised by international agreements, such as the Sendai Framework for Disaster Risk Reduction (UNISDR, 2015), which underlines the understanding of disaster risk including the hazard characteristics as a first priority. Developing adequate knowledge of past flood events is essential in order to sufficiently address this global problem (Dottori et al., 2016; Spaliviero et al., 2011) and to further reduce the consequences of future disastrous events.

Accurate data is key to developing a reliable representation of floods. While hydro-meteorological data are collected and made available in many places, most developing countries still struggle with limited availability due to inconsistent methodologies and datasets (Pozzi et al., 2013; Smith et al., 2015; Trigg et al., 2016). This may for example be because of the lack of rain and discharge gauges due to insufficient resources as a consequence of socio-economic issues (Hughes, 2006; Spaliviero et al., 2011). One of the regions where data availability is poor is (Southern) Africa (Kundzewicz et al., 2002; Naumann et al., 2014; Trigg et al., 2016; UNISDR, 2016). Not only is there a general lack of data, but the available data and resources are also not evenly distributed across the riparian countries, with most gauges in South Africa (Thiemig et al., 2011). While the country of South Africa is relatively rich in terms of data, technology and knowledge, many of its neighbouring countries are not (Spaliviero et al., 2011). This lack of spatially consistent datasets is a particular issue in this region, as many of the larger river basins are transboundary, and extreme events are often linked to phenomena on a wider, regional scale, such as cyclones (Biswas, 1999; Patt and Schro, 2008).

To address the issue of floods in data poor regions, increasingly available global datasets, such as global re-analysis data, may have significant potential. Re-analysis datasets are the result of a combination of earth observations, as well as various models and datasets containing in-situ measurements (Schellekens et al., 2017). Currently there are several re-analysis datasets available at a global scale and applicable to water resources, such as ERA-Interim/Land (Balsamo et al., 2015), GLDAS (Rodell et al., 2004), Global Water Cycle Reanalysis (van Dijk et al., 2014), GSWP-2 (Dirmeyer et al., 2006), WATCH (Haddeland et al., 2011) and WRR (Schellekens et al., 2017). These datasets provide consistent hydro-meteorological data with a global coverage, spanning several decades. Hence, they have significant potential to fill data gaps in regions such as Southern Africa (Sood and Smakhtin, 2015; Trigg et al., 2016; Ward et al., 2013; Wood et al., 2011). Datasets containing different global model outputs have thus far been used to determine climatic extremes as well as its uncertainties at the global or continental scale. For instance, Zhao et al. (2017) evaluated the influence of different river routing schemes in the various global hydrological models on peak discharge simulation. Dankers et al. (2014) compared the 30-year return period level of river discharge calculated using nine different global models regarding their projections of climate change impacts on flood hazards worldwide. Trigg et al. (2016) assessed the ability of six global models regarding their skill to produce hazard maps for the African continent. However, they note that there has thus far been limited validation of these global flood models against observed floods.

This study assesses the potential of a recently developed state-of-the-art global water resource re-analysis dataset in identifying damaging flood events for data poor regions such as the Limpopo River basin. The Limpopo River basin is a transboundary Southern African basin typical of the aforementioned data issues, including a general lack of data as well as an asymmetrical distribution of data availability across the riparian countries. The dataset used in this study is the open source global Water Resources Re-analysis (WRR) dataset developed in the eartH2Observe (E2O) research project, a collaborative project funded under the European Union's 7th Research Framework (Schellekens et al., 2017). The WRR dataset is assessed against two benchmarks. The first benchmark is developed using observed discharges from reliable gauges available in the region. As the upstream catchment area of these gauges varies, this provides insight into the skill of the global dataset in identifying the occurrence and magnitude of flood events in the basin, and how this skill is related to catchment scale. The second benchmark considers reported damaging flood events in the basin. Reported events from three disaster databases, including the Emergency Events Database (EM-DAT); the Global Active Archive of Large Flood Events (GAALFE); and the Natural Catastrophe Service (NatCatSERVICE), were collated to develop a chronology of damaging events. The ability of the global model datasets in identifying such damaging events provides insight into the potential of the global models to be able to predict the occurrence of impacting flood events. There is a critical need for both higher resolution re-analysis supporting data and flood forecasting systems to properly capture timing, intensity, and location of flood impacts. Global models such as those considered in this study are employed by several global and continental flood forecasting systems, such as the Global Flood Awareness System GloFAS (Alfieri et al., 2013) and the African Flood Forecasting System (Thiemig et al., 2011), and this assessment sheds light on the potential these have in providing information that is useful to managing floods at the regional and sub-basin scale. We also consider of scientific interest that for both the coarser and the finer resolution models, the threshold up to which the models are still able to capture the hydrology 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).

The remainder of this paper is structured as follows. Section 2 provides the materials and methods used, a description of the study area, data as well as verification methods. The results (Section 3) reveal the skill of the models in capturing the reported as well as modelled flood events. Section 4 provides a discussion of those results, as well as limitations and suggestions for further research. Conclusions are provided in Section 5.

2 Materials and methods

2.1 Study area

The Limpopo River basin is a transboundary river basin located in the east of Southern Africa, between latitudes 20°S – 26°S and longitudes 25°E – 35°E. With a length of approximately 4,000 km and a total drainage area of nearly 413,000 km^2, it is one of the largest basins in Southern Africa (Aich et al., 2014a; Maposa et al., 2014; Trambauer et al., 2015). The basin is shared by four riparian countries: South Africa, Botswana, Mozambique and Zimbabwe, as shown on Figure 1. The climate in the basin is predominantly dry, semi-arid and hot (FAO, 2004; Trambauer et al., 2015). The upstream part is located in the Kalahari Desert, while further downstream, the climate transitions from an arid desert to a hot and dry steppe and eventually to a dry tropical savannah.

Precipitation in the basin varies significantly and is highly seasonal (FAO, 2004). Mean annual rainfall is approximately 530 mm, ranging between circa 270 and 1,160 mm (Beck et al., 2017). Some 95% of the rainfall falls during the austral summer months between October and March, with the monsoonal rainfall events interspersed with dry spells. Precipitation events during the wet season are spatially as well as temporary isolated (FAO, 2004). The runoff ratio of the Limpopo River basin is low (Trambauer et al., 2014), which is characteristic for arid and semi-arid regions (Aich et al., 2014a), and is exacerbated in the Limpopo basin by water abstractions for irrigation and domestic use. The basin faces significant transmission losses, resulting in a decline of flow along the length of the river (WMO, 2012). Large sections of the main stem, especially near the mouth, have a dry river bed during the dry season (LBPTC, 2010). However, flood waters can rise quickly, especially in the floodplains around Chokwé in Mozambique, where the mean flood peak can raise water levels some five metres above normal levels, with levels twelve metres above normal observed during the severe floods of the year 2000 (WMO, 2012). Furthermore, the river basin has been modified to a large extent, with many dams, irrigation schemes, and storage reservoirs (Aich et al., 2014a; Ashton et al., 2001; LBPTC, 2010; Silva et al., 2010).

2.2 Input data

Input data in this research were provided by the publicly available WRR dataset that was developed within the E2O research initiative (Arduini et al., 2017; Dutra et al., 2015, 2017; Schellekens et al., 2017). This dataset includes the outputs of ten different global models that are available at two resolutions and time ranges; denoted WRR1 and WRR2. WRR1 has a 0.5 degree resolution (approximately 50 km at the equator) from 1979 to 2012 with the models forced by the Watch Forcing Data applied to ERA-Interim data (WFDEI) meteorological re-analysis dataset (Weedon et al., 2014). WRR2, on the other hand, has a 0.25 degree resolution from 1980 to 2014, and all models were forced using the Multi-Scale Weighted-Ensemble Precipitation (MSWEP) dataset (Beck et al., 2017). Apart from the different forcing and spatial resolution, the model algorithms were also improved, such as by a better representation of hydrological processes and by integrating earth observation data (Arduini et al., 2017; Dutra et al., 2017). More information on the WRR dataset and the improvements can be found in Arduini et al. (2017), Dutra et al. (2015, 2017), Schellekens et al. (2017) and Table 1.

As this research focusses on the occurrence of floods, simulated discharges of the ensemble of models included in the WRR datasets were used. Of the ten models, seven models provide daily discharge values, both Global Hydrological Models (GHMs) and Land Surface Models (LSMs). All apply different routing schemes to compute the discharges, see Table 1 for further information. The remaining three models do not include routing schemes and were therefore not considered. Discharge data for both WRR1 and WRR2 were downloaded at the locations of the river gauging stations in the model-grid. While modelled discharges were available for evenly spaced grid cells, river gauging stations are not equally distributed across the Limpopo River basin, resulting in multiple gauging stations in the same model cell in some cases. The daily modelled river discharges for each cell in the model-grid where one or multiple discharge gauging stations is located were downloaded from the E2O Water Cycle Integrator portal (https://wci.earth2observe.eu/). Modelled discharge data from the period 1980-2012 were used in this study as a common period in order to compare the differences between WRR1 and WRR2. Note that for three models simulated discharges were available for the higher 0.25 degrees resolution models, and for the SURFEX-TRIP model the discharge in WRR2 were only available at 0.5 degrees resolution (see Table 1).

2.3 Verification data

2.3.1 Discharge Data

Daily observed discharges from selected river gauging stations in the Limpopo River basin were used to verify the modelled discharges. Discharge records were collected from multiple sources and collated, including the Global Runoff Data Centre (GRDC), the South African Department of Water and Sanitation (DWAF) and the Regional Water Administration of Southern Mozambique (ARA Sul). In the entire Limpopo River basin, there are 196 accessible stations that contain data in the 1980 to 2012 time span. However, only 75 of these have daily data available for at least 25 years and passed the goodness of fit test by calculating the Kolmogorov Smirnoff statistic (Massey Jr., 1951) for the Gumbel Extreme Value Distribution (Gumbel, 1941) at the 5% significance level. These 75 stations are shown in Figure 1, and a detailed list is included in the Supplementary material (S1). The stations have upstream catchment areas that vary between 4 and 342,000 km^2.

2.3.2 Disaster data

Data from three disaster databases were compiled in order to determine a singular chronology of damaging flood events in the Limpopo River basin to be used as a benchmark: EM-DAT (CRED and Guha-Sapir, 2017), GAALFE (Brakenridge, 2017) and NatCatSERVICE (Munich-Re, 2017). This combined reference database contains the 48 damaging flood events that occurred in the basin over the time span that coincides with the period of record of the E2O dataset; from 1980 to 2012. A summary of this benchmark dataset is included in the Supplementary material (S2). To allow comparison of the reported events to the simulated and observed discharges, the severity or intensity levels of the reported damaging flood events were assessed. This was completed following the criteria from NatCatSERVICE (Kron et al., 2012), which are based on the number of fatalities and overall losses, and amended for the total number of fatalities in the entire basin. This resulted in severity levels ranging from 0 (natural events) to 5 (devastating catastrophes).

The basin is both affected by large-scale basin-wide flood events, as well as by smaller scale flood events that do not affect the whole basin at once. The three disaster databases are structured differently. Whereas EM-DAT and NatCatSERVICE report the flood events on a country basis, the GAALFE is ordered on an event basis. Apart from that, the level of detail regarding the location of where the flood took place varies, also within one database. Especially the flood events that occurred earlier often have only a broad administrative descriptions, rather than the (sub-)basin of where the flood occurred. The study area was therefore subdivided into seven administrative regions in order to be able to make a spatial distribution in areas exposed to flooding. These regions are the Limpopo basin with the riparian countries Botswana (BW), Mozambique (MZ) and Zimbabwe (ZW), and four regions within South Africa (ZA). South Africa was split into multiple regions since roughly half of the total basin area is located within South Africa, while nearly all of the available stations are within this part of the basin, allowing a higher level of detail in identifying the spatial occurrence of flood events. The four different regions in South Africa identified are the North West Province (ZA1), the Gauteng Province (ZA2), and the combined provinces of Limpopo and Mpumalanga, subsequently divided into a western (ZA3) and an eastern part (ZA4). The different regions can be seen in Figure 1.

2.4 Evaluating the model performance

2.4.1 Hydrological performance

Hydrological performance of the daily simulated discharges from all models was assessed using commonly used model evaluation statistics, considering Nash-Sutcliffe Efficiency (NSE), Percent Bias (PBIAS) and Pearson's correlation coefficient (r). For a fuller description of these statistics and their application see Moriasi et al. (2007). NSE ranges between $-\infty$ to 1, where 1 indicates a perfect representation of observed discharges, with values above zero meaning the simulated discharges have better skill than simply taking the average of the observed. PBIAS determines the tendency of the simulated discharge to underestimate or overestimate observed discharges (Gupta et al., 1999), normalised with the mean discharge. Ideal values of PBIAS are zero, with acceptable values considered to be below ±25 percent (Moriasi et al., 2007). Pearson's correlation coefficient (r) provides an indication of the linear relationship between simulated and observed discharges data. Ranging from -1 to 1, which indicate a perfect negative or perfect positive relationship respectively, a correlation coefficient of 0 shows no relationship whatsoever. Correlation coefficients are widely used to describe the proportional decrease or increase of two variables, and have the advantage to be sensitive to large values (Beck et al., 2017; Legates and McCabe Jr., 1999), which is important for analysing hydrological extremes (we use the term extremes in this paper to indicate the high river flows).

2.4.2 Hydrological extremes

Flood Frequency Analysis

Flood frequency analysis was performed in order to obtain the magnitudes of the hydrological extremes (Mujere, 2011). By fitting a Gumbel distribution using the method of moments, the daily river discharge values were converted to annual exceedance probabilities or return periods (Ward et al., 2011). This allows the occurrence and severity of flood events to be identified in both the observed and modelled discharge time series. Observed flood events were identified as events with a low annual exceedance probability (or high return period) at the river gauging stations, with discharges associated to progressively smaller probability thresholds used to identify increasingly severe flood events. Flood events in the modelled discharge time series were identified in two ways; using either the model climatology or the observed climatology. When using the model climatology, the discharge values for the selected probability thresholds were derived using the Gumbel distribution applied to the modelled discharges, providing the skill of the model in simulating the variability of extreme discharges. When using the observed climatology, the discharge values for the thresholds were derived using the observed discharges, which represents the skill of the model in determining the absolute discharges. The severity of the reported damaging flood events retrieved from the three disaster databases (Section 2.3.2) are then compared to the severity of the flood events identified in observed and modelled time series. To allow this comparison, the reported damaging flood events, the annual exceedance probabilities or return periods were converted to flood intensity levels, according to Table 2. In order to determine the possible added value of the higher resolution global models, modelled flood events were assessed both for WRR1 and WRR2, as well as for each of the individual models, and the model ensemble.

Skill Scores

The ability of the models to detect the flood severity was assessed using a contingency table in combination with three skill scores that were based on the model climatology and derived from the table as performance measures. The annual exceedance probabilities (or return periods) for both the observed and modelled discharges extracted from the model-grid cell corresponding to the location of the gauge, were computed using the Gumbel distribution which was estimated using the method of moments. A moving window of seven days for both the observed as well as the modelled discharge was applied to select the maximum discharge of a given event. This window was chosen to disregard possible small time lags between the modelled and observed discharges (Thiemig et al., 2012). The annual exceedance probability thresholds were then used to assess whether or not the modelled discharge is able to capture the timing and intensity of the extreme discharge events. To compare the relative performance of the models, different annual exceedance probability thresholds were used for the modelled as well as for the observed discharges, ranging between 0.342 and 0.005, equivalent to return periods of 1.5-year and 200 years, respectively. These thresholds were used to establish the contingency table for the observed discharge at each gauging station with the discharge from its matching model cell, as shown in Table 3. The table identifies the hits (H, flood events are both modelled and observed in the gauged data), misses (M, flood events are observed but not modelled), false alarms (FA, flood events are modelled but not observed) and correct negatives (CN, flood events are neither observed nor modelled).

Skill scores to quantify the ability of the models to identify flood events were derived from these contingency tables, and include the Critical Success Index (CSI), the Probability of Detection (POD) and the False Alarm Ratio (FAR). These were assessed for each model using either the model or the observed climatology. The CSI and POD determine the percentage of successfully forecasted events of all events observed, whereas the FAR identifies the percentage of incorrectly forecasted flood events out of all events forecasted. The ideal value for CSI and POD is at 100%, while for FAR it is at 0%. The CSI, POD and FAR are calculated using Equations 1, 2 and 3:

$$CSI = \frac{H}{H+M+FA} * 100 \qquad (1)$$

$$POD = \frac{H}{H+M} * 100 \qquad (2)$$

$$FAR = \frac{FA}{H+FA} * 100 \qquad (3)$$

2.4.3 Damaging hydrological extremes

The capability of the models in capturing the flood events that resulted in reported damages was illustrated graphically. The relationship of the severity levels of the damaging flood events that were reported by the disaster databases, and the corresponding annual exceedance probabilities of the observed as well as the modelled discharges at the gauging stations was illustrated. For each reported event, the corresponding maximum discharge (and thus the lowest annual exceedance probability) in either the observed or simulated time series was determined with a moving average of three days before and after the start and end date of the reported flood event (corresponding to a window of seven days for flood events reported to occur on a single date). The reported damaging flood events are reported as occurring in one or more of the seven defined regions. However, as the disaster databases typically report only the broad administrative region of where the flood took place, there was often not enough information available on the sub-basin scale. Therefore, to associate the reported flood events in a region to a flood event being identified in either the observed or the modelled discharges, the lowest annual exceedance probability for every event was determined for each observed river gauging station and corresponding model-grid cell in WRR1 for all stations with an area larger than 2,500 km^2, and in WRR2 for all stations with an area larger than 520 km^2. These sizes of the catchment areas for WRR1 as well as WRR2 were assessed using the NSE statistic in Section 3.1, and are predominantly related to the cell size in WRR1 and WRR2. This process was repeated for all events and for every region in the basin.

3 Results

3.1 Hydrological Performance

The relationship between the upstream catchment area of the river gauging stations in the Limpopo River basin and the error statistics for the models in WRR1 and WRR2, is illustrated in Figure 2 and Table 4. Figure 2 and Table 4a show the three models that are available both in WRR1 and WRR2, whereas Table 4b also provides the performance statistics for the models that are available only in WRR1, as well as the results using the mean of the seven-member ensemble based on the models in WRR1. The different results demonstrate the improvement of model simulations for stations with a large upstream catchment area, when compared to those with smaller ones. This can be best observed by looking at the NSE statistic, from which it is evident that the models are generally able to capture the hydrology for stations with an upstream catchment area that is larger than 2,500 km^2 for WRR1 (Figure 2a), and larger than 520 km^2 for WRR2 (Figure 2d). This provides an indication of the catchment size at which the models are capable of capturing the hydrology, and also illustrates the difference in forcing, resolution and the improvements made in WRR2 as compared to WRR1. The NSE values in Table 4a show that for WRR1 as well as for WRR2, the HTESSEL-CaMa and WaterGAP3 models both perform reasonably well and had roughly equal NSE values, even though the structure of the models is quite different, as the former is a land surface model, while the latter is a global hydrological model.

PBIAS (Figure 2b and 2e, and Table 4) largely shows negative values, indicating an overestimation of the models compared to the observed discharges. This overestimation is visible for all models, and is more dominant at the stations with the smallest upstream catchment areas. This can be expected, as the models take the discharge accumulated over a large area (approximately 2,600 km^2 and 650 km^2 for WRR1 and WRR2 respectively) as the value for one model-grid cell, whereas the true upstream catchment areas of the stations may be as small as 4 km^2. The models for which the overestimation is lower, and which thus generally perform better, are again HTESSEL-CaMa and WaterGAP3, both in WRR1 and WRR2. Furthermore, HTESSEL-

CaMa is the only model that frequently under predicted the discharges, reflected by a positive PBIAS value. The seven models and model ensemble mean that were available in WRR1 (Table 4) have quite distinct differences. The models in WRR1 ranked from best to worst for NSE and PBIAS for only the largest catchment areas were; HTESSEL-CaMa; SURFEX-TRIP; WaterGAP3; the ensemble mean; ORCHIDEE; PCR-GLOBWB; LISFLOOD; and W3RA. The poor performance of W3RA was attributed by consistent severe overestimation of modelled discharges.

The last error statistic considered is Pearson's correlation coefficient, r, displayed in Figure 2c, Figure 2f and Table 4. This error statistic shows relatively consistent correlations for each model, irrespective of the upstream catchment areas. For WRR1, the models that performed best are respectively; SURFEX-TRIP; LISFLOOD; the model ensemble mean; and WaterGAP3, whereas the poorest performance is found for PCR-GLOBWB, and to a lesser extent HTESSEL-CaMa. For WRR2, WaterGAP3 performs significantly better, and also the improvement of WRR2 over WRR1 is notable for both HTESSEL-CaMa and SURFEX-TRIP. LISFLOOD, on the other hand, has a lower r value for WRR2 compared to WRR2. The WRR1 models scores differently for the r values when compared to ranking for NSE and PBIAS. The order, ranking from best to poorest order is; SURFEX-TRIP; LISFLOOD; the ensemble mean; WaterGAP3; W3RA; ORCHIDEE; HTESSEL-CaMa; and lastly PCR-GLOBWB.

Even though some models perform relatively well, the overall performance of the models is, however, quite poor. Average NSE remains negative for all models and upstream catchment areas. Average PBIAS was below 25% in only a few instances for the models HTESSEL-CaMa and WaterGAP3, and the average r value rarely exceeded 0.5.

3.2 Hydrological extremes

3.2.1 Flood Frequency Analysis

The ability of the models in predicting hydrological extremes was analysed by comparing the modelled hydrological extremes to the hydrological extremes that were observed at the river gauging stations (Spookspruit and Limpopo River), as well to the chronology of reported damaging flood events. Results are illustrated for two stations selected as an example in Figure 3. Modelled extremes were analysed using the discharge thresholds derived from either observed climatology or the modelled climatology. The locations of the two river gauging stations are shown in Figure 1, and the model selected is the WaterGAP3 model. Similar patterns were observed at stations with similar sizes and for the other models. Comparing the pattern of flood events identified by MM1 (WRR1 using the modelled climatology), as well as by MM2 (WRR2 using the modelled climatology) to the observed (Obs) or reported (Rep) flood events, it is clear that the WaterGAP3 model is relatively well capable of capturing the variation of the discharge in the observed data, as well as the occurrence of reported damaging events, particularly at the station with a large upstream catchment area, though even at the station with a small upstream catchment area the correspondence in the patterns is reasonable.

Another result derived from Figure 3 is the ability of the models to capture the actual intensity of the identified flood events. This is indicated in the bottom two lines; MO1 and MO2, in which the severity thresholds were established using the observed climatology. The frequency of flood events for WaterGAP3 is quite a bit higher than the observed frequency, with the severity when observed and simulated events do line up also being quite a bit higher. This is clearly the result of the over-prediction of observed discharges. However, there is a marked improvement from the station situated in the river with a small upstream catchment area to the station with a large upstream catchment area, as well as when comparing the higher resolution WRR2 to WRR1. Similar results were found for other models and stations pairs, and accordingly also in the model performance statistics discussed in the previous section.

3.2.2 Skill scores

The upper panel in Figure 4 shows the CSI for each of the models in WRR1, as well as for the seven-model ensemble mean, with discharge thresholds based on model climatology. The score for the models in WRR1 was found to be quite constant for discharges that occur more frequently, i.e. Annual Exceedance Probabilities higher than 0.09, equivalent to a return period of 10 years. The relative performance of the models from best to worst for these discharges is W3RA; the ensemble mean; SURFEX-TRIP; LISFLOOD; WaterGAP3; PCR-GLOBWB; HTESSEL-CaMa; and ORCHIDEE. The pattern, however, changes for the more extreme (low probability) discharges. The discharges with an annual exceedance probability that was less than 0.09 showed a greater spread, as well as changes in the order of performance of the models. For example, SURFEX-TRIP and LISFLOOD now perform better, while W3RA performs worse for these more extreme discharge events. The model ensemble mean though has a remarkably high CSI score which is independent of the return period.

The differences in performance of WRR2 compared to WRR1 as a result of increased spatial resolution, different forcing and model improvements becomes evident from the lower panel of Figure 4. For WaterGAP3, HTESSEL-CaMa and SURFEX-TRIP, using WRR2 yields higher CSI values. For LISFLOOD, on the other hand, the performance of WRR1 is better than that in WRR2. Again, it appears that WaterGAP3 WRR2 performs best overall. These same patterns are observed regarding the error statistics, as shown in Figure 2 and discussed in Section 3.1.

The underlying reason for the observed patterns of the CSI can be explained by taking a closer look at the POD and FAR. The performances of all three skill scores with respect to the upstream catchment area of each individual station are shown in Figure 5. Skill scores are shown here for events with an annual exceedance probability of 0.164, equivalent to a return period of 5 years. As can be observed by looking at the models in WRR1 (upper panel), the average POD is around 25% and the average FAR is around 70%, resulting in an average CSI of roughly 15%. The CSI, POD and FAR all have a relatively large spread, with little relationship to the upstream catchment area of the stations. Stations with a larger upstream catchment area do not necessarily result in better skill scores. An explanation for the lack of a relationship with catchment areas is that the three skill scores are based on model climatology, and thus the relative flood intensity, while the error statistics are based on the observed climatology and thus the absolute intensities. This clarifies the notable difference with the error statistics, such as the NSE (as shown in Figure 2a and d), where the improvement of stations with a larger upstream catchment area is clear. This suggests that the performance of the models in estimating the relative intensity, is not highly influenced by the upstream catchment area of the river gauging stations. The difference in performance between WRR1 and WRR2 is, however, apparent. Both HTESSEL-CaMa and WaterGAP3 display improved values for the CSI, POD and FAR. Again, the notable exception is LISFLOOD, where WRR1 performs better than for WRR2, independent of the skill score. This again reflects the error statistics discussed in Section 3.1 and is in correspondence with Arduini et al. (2017) and Dutra et al. (2017). There are a number of factors that could contribute to this observation, such as the model modifications (see Table 1) and that the same calibration parameterization were used as in WRR1, even though the alterations to the model require an updated calibration (Arduini et al., 2017).

3.3 Damaging hydrological extremes

Scatter plots were used to demonstrate the relationship between the reported severity of the reported flood events with the severity of the corresponding events identified in the observed as well as the modelled discharges. These scatter plots are shown in Figure 6, illustrating the reported flood severity in discrete classes (x-axis), as well as the annual exceedance probability for the events identified using the maximum of the modelled or observed discharges in a seven day window around the reported event (y-axis). The exceedance probability found at each station is plotted. Ideally the events should be clustered along the diagonal from top left (higher probability, lower severity), to bottom right (lower probability, higher severity), reflecting that lower impact flood events typically occur in only a few stations and have higher probabilities (low return periods), while high impact severe flood events are often basin wide, occurring at most stations across the basin with lower probabilities. For medium severity reported events, a wider scatter would be expected, as these events may occur only in a part of the basin.

The figure shows that when a reported flood event is classified at the most severe category 5, impacts were observed throughout the basin, as all observed as well as modelled probabilities indicate above normal river discharge, many of which with extreme (low probability) discharges. Small-scale flood events that resulted in low as well as localised damages, on the other hand, were classified either as category 0 or 1. As can be seen in Figure 6, the annual exceedance probabilities corresponding to these events have a larger spread. The reason for this is that small-scale events are not noticeable throughout the entire region, but only locally, as many gauges were still measuring normal flow, while those where the event does occur show more extreme discharges. It can be observed though that part of the gauges measured an above normal discharge, whereas this was frequently not observed by the models. Only WaterGAP3 was able to detect extreme discharges for the floods with a severity level of zero. Apart from that, the four different models displayed comparable results, although HTESSEL-CaMa generally had lower annual exceedance probabilities for the same flood events when compared to the other models.

4 Discussion

The potential of the global Water Resources Re-analysis dataset was assessed by studying the hydrological performance, identification of hydrological extremes, as well as of damaging flood events, and was evaluated by means of commonly used error statistics and verification skill scores (CSI, POD and FAR). The verification of the models within the WRR dataset was largely dependent on the observed river discharge data. Access to these data proved to be quite challenging, and the quality of the discharge data that was obtained was often insufficient. Only 75 of the 196 river gauging stations for which at least some data available in the Limpopo for the desired time range were used in this research, with most of these in South Africa. This has implications for the conclusions drawn from the research, especially for the PBIAS as it is highly influenced by the uncertainty in the observed data (Moriasi et al., 2007). Despite these limitations, this research shows that the discharges that were estimated by the different global models are to some extent able to capture the variability of observed discharges, as indicated by the different error statistics. For instance, the NSE demonstrated that for WRR1 as well as WRR2, both the HTESSEL-CaMa and the WaterGAP3 models performed well with roughly similar NSE values, despite the different structure of these models. HTESSEL-CaMa is a LSM and does not include lakes and reservoirs or water usage, whereas WaterGAP3 is a GHM and does include both lakes and reservoirs, as well as water usage (Table 1). The differences between the model structures is illustrated by the PBIAS and r values. HTESSEL-CaMa has reasonable PBIAS, while WaterGAP3 has a relatively good r. As noted in Section 2.1, the basin is highly altered due to human influences, in particular by a large number of storage reservoirs. Models that capture only natural flow conditions, and do not take the reservoirs and water usage into account, may be able to reasonably estimate runoff volumes, though they do tend to largely overestimate the actual magnitude of the discharges. Not including human influences such as regulation, however, results in low correlations. The relative intensity of flood events, on the other hand, can still be well captured by the same model when using the model climatology instead of the observed climatology as a reference. An example of such a model is W3RA, which performs poorly when considering the error statistics, but relatively well for the CSI.

Global models are best suited for the modelling of large-scale processes, but poorly represent the small-scale ones such as the variability associated with convection (Beck et al., 2017). These conclusions have been drawn in similar research, such as Asante et al. (2008), Thiemig et al. (2015) and Trigg et al. (2016). This study indicates that the small-scale flood events were generally not well captured by the global models that were analysed in this research. The results do show, however, that the performance of these global models improves with model developments in terms of resolution, forcing and model parameterization. The statistics for model performance measures for the higher resolution WRR2 starts to approach reasonable values for gauges with upstream areas of some 500 km^2, while for the lower resolution WRR1 these same values are attained only at areas of some 2,500 km^2. The higher resolution WRR2 also shows for two of the three models better skill in identifying reported flood events, represented in the chronology of reported flood events developed. Whether the improved performance of the higher resolution is due to the improved and higher resolution MSWEP forcing data (Beck et al., 2017), or due to improved representation of hydrological processes is unclear. However, as the improvements vary between the models, it is clear that model structure has an influence.

That there is skill in these global models in identifying flood events that have impacts, and that this skill improves as the resolution of these large-scale models improves, is significant. Global scale forecasting systems (Alfieri et al., 2013) as well as those at continental scale (Thiemig et al., 2011) typically employ such large-scale models for developing forecasts, using thresholds based on model climatology to inform the severity of predicted events and subsequent issuing of flood warnings. Such warnings may be issued where there are no (reliable) river gauges, as is the case in much of the Limpopo basin, making calibration of a local model difficult. The ability of these global and or continental models to predict the occurrence of flood events that have impacts bolsters the confidence of using these warnings to initiate response, though the high false alarm rate found could again diminish confidence.

It is important to note, though, that likely not all small-scale flood events that occurred between 1980 and 2012 will have been included in the chronology of reported flood events that was developed. As has more often found to be the case in the Global South (Brakenridge, 2017), the availability of disaster data in the Limpopo River basin is fairly limited. In order to construct a basin-wide timeline of historic damaging floods, events reported in the EM-DAT, GAALFE and NatCatSERVICE databases were collated. Even though the three used here are currently the most comprehensive databases containing reported damaging historic flood events in Africa (Aich et al., 2014b), several shortcomings are noted. These include inconsistencies between events reported, gaps, and limited reporting in some areas (Guha-Sapir et al., 2016). Additionally, most disaster databases are available at the country scale, whereas flood events occur at the basin or finer scales. It is recommended to enhance the reporting of flood disasters by providing more details on the losses that were incurred as well as a more precise description of the location and extent of the floods. The basin-wide approach to identify past flood events by using empirical disaster databases used in this research has also been applied in other research (Aich et al. (2014b), Asante et al. (2008), Bischiniotis et al. (2018), Huggel et al. (2015) and Thiemig et al. (2015)), noting similar deficiencies.

The flood classification that was used in this research is a discrete classification, taking the number of fatalities and overall losses into account. However, it is expected that a continuous flood severity classification would be better able to reveal the relationship between extreme river discharges and the intensity of reported damaging flood events. However, due to the gaps in the reported damaging flood event data as well as broad area descriptions, this could not be assessed at this point. In order to identify the added value of such a classification, additional research is required in addition to improving disaster loss data.

In this study, the Gumbel distribution is used to determine the annual exceedance probability thresholds of both the modelled and observed discharge data. Different extreme value distribution, however, can significantly influence the probability of the extreme discharges (Dankers and Feyen, 2008). The Gumbel distribution is a two-parameter distribution and was applied due its simplicity and robustness, though some authors (e.g. Ponce (1989), argue that a three-parameter distribution such as the GEV or the Log Pearson type III should be used for flood frequency analysis. However, the goodness of fit of the distributions found was tested using the Kolmogorov Smirnoff test, with stations that did not meet the 5% significance threshold not considered. Further inspection of these stations revealed that these were often directly downstream of a dam, or otherwise strongly influenced by human activities. Additional research could additionally explore the influence of using more complex extreme value distributions. This could also consider the influence of the length of the moving window that was used to identify the maximum discharge in the observed and modelled time series. This moving window was chosen to allow for the travel time from the upstream parts of the sub catchment. In reality, however, the catchment upstream of each gauging station has its own time of concentration, and the window used could be made specific for each station accordingly.

Of the global models considered in this study, the higher resolution WaterGAP3 in WRR2 demonstrated the best performance, both for capturing the hydrological behaviour across the Limpopo basin, as indicated by good values for the error statistics, as well as for identifying the occurrence and severity of hydrological extremes, which was indicated by the skill scores. It was also observed that WaterGAP3 in WRR2 is reasonably good at estimating low annual exceedance probabilities for the damaging flood events for the stations with a large upstream catchment area. One reason for this improved performance may be the inclusion of lakes and reservoirs, as well as water abstractions in the model. However, results for other models, such as W3RA, which has the worst model performance error statistics, may rank higher than other models when used to identifying the occurrence of flood events, where these are identified using the model's own climatology. It is also important to note that if similar research would be applied elsewhere, the ranking of model performance may be quite different. The ranking of the models also clearly depends on the aim of the research. WaterGAP3 for instance performed poorly in respect to other global hydrological models in research focussing on a snowmelt driven catchment (Casson et al., 2018). Furthermore, when the key interest is the relative performance of the model for the Limpopo River basin, taking only the model climatology into account, the W3RA model would be the preferred model, as it has a high CSI. However, when the main goal would be the absolute magnitude of discharges, the W3RA model would not be considered, as it is found to severely overestimate the discharges in the Limpopo River basin. The seven-model ensemble mean, on the other hand, proved to be quite consistent in its performance. For the CSI values particularly it scores remarkably high, but it also scores relatively well for the Pearson's correlation coefficient r. Though it should carefully be assessed which model would be the best applicable for each instance, the model ensemble mean would be the safest bet in an area where no model clearly stands out.

5 Conclusion

The study explores the use of a global re-analysis dataset developed within the EU-FP7 EarH2Obverve (E2O) project, which is constructed using a set of global hydrological and land surface models, to support flood risk analysis in data sparse regions, such as the Limpopo River basin. There is a necessity for such re-analysis data, since measured river discharge data in this basin and others like it are currently insufficient, poorly spatially distributed, have an insufficient period of record, or are partly inaccessible. The E2O re-analysis dataset provides hydro-meteorological data of sufficient length and coverage required for statistical analysis. When the variability of the discharge results of the ensemble of models included in the re-analysis dataset is evaluated, the error statistics found show that the models all have reasonable skill in capturing the variability of the observed discharges, though there may be significant bias in magnitude. This was indicated by strong correlations, low Nash-Sutcliffe Efficiency and high percent-bias values. Furthermore, the error statistics revealed that the variability is better captured by the models at hydrological gauging stations that have larger upstream catchment areas compared to those in smaller catchments. The upstream catchment areas of the river gauging stations at which WRR1 and WRR2 are able to provide representation of the hydrological behaviour that is better than the average of the observed is found for catchment areas of some 2,500 km^2 and 520 km^2 and above respectively, with significantly poorer performance for smaller catchment sizes. This shows that the continued improvements in the global models with a higher resolution, either due to improved higher resolution forcing, or due to improved model structures, can be expected to lead in most cases to better capabilities of capturing the variability of the observed discharge as well as the magnitude of observed discharges.

A novel aspect of this study is in exploring the skill of the global models in identifying the occurrence and severity of flood events in two benchmark chronologies of flood events. The first was developed through flood frequency analysis, with flood events identified to occur at selected probabilities, while the second was developed through collating reported flood events in three disaster impact databases. This shows that the global models do have skill in capturing the observed as well as reported damaging floods. This is, however, only the case when the thresholds of the discharges corresponding to the flood events are determined using the model climatology, and not the observed climatology. The simulated discharges of these global models are thus found to better represent the variability of the observed discharges, than the magnitude; though this is less an issue for the better performing higher resolution models of WRR2.

Despite the absence of high-quality data in the Limpopo River basin and the coarse resolution of the models in the global re-analysis dataset, this research shows that regardless these limitations, the global re-analysis dataset can provide valuable information for flood risk assessment in data sparse regions. The skill of the models to predict flood events in the basin that have led to flood damage, as recorded in the chronology of reported floods is an important finding, as global models such as those assessed here are often used in global and continental forecasting systems to generate flood forecasts and issue warnings in basins with little or no gauged data, but where floods and consequent impacts do occur. This indicates that openly available global scale hydro-meteorological data can provide valuable information regarding extreme events in data sparse regions and may therefore be of use to local decision makers in mitigating the negative consequences of future flood events, and that this may improve as the resolution of these global models improves.

Competing interests. The authors declare that they have no competing interests.

Acknowledgements. This work received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 603608, Global Earth Observation for Integrated Water Resource Assessment (eartH2Observe). We are grateful to Munich Re for providing historic flood events from their NatCatSERVICE database in the framework of NWO VIDI grant 016-161-324, and EU FP7 IMPREX project, grant agreement no. 641811.

References

Aich, V., Koné, B., Hattermann, F. F., and Müller, E. N.: Floods in the Niger basin - analysis and attribution, Natural Hazards and Earth System Sciences Discussions, 2, 5171–5212, https://doi.org/10.5194/nhessd-2-5171-2014, http://www.nat-hazards-earth-syst-sci-discuss.net/2/5171/2014/, 2014a.

Aich, V., Liersch, S., Vetter, T., Huang, S., Tecklenburg, J., Hoffmann, P., Koch, H., Fournet, S., Krysanova, V., Müller, E. N., and Hattermann, F. F.: Comparing impacts of climate change on streamflow in four large African river basins, Hydrology and Earth System Sciences, 18, 1305–1321, https://doi.org/10.5194/hess-18-1305-2014, 2014b.

Alfieri, L., Burek, P., Dutra, E., Krzeminski, B., Muraro, D., Thielen, J., and Pappenberger, F.: GloFAS-global ensemble streamflow forecasting and flood early warning, Hydrology and Earth System Sciences, 17, 1161–1175, https://doi.org/10.5194/hess-17-1161-2013, 2013.

Arduini, G., Fink, G., Martinez de la Torre, A., Nikolopoulos, E., Anagnostou, E., Balsamo, G., and Boussetta, S.: End-user-focused improvements and descriptions of the advances introduced between the WRR tier1 and WRR tier2, Tech. rep., eartH2Observe, 2017.

Asante, K. O., Artan, G. a., Pervez, S., and Rowland, J.: A linear geospatial streamflow modeling system for data sparse environments, International Journal of River Basin Management, 6, 233–241, https://doi.org/10.1080/15715124.2008.9635351, 2008.

Ashton, P., Love, D., Mahachi, H., and Dirks, P.: An overview of the impact of mining and mineral processing operations on water resources and water quality in the Zambezi, Limpopo and Olifants Catchments in Southern Africa, Tech. rep., CSIR- Environmentek, Pretoria, South Africa and Geology Department, University of Zimbabwe, Harare, Zimbabwe, https://doi.org/ENV-P-C 2001-042., 2001.

Balsamo, G., Viterbo, P., Beljaars, A., van den Hurk, B., Hirschi, M., Betts, A. K., and Scipal, K.: A Revised Hydrology for the ECMWF Model: Verification from Field Site to Terrestrial Water Storage and Impact in the Integrated Forecast System, Journal of Hydrometeorology, 10, 623 – 643, https://doi.org/10.1175/2008JHM1068.1, 2009.

Balsamo, G., Albergel, C., Beljaars, A., Boussetta, S., Brun, E., Cloke, H., Dee, D., Dutra, E., Munõz-Sabater, J., Pappenberger, F., De Rosnay, P., Stockdale, T., and Vitart, F.: ERA-Interim/Land: A global land surface reanalysis data set, Hydrology and Earth System Sciences, 19, 389–407, https://doi.org/10.5194/hess-19-389-2015, 2015.

Beck, H. E., van Dijk, A. I. J. M., Levizzani, V., Schellekens, J., Miralles, D. G., Martens, B., and de Roo, A.: MSWEP: 3-hourly 0.25° global gridded precipitation (1979-2015) by merging gauge, satellite, and reanalysis data, Hydrology and Earth System Sciences, 21, 589–615, https://doi.org/10.5194/hess-2016-236, 2017.

Bischiniotis, K., Van Den Hurk, B., Jongman, B., Coughlan De Perez, E., Veldkamp, T., De Moel, H., and Aerts, J.: The influence of antecedent conditions on flood risk in sub-Saharan Africa, Natural Hazards and Earth System Sciences, 18, 271–285, https://doi.org/10.5194/nhess-18-271-2018, 2018.

Biswas, A. K.: Management of International Waters: Opportunities and Constraints, Water Resources Development, 15, 429–441, 1999.

Brakenridge, G. R.: Global Active Archive of Large Flood Events, Dartmouth Flood Observatory, University of Colorado, USA, http://floodobservatory.colorado.edu/Archives/index.html, 2017.

Casson, D. R., Werner, M., Weerts, A., and Solomatine, D.: Global re-analysis datasets to improve hydrological assessment and snow water equivalent estimation in a Sub-Arctic watershed, Hydrology and Earth System Sciences Discussions, 5194, 1–20, https://doi.org/10.5194/hess-2018-82, 2018.

CRED and Guha-Sapir, D.: EM-DAT: The Emergency Events Database, Université catholique de Louvain (UCL), Brussels, Belgium, www.em-dat.be, 2017.

Dankers, R. and Feyen, L.: Climate change impact on flood hazard in Europe: An assessment based on high-resolution climate simulations, Journal of Geophysical Research Atmospheres, 113, 1–17, https://doi.org/10.1029/2007JD009719, 2008.

Dankers, R., Arnell, N. W., Clark, D. B., Falloon, P. D., Fekete, B. M., Gosling, S. N., Heinke, J., Kim, H., Masaki, Y., Satoh, Y., Stacke, T., Wada, Y., and Wisser, D.: First look at changes in flood hazard in the Inter-Sectoral Impact Model Intercomparison Project ensemble., Proceedings of the National Academy of Sciences of the United States of America, 111, 3257–61, https://doi.org/10.1073/pnas.1302078110, 2014.

Decharme, B., Alkama, R., Douville, H., Becker, M., and Cazenave, A.: Global Evaluation of the ISBA-TRIP Continental Hydrological System. Part II: Uncertainties in River Routing Simulation Related to Flow Velocity and Groundwater Storage, Journal of Hydrometeorology, 11, 601–617, https://doi.org/10.1175/2010JHM1212.1, 2010.

Dirmeyer, P. A., Gao, X., Zhao, M., Guo, Z., Oki, T., and Hanasaki, N.: GSWP-2: Multimodel analysis and implications for our perception of the land surface, Bulletin of the American Meteorological Society, 87, 1381–1397, https://doi.org/10.1175/BAMS-87-10-1381, 2006.

Dottori, F., Salamon, P., Bianchi, A., Alfieri, L., and Hirpa, F. A.: Advances in Water Resources Development and evaluation of a framework for global flood hazard mapping, Advances in Water Resources, 94, 87–102, https://doi.org/10.1016/j.advwatres.2016.05.002, 2016.

Dutra, E., Balsamo, G., Calvet, J.-C., Minvielle, M., Eisner, S., Fink, G., Pessenteiner, S., Orth, R., Burke, S., van Dijk, A. I., Polcher, J., Beck, H. E., and Martinez de la Torre, A.: Report on the current state-of-the-art Water Resources Reanalysis, Tech. Rep. Tech. Rep. D.5.1, eartH2Observe, 2015.

Dutra, E., Balsamo, G., Calvet, J.-C., Munier, S., Burke, S., Fink, G., van Dijk, A. I. J. M., Martinez de la Torre, A., van Beek, R., de Roo, A., and Polcher, J.: Report on the improved Water Resources Reanalysis (WRR2), Tech. Rep. Tech. Rep. D.5.2, eartH2Observe, 2017.

FAO: Drought impact mitigation and prevention in the Limpopo River Basin: A situation analysis, Land and Water Discussion Paper 4, 4, 1–160, https://doi.org/10.1016/S0009-2509(96)00385-5, 2004.

Flörke, M., Kynast, E., Bärlund, I., Eisner, S., Wimmer, F., and Alcamo, J.: Domestic and industrial water uses of the past 60 years as a mirror of socio-economic development: A global simulation study, Global Environmental Change, 23, 144–156, https://doi.org/10.1016/j.gloenvcha.2012.10.018, 2013.

Guha-Sapir, D., Hoyois, P., and Below, R.: Annual Disaster Statistical Review 2015: The numbers and trends, Tech. rep., Centre for Research on the Epidemiology of Disasters (CRED), Brussels, Belgium, https://doi.org/10.1093/rof/rfs003, 2016.

Gumbel, E.: The Return Period of Flood Flows, The Annals of Mathematical Statistics, 12, 163–190, 1941.

Gupta, H. V., Sorooshian, S., and Yapo, P. O.: Status of automatic calibration for hydrologic models: Comparison with multilevel expert calibration, Journal of Hydrologic Engineering, 4, 135–143, https://doi.org/10.1002/fut.20174, 1999.

Haddeland, I., Clark, D. B., Franssen, W., Ludwig, F., Voß, F., Arnell, N. W., Bertrand, N., Best, M., Folwell, S., Gerten, D., Gomes, S., Gosling, S. N., Hagemann, S., Hanasaki, N., Harding, R., Heinke, J., Kabat, P., Koirala, S., Oki, T., Polcher, J., Stacke, T., Viterbo, P., Weedon, G. P., and Yeh, P.: Multimodel estimate of the global terrestrial water balance: setup and first results, Journal of Hydrometeorology, 12, 869–884, https://doi.org/10.1175/2011JHM1324.1, 2011.

Huggel, C., Raissig, A., Rohrer, M., Romero, G., Diaz, A., and Salzmann, N.: How useful and reliable are disaster databases in the context of climate and global change? A comparative case study analysis in Peru, Natural Hazards and Earth System Sciences, 15, 475–485, https://doi.org/10.5194/nhess-15-475-2015, 2015.

Hughes, D. A.: Comparison of satellite rainfall data with observations from gauging station networks, Journal of Hydrology, 327, 399–410, https://doi.org/10.1016/j.jhydrol.2005.11.041, 2006.

IPCC: Managing the risks of extreme events and disasters to advance climate change adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change, Tech. rep., Cambridge, United Kingdom and New York, NY, USA, https://doi.org/10.1596/978-0-8213-8845-7, 2012.

IPCC: Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Tech. rep., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2014.

Jongman, B., Winsemius, H. C., Aerts, J. C. J. H., Coughlan, E., Perez, D., and Aalst, M. K. V.: Declining vulnerability to river floods and the global benefits of adaptation, Proceedings of the National Academy of Sciences, 112, 2271–2280, https://doi.org/10.1073/pnas.1414439112, 2015.

Krinner, G., Viovy, N., de Noblet-Ducoudre, N., Ogée, J., Polcher, J., Friedlingstein, P., Ciais, P., Sitch, S., and Prentice, I. C.: A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system, Global Biochemical Cycles, 19, 1–33, https://doi.org/10.1029/2003GB002199, 2005.

Kron, W., Steuer, M., Löw, P., and Wirtz, A.: How to deal properly with a natural catastrophe database - Analysis of flood losses, Natural Hazards and Earth System Science, 12, 535–550, https://doi.org/10.5194/nhess-12-535-2012, 2012.

Kundzewicz, Z. W., Budhakooncharoen, S., Bronstert, A., Hoff, H., Lettenmaier, D., Menzel, L., and Schulze, R.: Floods and Drougts: Coping with Variability and Climate Change, Natural Resources Forum, 26, 263–274, 2002.

LBPTC: Joint Limpopo River Basin Study - Scoping Phase - Final Report, Tech. rep., Limpopo Basin Permanent Technical Committee, Mozambique, 2010.

Legates, D. R. and McCabe Jr., G. J.: Evaluating the use of "goodness of fit" measures in hydrologic and hydroclimatic model validation, Water Resources Research, 35, 233–241, https://doi.org/10.1029/1998WR900018, 1999.

Maposa, D., Cochran, J. J., Lesaoana, M., and Sigauke, C.: Estimating high quantiles of extreme flood heights in the lower Limpopo River basin of Mozambique using model based Bayesian approach, Natural Hazards and Earth System Science Discussions, 2, 5401–5425, https://doi.org/10.5194/nhessd-2-5401-2014, 2014.

Massey Jr., F. J.: The Kolmogorov-Smirnov Test for Goodness of Fit, Journal of the American Statistical Association, 46, 68–78, 1951.

Moriasi, D., Arnold, J., Van Liew, M., Binger, R., Harmel, R., and Veith, T.: Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, American Society of Agricultural and Biological Engineers (ASABE), 50, 885–900, https://doi.org/10.13031/2013.23153, 2007.

Mujere, N.: Flood Frequency Analysis Using the Gumbel Distribution, International Journal on Computer Science and Engineering, 3, 2774–2778, 2011.

Munich-Re: NatCatSERVICE Database. Munich: Munich Reinsurance Company Geo Risk Research, 2017.

Naumann, G., Dutra, E., Barbosa, P., Pappenberger, F., Wetterhall, F., and Vogt, J. V.: Comparison of drought indicators derived from multiple data sets over Africa, Hydrology and Earth System Sciences, 18, 1625–1640, https://doi.org/10.5194/hess-18-1625-2014, 2014.

Patt, A. G. and Schro, D.: Perceptions of climate risk in Mozambique : Implications for the success of adaptation strategies, Global Environmental Change, 18, 458–467, https://doi.org/10.1016/j.gloenvcha.2008.04.002, 2008.

Ponce, V.: Engineering Hydrology, Principles and Practices, Prentice Hall, 1989.

Pozzi, W., Sheffield, J., Stefanski, R., Cripe, D., Pulwarty, R., Vogt, J. V., Heim, R. R., Brewer, M. J., Svoboda, M., Westerhoff, R., Van Dijk, A. I. J. M., Lloyd-Hughes, B., Pappenberger, F., Werner, M., Dutra, E., Wetterhall, F., Wagner, W., Schubert, S., Mo, K., Nicholson, M., Bettio, L., Nunez, L., Van Beek, R., Bierkens, M., De Goncalves, L. G. G., De Mattos, J. G. Z., and Lawford, R.: Toward global drought early warning capability: Expanding international cooperation for the development of a framework for monitoring and forecasting, Bulletin of the American Meteorological Society, 94, 776–785, https://doi.org/10.1175/BAMS-D-11-00176.1, 2013.

Rodell, M., Houser, P. R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., Entin, J. K., Walker, J. P., Lohmann, D., and Toll, D.: The global land data assimilation system, Bulletin of the American Meteorological Society, 85, 381–394, https://doi.org/10.1175/BAMS-85-3-381, 2004.

Schellekens, J., Dutra, E., Martínez-De La Torre, A., Balsamo, G., Van Dijk, A., Sperna Weiland, F., Minvielle, M., Calvet, J. C., Decharme, B., Eisner, S., Fink, G., Flörke, M., Peßenteiner, S., Van Beek, R., Polcher, J., Beck, H., Orth, R., Calton, B., Burke, S., Dorigo, W., and Weedon, G. P.: A global water resources ensemble of hydrological models: The eartH2Observe Tier-1 dataset, Earth System Science Data, 9, 389–413, https://doi.org/10.5194/essd-9-389-2017, 2017.

Silva, J. A., Eriksen, S., and Ombe, Z. A.: Double exposure in Mozambique's Limpopo River Basin, The Geographical Journal, 176, 6–24, https://doi.org/10.1111/j.1475-4959.2009.00343.x, 2010.

Smith, A., Sampson, C., and Bates, P.: Regional flood frequency analysis at the global scale, Water Resources Research, 51, 539–553, https://doi.org/10.1002/ 2014WR015814, 2015.

Sood, A. and Smakhtin, V.: Global hydrological models: a review, Hydrological Sciences Journal, 60, 549–565, https://doi.org/10.1080/02626667.2014.950580, 2015.

Spaliviero, M., De Dapper, M., Mannaerts, C. M., and Yachan, A.: Participatory approach for integrated basin planning with focus on disaster risk reduction: The Case of the Limpopo River, Water, 3, 737–763, https://doi.org/10.3390/w3030737, 2011.

Thiemig, V., de Roo, A., and Gadain, H.: Current status on flood forecasting and early warning in Africa, International Journal of River Basin Management, 9, 63–78, https://doi.org/10.1080/15715124.2011.555082, 2011.

Thiemig, V., Rojas, R., Zambrano-Bigiarini, M., Levizzani, V., De Roo, A., Thiemig, V., Rojas, R., Zambrano-Bigiarini, M., Levizzani, V., and Roo, A. D.: Validation of Satellite-Based Precipitation Products over Sparsely Gauged African River Basins, Journal of Hydrometeorology, 13, 1760–1783, https://doi.org/10.1175/JHM-D-12-032.1, http://journals.ametsoc.org/doi/abs/10.1175/JHM-D-12-032.1, 2012.

Thiemig, V., Bisselink, B., Pappenberger, F., and Thielen, J.: A pan-African medium-range ensemble flood forecast system, Hydrology and Earth System Sciences, 19, 3365, https://doi.org/10.5194/hess-19-1-2015, 2015.

Trambauer, P., Maskey, S., Werner, M., Pappenberger, F., Beek, L. P. H. V., Uhlenbrook, S., Park, S., and Section, W. R.: Identification and simulation of space – time variability of past hydrological drought events in the Limpopo River basin , southern Africa, Hydrology and Earth System Sciences, 18, 2925–2942, https://doi.org/10.5194/hess-18-2925-2014, 2014.

Trambauer, P., Werner, M., Winsemius, H. C., Maskey, S., Dutra, E., and Uhlenbrook, S.: Hydrological drought forecasting and skill assessment for the Limpopo River basin, southern Africa, Hydrology and Earth System Sciences, 19, 1695–1711, https://doi.org/10.5194/hess-19-1695-2015, 2015.

Trigg, M. A., Birch, C. E., Neal, J. C., Bates, P. D., Smith, A., Sampson, C. C., Yamazaki, D., Hirabayashi, Y., Pappenberger, F., Dutra, E., Ward, P. J., Winsemius, H. C., Salamon, P., Dottori, F., Rudari, R., Kappes, M. S., Simpson, A. L., Hadzilacos, G., and Fewtrell, T. J.: The credibility challenge for global fluvial flood risk analysis, Environmental Research Letters, 11, 094 014, https://doi.org/10.1088/1748-9326/11/9/094014, 2016.

UNISDR: Sendai Framework for Disaster Risk Reduction 2015 - 2030, https://doi.org/A/CONF.224/CRP.1, www.unisdr.org/we/inform/publications/43291, 2015.

UNISDR: Disaster Risk Reduction in Africa, Status Report - 2015, Executive summary, https://doi.org/10.1002/aehe.3640230702, 2016.

UNISDR & CRED: The Human Cost of Weater Related Disasters, 1995-2015, https://doi.org/10.1017/CBO9781107415324.004, 2015.

van Beek, L. P. H. and Bierkens, M. F. P.: The Global Hydrological Model PCR-GLOBWB: Conceptualization, Parameterization and Verification, http://vanbeek.geo.uu.nl/suppinfo/vanbeekbierkens2009.pdf, 2009.

van der Knijff, J. M., Younis, J., and de Roo, A. P. J.: LISFLOOD : a GIS - based distributed model for river basin scale water balance and flood simulation, International Journal of Geographical Information Science, 24, 189 – 212, https://doi.org/10.1080/13658810802549154, 2008.

van Dijk, A. I. J. M., Renzullo, L. J., Wada, Y., and Tregoning, P.: A global water cycle reanalysis (2003-2012) merging satellite gravimetry and altimetry observations with a hydrological multi-model ensemble, Hydrology and Earth System Sciences, 18, 2955–2973, https://doi.org/10.5194/hess-18-2955-2014, 2014.

Ward, P., Jongman, B., Sperna Weiland, F., Bouwman, A., van Beek, R., Bierkens, M., Ligtvoet, W., and Winsemius, H. C.: Assessing flood risk at the global scale: Model setup, results, and sensitivity, Environmental Research Letters, 8, 044 019, https://doi.org/10.1088/1748-9326/8/4/044019, 2013.

Ward, P. J., De Moel, H., and Aerts, J. C. J. H.: How are flood risk estimates affected by the choice of return-periods?, Natural Hazards and Earth System Science, 11, 3181–3195, https://doi.org/10.5194/nhess-11-3181-2011, 2011.

Weedon, G., Balsamo, G., Bellouin, N., Gomes, S., Best, M., and Viterbo, P.: The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data, Water Resources Research, 50, 7505–7514, https://doi.org/10.1002/2014WR015638.Received, 2014.

WMO: Limpopo River Basin: A proposal to improve the flood forecasting and early warnign systems, 2012.

Wood, E. F., Roundy, J. K., Troy, T. J., Van Beek, L. P. H., Bierkens, M. F. P., Blyth, E., de Roo, A., Doell, P., Ek, M., Famiglietti, J., Gochis, D., van de Giesen, N., Houser, P., Jaffe, P. R., Kollet, S., Lehner, B., Lettenmaier, d. p., Peters-Lidard, C., Sivapalan, M., Sheffield, J., Wade, A., and Whitehead, P.: Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water, Water Resources Research, 47, 1–10, https://doi.org/10.1029/2010WR010090, 2011.

Zhao, F., Veldkamp, T. I. E., Frieler, K., Schewe, J., Ostberg, S., Willner, S., Schauberger, B., Gosling, S. N., Schmied, H. M., Portmann, F. T., Leng, G., Huang, M., Liu, X., Tang, Q., Hanaski, N., Bemoans, H., Gerten, D., Satoh, Y., Pokhrel, Y., Stacke, T., Ciais, P., Chang, J., Ducharne, A., Guimberteau, M., Wada, Y., Kim, H., and Yamazaki, D.: The critical role of the routing scheme in simulating peak river discharge in global hydrological models, Environ. Res. Lett, 12, 075 003, https://doi.org/10.1088/1748-9326/aa7250, http://iopscience.iop.org/1748-9326/12/7/075003, 2017.

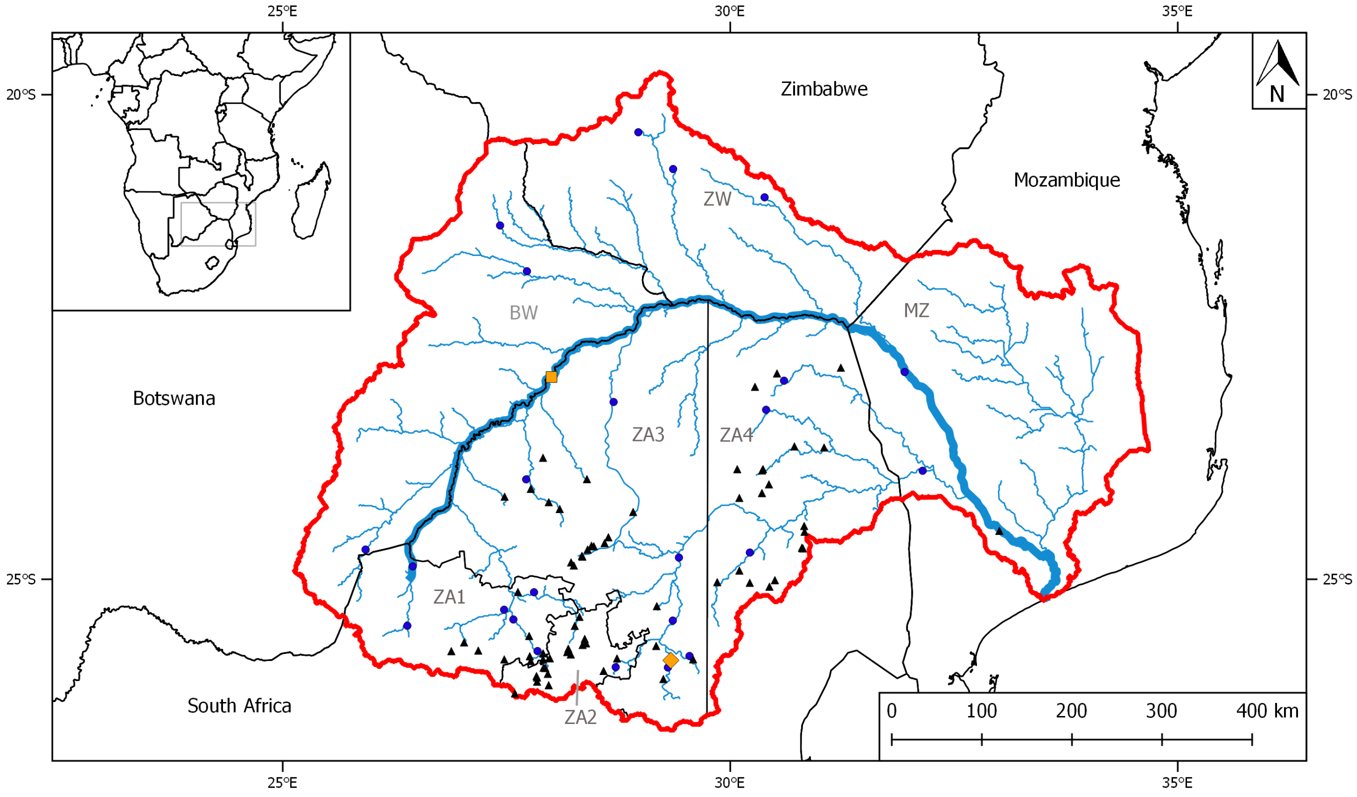


Figure 1. Map of Limpopo River basin with the riparian countries, major tributaries, and the seven regions in the basin that were identified for this research. Also shown are the major dams (blue circle) and the river gauging stations with at least 25 years of data between 1980 and 2012 (black triangle). The stations used to illustrate the flood frequency analysis in Section 3.2.1 are shown by a square (located upstream in the Spookspruit tributary; 252 km^2), and a diamond (located at the main stem of the Limpopo River; 98,240 km^2).

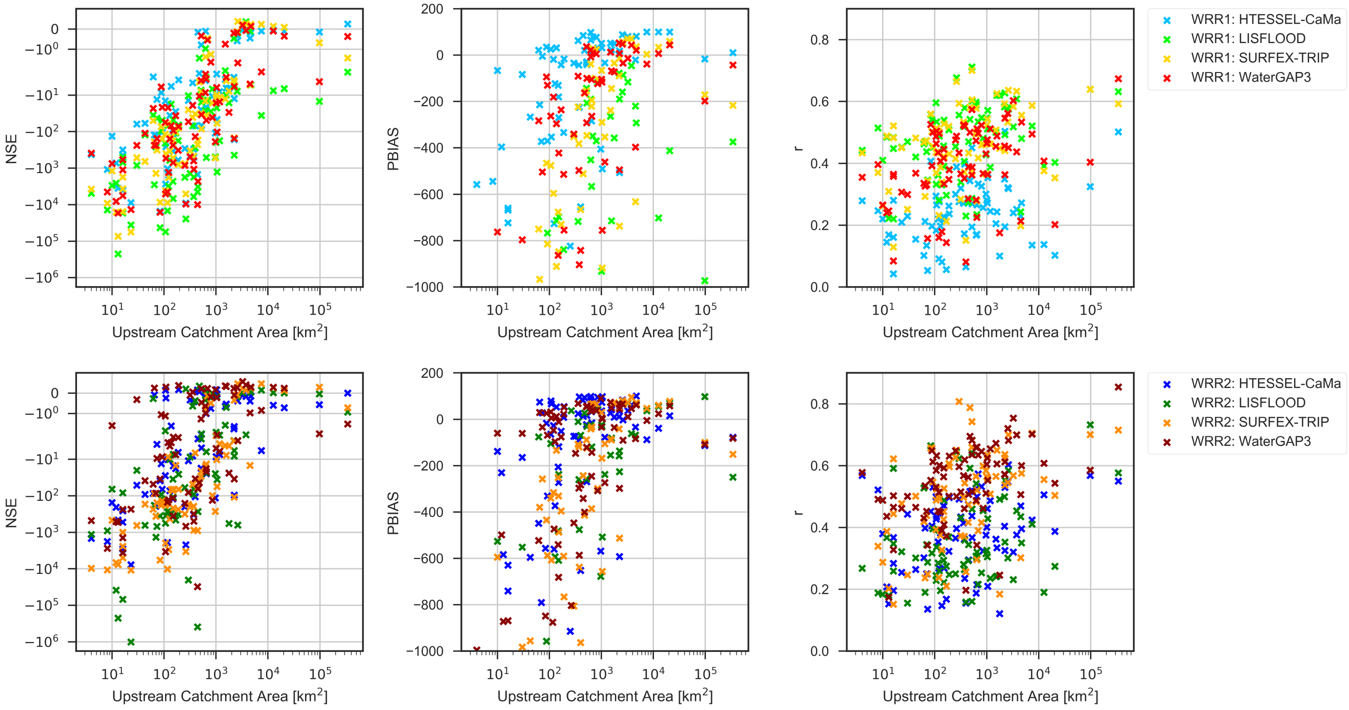


Figure 2. Performance statistics for the four models available in both WRR1 (top) and WRR2 (bottom) for each of the 75 gauging stations in the Limpopo River basin, ordered by upstream catchment area. The error statistics displayed include (a) the Nash Sutcliffe Efficiency (NSE) for WRR1, (b) the Percent Bias (PBIAS) for WRR1, (c) Pearson's r for WRR1, (d) NSE for WRR2, (e) the PBIAS for WRR2, and (f) Pearson's r for WRR2. For clarity, the lower limit of the y-axis of the PBIAS has been set to -1,000.

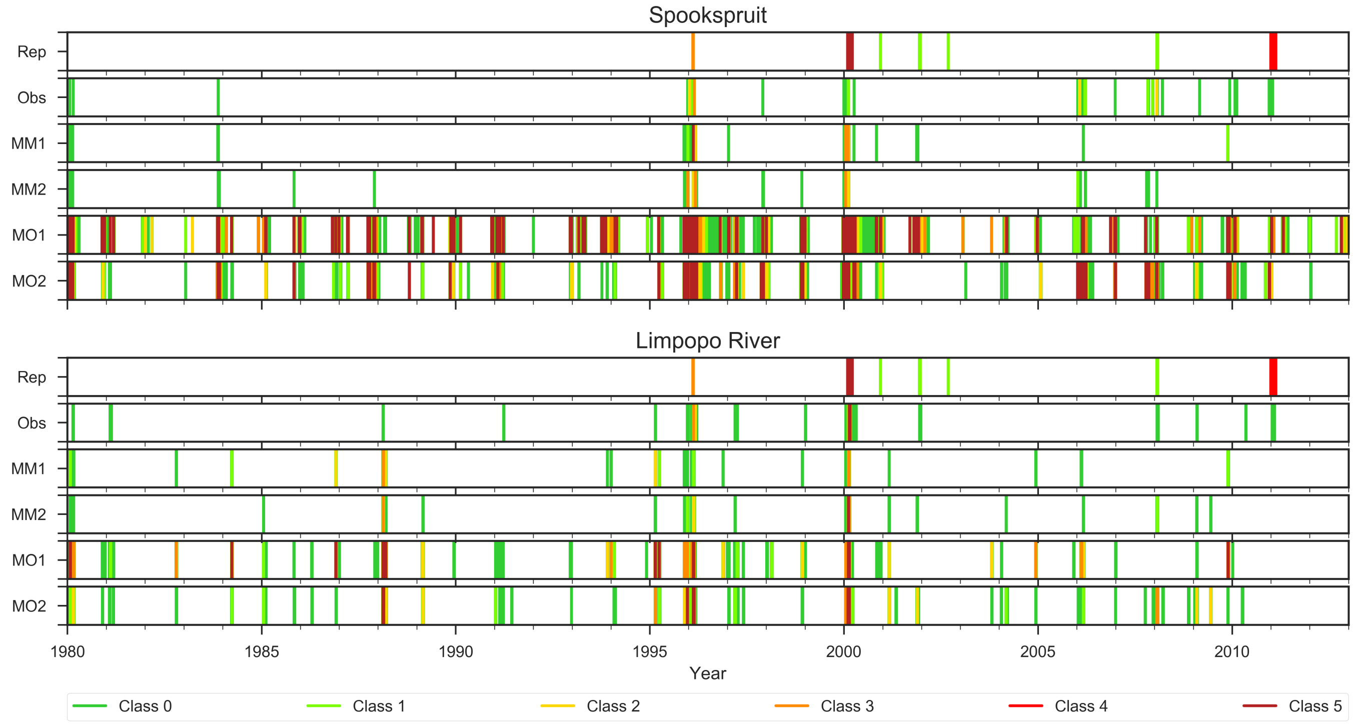


Figure 3. Occurrence of flood events of increasing severity classes at the Spookspruit gauge (252 km^2; upper panel) and in the main Limpopo River (98,240 km^2; lower panel). Model flood events were identified using model climatology (MM1 & MM2) or observed climatology (MO1 & MO2), and were compared to benchmarks based on a compiled disaster impact database (Rep) and observed river discharge data (Obs). The index value refer to models with 0.5 degree resolution (MM1 and MO1) and 0.25 degree resolution (MM2 and MO2). Results are shown for the WaterGAP3 model, which is available in the eartH2Observe Water Resources Re-analysis dataset.

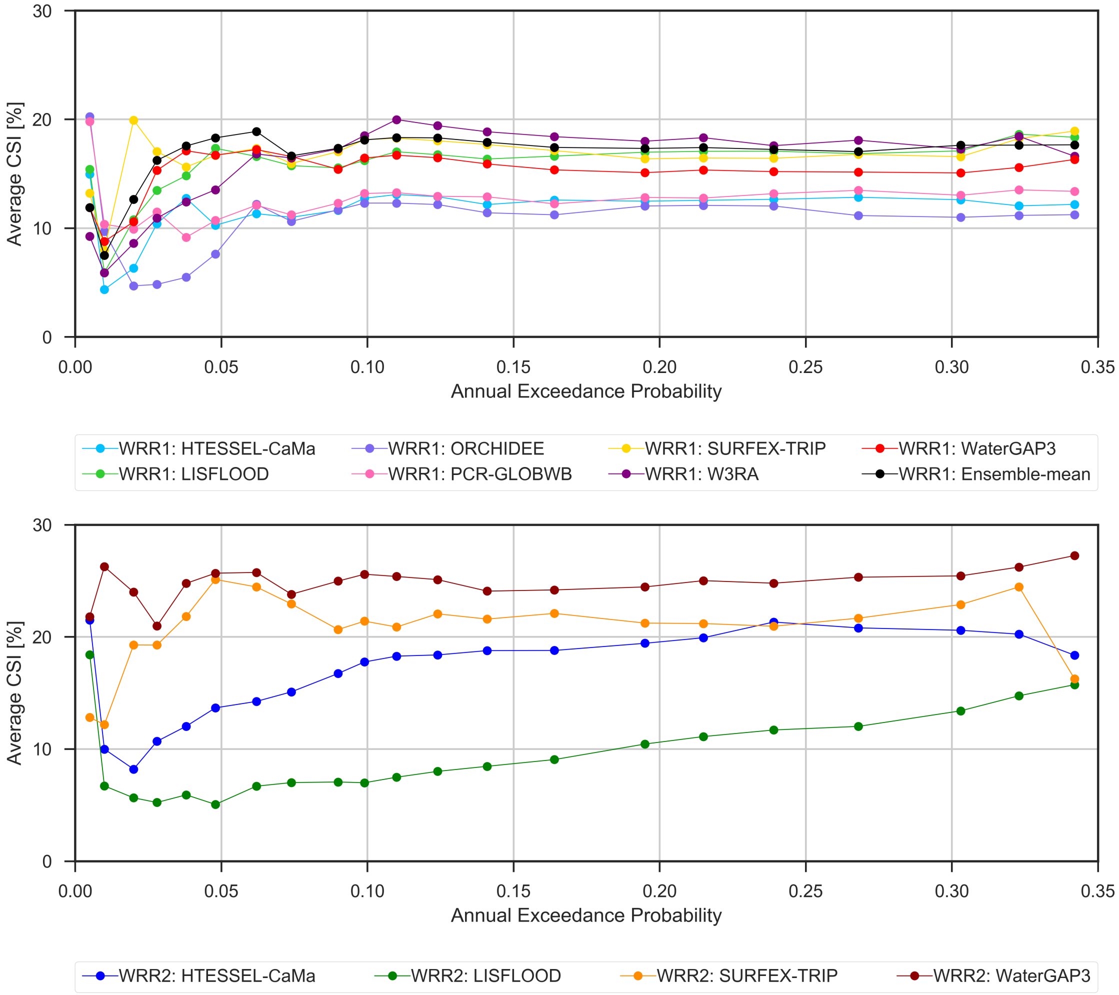


Figure 4. The Critical Success Index (CSI) using different annual exceedance probability thresholds averaged over all gauging stations for the seven models and ensemble mean available in WRR1 (upper panel), and the four models that are also available in WRR2 (lower panel).

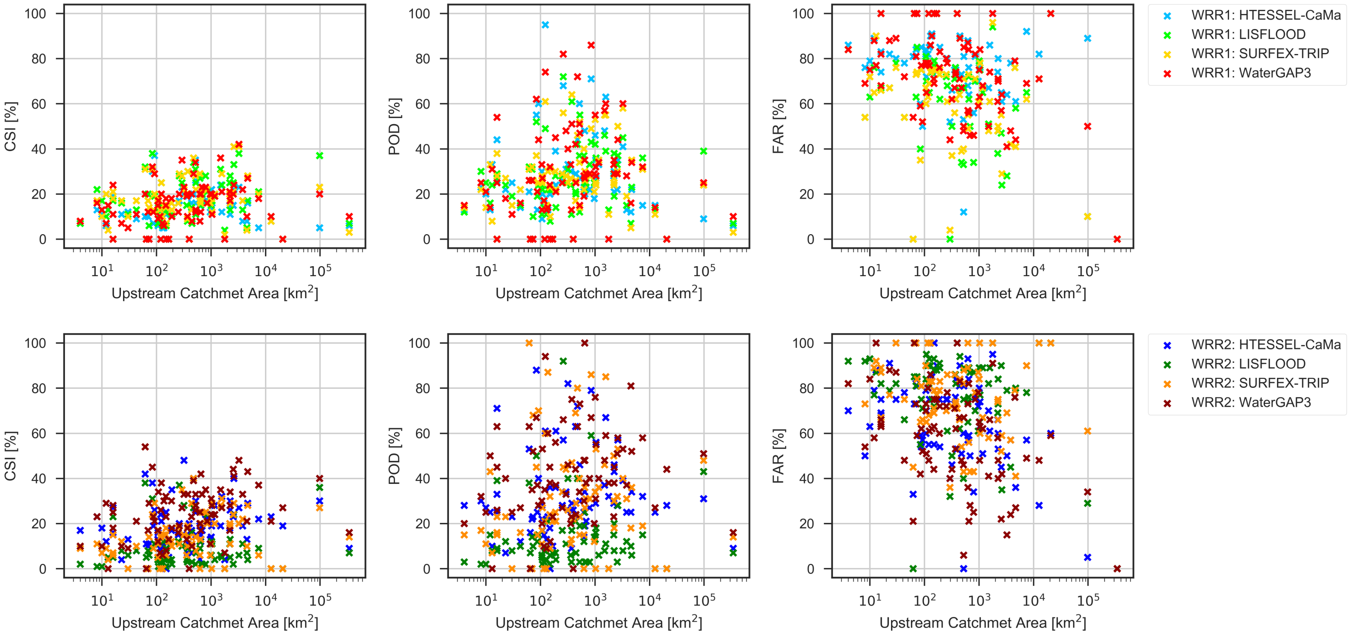


Figure 5. The Critical Success Index, Probability of Detection and False Alarm Ratio determined using the annual exceedance probability threshold of 0.164 (return period of 5 years) for all gauging stations for the three models available in WRR1 (upper panel), and the models that are also available in WRR2 (lower panel).

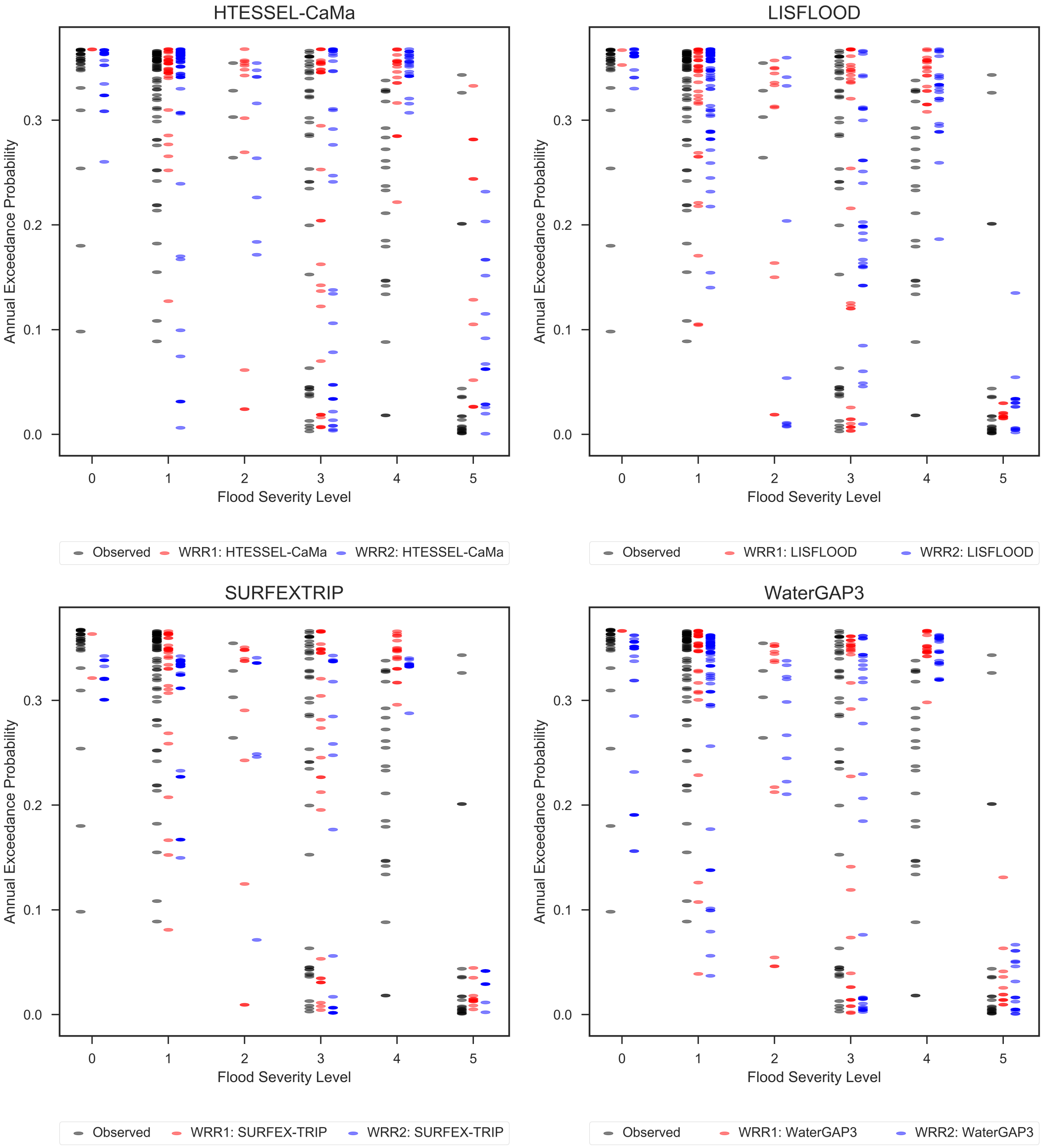


Figure 6. The relationship of the flood event severity for the reported flood events, and the corresponding annual exceedance probabilities that were observed and modelled for (a) HTESSEL-CaMa, (b) LISFLOOD, (c) SURFEX-TRIP and (d) WaterGAP3.

Table 1. Overview of the seven global models in the Water Resources Re-analysis dataset that include daily river discharges.

Model	Model Type	Changes in WRR2	Lakes-Reservoirs	Water use	Routing	Reference
HTESSEL-CaMa	LSM	Multi-layer snow scheme, increased no of soil layers.	No	No	CaMa-Flood	(Balsamo et al., 2009)
LISFLOOD	GHM	Increased no of soil layers, groundwater abstraction.	Yes	Yes	Double kinematic wave	(van der Knijff et al., 2008)
ORCHIDEE	LSM	N/A	No	No	Linear cascade of reservoirs	(Krinner et al., 2005)
PCR-GLOBWB	GHM	N/A	WRR1 only lakes	Not in WRR1	Travel time	(van Beek and Bierkens, 2009)
SURFEX-TRIP	LSM	Ground water, flood plains, land use, plant growth, surface energy and snow.	No	No	TRIP with stream	(Decharme et al., 2010)
WaterGAP3	GHM	Assimilation of soil water estimates, reservoir management.	Yes	Yes	Manning-Strickler	(Flörke et al., 2013)
W3RA	GHM	N/A	No	No	Cascading linear reservoirs	(van Dijk et al., 2014)
Ensemble of 7 models	GHM & LSM	N/A	Various	Various	Various	N/A

[*Source:* Schellekens et al. 2017; Dutra et al., 2015, 2017]

Table 2. Thresholds that were used to classify the exceedance probabilities according to flood severity levels.

Flood Severity Level	**Annual Exceedance Probability**	**Return Period [years]**
0	≤ 0.303	≥ 2
1	≤ 0.164	≥ 5
2	≤ 0.090	≥ 10
3	≤ 0.038	≥ 25
4	≤ 0.010	≥ 100
5	≤ 0.005	≥ 200

Table 3. Contingency table for flood events

		Observed	
		Yes	**No**
Modelled	**Yes**	Hits (H)	False Alarms (FA)
	No	Misses (M)	Correct Negatives (CN)

[*Source:* Thiemig et al., 2015]

Table 4. Performance statistics for the four models available in WRR1 and WRR2 (Table 4, upper panel) for each of the 75 gauging stations in the Limpopo River basin, and the three models and ensemble mean only available in WRR1 (Table 4, lower panel). Statistics displayed for WRR1 and WRR2 include the Nash Sutcliffe Efficiency (NSE), the Percent Bias (PBIAS) and Pearson's r (r). The three different upstream catchment areas indicate that an average is taken of the error statistics of the stations that are larger than indicated. Stations $\geq$ 4 km^2 indicates all 75 stations, $\geq$ 520 km^2 are the largest 31 stations, and $\geq$ 2,500 km^2 are the largest 11 stations.

	Stations with upstream catchment area [km^2]	**HTESSEL-CaMa**		**LISFLOOD**		**SURFEX-TRIP**		**WaterGAP3**	
		WRR1	*WRR2*	WRR1	*WRR2*	WRR1	*WRR2*	WRR1	*WRR2*
NSE	$\geq$ 4	-734.77	*-294.83*	-6,473.49	*-23,616.81*	-2,938.32	*-1,147.33*	-1,445.34	*-628.78*
	$\geq$ 520	-16.38	*-9.62*	-107.99	*-43.12*	-31.33	*-21.73*	-21.94	*-8.16*
	$\geq$ 2,500	-0.16	*-0.82*	-7.31	*-57.94*	-0.54	*-1.21*	-1.27	*-0.58*
PBIAS	$\geq$ 4	-595.10	*-335.27*	-6,359.73	*-9,996.76*	-2,415.21	*-1,176.01*	-1,680.21	*-889.04*
	$\geq$ 520	-17.96	*-25.19*	-987.01	*-361.25*	-243.26	*-167.95*	-143.29	*-32.72*
	$\geq$ 2,500	58.66	*-5.18*	-402.14	*-476.37*	-57.74	*-74.57*	-51.23	*-9.59*
r	$\geq$ 4	0.24	*0.38*	0.47	*0.35*	0.45	*0.50*	0.39	*0.54*
	$\geq$ 520	0.26	*0.42*	0.51	*0.37*	0.50	*0.56*	0.43	*0.60*
	$\geq$ 2,500	0.26	*0.47*	0.51	*0.41*	0.52	*0.60*	0.45	*0.66*

	Stations with upstream catchment area [km^2]	**ORCHIDEE**	**PCR-GLOBWB**	**W3RA**	**Ensemble mean**
		WRR1	WRR1	WRR1	WRR1
NSE	$\geq$ 4	-4,301.50	-176,842.51	-35,536,946.62	-5,645.16
	$\geq$ 520	-576.83	-220.57	-44,933.26	-59.92
	$\geq$ 2,500	-1.30	-8.80	-1,878.46	-1.17
PBIAS	$\geq$ 4	-7,049.09	-8,661.80	-235,229.53	-5,108.67
	$\geq$ 520	-1,954.28	-714.12	-21,748.70	-804.86
	$\geq$ 2,500	-103.64	-91.09	-5,014.81	-188.17
r	$\geq$ 4	0.32	0.12	0.31	0.46
	$\geq$ 520	0.34	0.13	0.33	0.49
	$\geq$ 2,500	0.31	0.13	0.36	0.49

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 #1:

General Comments
This paper assesses the potential of using re-analysis datasets with hydrological models to identify flood events in the Limpopo River basin. They evaluate climatological forcing's at 0.5 and 0.25 deg spatial resolution, and different hydrological and land surface models of the WRR datasets. While it is a model intercomparison paper, the objective is to identify timing and magnitude of floods. The novel aspect of the article is a flood detection comparison with reported observed flood damaged, which tries to link what is modeled to the actual impacts. The analysis focus on evaluating coarse spatiotemporal resolution dataset and models, which are not the most up to date and appropriated to assess local scale floods in small catchments. As the current generation of land surface and hydrological models are currently available at much higher resolution (i.e., 5-10-5 km), these models could potentially be more appropriated and yield better skill in detecting floods. Nonetheless, I understand that the authors are constrained by the data and models available at the WRR dataset. However, WRR could have been updated for a more novel study. The paper is clear and concise, and the authors acknowledge limitations on data, models, and analysis, and well as listed aspects for improvement. Despite the limitations, this study intended to inform the scientific community on the potentials and limitation of currently available data and model for flood applications.
We thank the referee for taking the time to review our manuscript thoroughly and for his/her constructive comments. We are pleased that the referee values our work and is generally positive. The referee provided helpful comments in order to improve our manuscript. We have considered each of the comments carefully, which we will address here in detail. For the ease of reading we have copied the referee comments (in bold), and respond to each of the comments below.

Specific comments
1. Page 2 Line 19 and Page 3 line 8: Can you expand the explanation of the term "spatially symmetrical"?
We thank the referee for pointing out that this was not clear. By observational data being "spatially symmetrical" we mean that the data is evenly distributed across the study area. In the case of the Limpopo River Basin most of the available data as well as resources (financial, institutional) are located in South Africa, thus the data is not spatially symmetrical, which is mentioned in the next sentence (page 2 lines 19-20). We will change this in our revised manuscript to "not evenly distributed across the riparian countries, with most gauges in South Africa".

2. Page 3 Line 21: I would say … managing floods at the regional and basin scales …. I'm not sure to what extent forecasting at 0.25deg resolution is aiding flooding management at local scales.
We thank the referee for the comment. As the scope of our paper is to assess the scale up to which global models are able to provide useful information for small-scale flood risk management in areas with insufficient observational data, we used the word "local". As this was not completely clear, we will modify the text to "… managing floods at the regional and sub-basin scales.".

3. Section 2.3.2 Disaster Data: I understand flood damage data is scarce and has its several limitations, which leads to data aggregation as an alternative to consolidate a standard analysis.

However, it would be interesting to see few point results for maybe one or two cases where the location and time of the flood events are reported, and how do the models perform regarding flood timing, magnitude, and detection. This additional analysis would bring more meaningful insights on the potential use of these models for flooding management and flood detection, rather than a sub-basin aggregation.
Indeed, we agree that it indeed would be highly interesting. We will investigate the possibilities for adding this aspect in our revised manuscript.

4. Section 3.1: The models, in the context they were applied in this study, were not designed to evaluate discharge and floods at small catchments with < 4km2, as the grid size is of at least _625 km2 (0.25deg). As an (expected) result, the timing, magnitude, and flood detection are poorly captured. As these models are not appropriated to be applied in small catchments, can you expand on what is the purpose of evaluating the small catchments in this study, why is it a reasonable approach, and which knowledge/information do you expect the scientific community will gain from it?
Many thanks for this thought-provoking point. 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).

5. Section 3.1: Coefficient of determination could be used instead of linear Pearson to represent how much of the variability can be represented by the proposed models.
Indeed, we agree that the coefficient of determination could have also been used, as they both describe the degree of collinearity between the modelled and observed data, though in a different way. The coefficient of determination describes the proportion of the variance in the measured data explained by the model, whereas Pearson's correlation coefficient is an index of the degree of linear relationship between observed and simulated data. We were interested in this linear relationship, in addition to the NSE and PBIAS, which is why we chose for Pearson's correlation coefficient above the coefficient of determination.

6. Page 11 Line 11. Can you expand on why do you think LISFLOOD perform worst using higher resolution forcing's data? I'd guess something related to model calibration.
We thank the referee for raising this interesting point. We have greatly looked into the reason why LISFLOOD could be performing worse in a higher resolution. There are a number of factors that could contribute to this. First of all, the models in WRR2 had further modifications in respect to WRR1, apart from the different forcing and higher resolution. For LISFLOOD the modifications include an increased number of lakes and reservoirs, improvement of irrigation water demand and groundwater abstraction and an increased number of soil layers (Arduini et al., 2017; Dutra et al., 2017). This could be part of

the answer as to why the improved forcing and higher spatial resolution did not result in an improved performance. Secondly, we agree with the referee that it could indeed also be related to the model calibration. LISFLOOD was somewhat calibrated for WRR1 using eleven parameters for 24 large catchments (Dutra et al., 2015), but LISFLOOD was not calibrated for WRR2 (Dutra et al., 2017). Instead, the parameterisation of WRR1 was also used for WRR2, even though the alterations to the model require an updated calibration (Arduini et al., 2017; Dutra et al., 2017). We will include the reasons for the worse performance of LISFLOOD in WRR2 in our revised manuscript.

7. This study was conducted considering data and models at daily time resolution, can you comment about the implications of temporal resolution on the forcing's data and modeling on the identification of short flashy floods. To what extent does it play a role in correctly identifying the flood category in places like southern Africa where rainfall if general driven by short and intense convective cells.
The forcing data of MSWEP is actually available at a 3-hourly resolution. However, as most models operate at the daily timestep (in the WRR dataset LISFLOOD, PCR-GLOBWB, W3RA and WaterGAP3) and since the output is also at the daily step, the forcing data with a sub-daily resolution is generally not used. Furthermore, sub-daily data is generally not available in these areas. Some of the rain- and discharge gauges report hourly values, but most only at the daily timestep. In order to analyse the short flash floods, even shorter reporting time-steps (every minute or 5-minutes) would be preferred. Apart from the high resolution data a very high resolution local model is also needed to model the short flash floods (López et al., 2016). For these models, the forcing data used in the WRR dataset would be insufficient. Our scope, however, is to look at the potential of available datasets at the global scale, and not to model at the high resolution the occurrence of flash floods. Indeed, these flash floods are likely not well captured in the data, and most research does not take this kind of flood events into consideration. That is the exact reason why we are also interested in including this in our research.

Technical Corrections / Minor Points
1. Page 2 Line 32: … determine climatic extremes as well as its uncertainties at global…
We thank the referee for pointing this out, the sentence will be altered.

2. Page 3 Line 20: as an illustration, I'd list some examples of currently available flooding forecast systems currently available.
The suggestion is highly appreciated, we will look into it.

3. Page 4 Line 15: it would be nice to see the location of dams and reservoirs mapped in Figure 1., as it expands our sense about the basin dynamics and importance of representation of lake and reservoirs in hydrological models.
We agree with this comment, we will add this in the revised version of our paper.

4. Page 14 Line 22: …whereas flood events occur at the basin or finer scales…
We assume the referee is referring to Page 13 Line 22: "whereas flood events occur at the basin scale". In this case, we agree with the referee and will modify this sentence text accordingly.

5. Page 15 Line 4. I'd say there is a critical need for both higher resolution re-analysis supporting data and flood forecasting systems to properly capture timing, intensity, and location of flood impacts.
We agree with the referee that there is a critical need for both the re-analysis data as well as flood forecasting systems. Such global models as used in this re-analysis are also employed in global forecasting systems, such as GloFas (Alfieri et al., 2013). GloFas uses the HTESSEL model which is also considered here, but then at a resolution of 0.1 degrees. We will include this in our revised manuscript.

6. Page 15 Line 22. I would change to something like: "This shows that some largescale models (i.e., WaterGAP3) have some skill in capturing observed and reported damaging floods, while others

perform poorly. The finds here presented, highlights the importance of model intercomparison and evaluation studies to inform the scientific community on model's strengths and weakness as well as plausible applications".

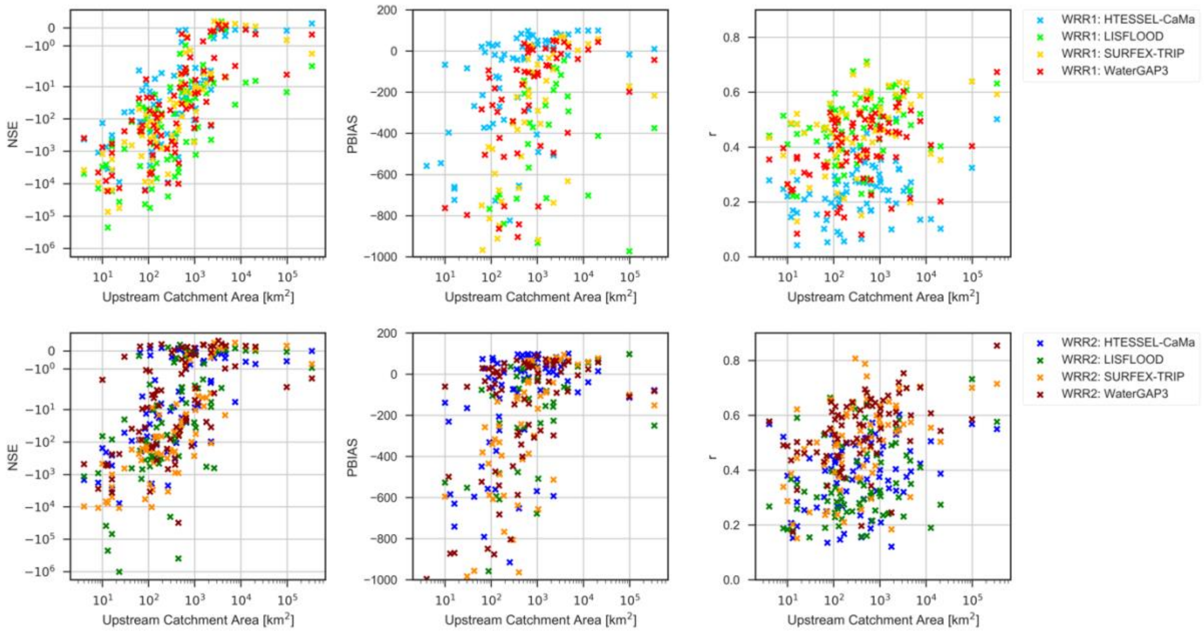
Many thanks for this comment, however, we do not fully agree here. In this particular study that focussed on the Limpopo basin WaterGAP3 performed better, but it is not sure if this is a general conclusion. There are many instances possible where other models would perform better. For instance, if this study would be repeated elsewhere, or if the discharge outputs of other models would be compared, or if the models would have had another forcing, or if all models would have been calibrated.

7. Figure 1: Use a different color for the basin delineation; include the other river tributaries of a lower order, especially the ones where gauges were evaluated; use a different color for the circle and square.

We thank the author for the suggestions to improve this figure, we will alter the figure where possible to include these points, as well as add the dams and/or reservoirs (as was pointed out in the Technical Corrections / Minor Points 3).

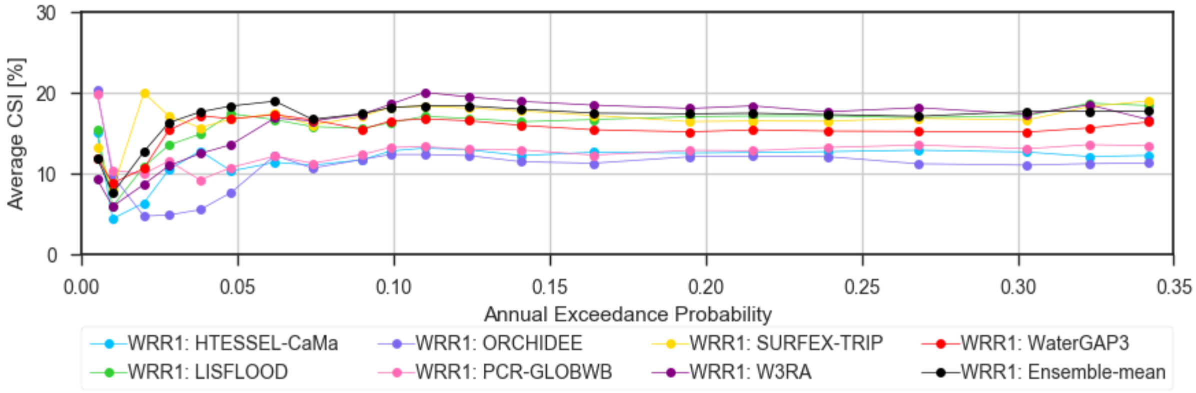
8. Figure 2: Maybe you can check if a log scale in the y-axis of the NSE plots would improve the representation of the values around zero.

Thank you for your comment, which was also raised by the second referee. We will take a closer look at improving the graphs for the NSE in this figure.



9. Figure 4: The 'X' points are very confusing and hard to follow through, maybe consider dots with a light line connecting them in the background.

Thank you, we have modified this figure, see below for WRR1. This indeed makes the figure clearer.



10. Figure 5: Full name of POD and FAR in the figure caption.
Thank you for pointing this out, we will include the full name in the paper.

11. Figure 6: Improve the figure quality regarding dpi. Change color scheme to opposite colors (i.e., red and blue) rather than light and dark (i.e., light blue and dark blue). Otherwise it's hard to see the differences.
Thank you for raising this issue, which was also mentioned by the second referee. We will change the figure in the revised version of the paper.

References used in this revision response:

Alfieri, L., Burek, P., Dutra, E., Krzeminski, B., Muraro, D., Thielen, J. and Pappenberger, F.: GloFAS-global ensemble streamflow forecasting and flood early warning, Hydrol. Earth Syst. Sci., 17(3), 1161–1175, doi:10.5194/hess-17-1161-2013, 2013.

Arduini, G., Fink, G., Martinez de la Torre, A., Nikolopoulos, E., Anagnostou, E., Balsamo, G. and Boussetta, S.: End-user-focused improvements and descriptions of the advances introduced between the WRR tier1 and WRR tier2., 2017.

Dutra, E., Balsamo, G., Calvet, J.-C., Minvielle, M., Eisner, S., Fink, G., Pessenteiner, S., Orth, R., Burke, S., van Dijk, A. I. J. M., Polcher, J., Beck, H. E. and Martinez de la Torre, A.: Report on the current state-of-the-art Water Resources Reanalysis., 2015.

Dutra, E., Balsamo, G., Calvet, J.-C., Munier, S., Burke, S., Fink, G., van Dijk, A. I. J. M., Martinez de la Torre, A., van Beek, R., de Roo, A. and Polcher, J.: Report on the improved Water Resources Reanalysis (WRR2)., 2017.

Huh, S., Dickey, D. A., Meador, M. R. and Ruhl, K. E.: Temporal analysis of the frequency and duration of low and high streamflow : years of record needed to characterize streamflow variability, J. Hydrol., 78–94, doi:10.1016/j.jhydrol.2004.12.008, 2005.

López, P.L., Wanders, N., Schellekens, J., Renzullo, L. J. and Sutanudjaja, E. H.: Improved large-scale hydrological modelling through the assimilation of streamflow and downscaled satellite soil moisture observations, Hydrol. Earth Syst. Sci., 20, 3059–3076, doi:10.5194/hess-20-3059-2016, 2016.

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.