Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC2
https://doi.org/10.5194/hess-2021-153-RC2, 2021

Comment on hess-2021-153

Paul Bates (Referee)

Referee comment on "Intercomparison of global reanalysis precipitation for flood risk modelling" by Fergus McClean et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-153-RC2, 2021

This paper looks at how four different global reanalysis precipitation products impact flood risk calculations made using a two-dimensional hydrodynamic model for five extreme floods which occurred in river basins in northern England. A national gauge-based gridded rainfall product (CEH-GEAR1hr) is used to compute the benchmark simulation, and comparisons relative to the benchmark are made in terms of inundated area, median floodplain depth, number of buildings inundated, peak stage at the most downstream gauge site in each catchment and absolute peak time error at the same site. In addition, all simulations are compared to the observed stage for each event at the downstream gauge sites.

This is definitely a useful thing to do, and whilst there have been a handful of previous papers on this topic, reanalysis datasets have advanced in recent years such that a new study is appropriate. Moreover, as errors in reanalysis products are likely to vary markedly in space and time, more information for different catchments, for different event types and using a wider range of metrics is needed.

However, I think there are a number of issues to address before publication. The main ones are:

- The key assumption underpinning the paper is that the national gauge-based rainfall product is significantly more accurate than the global reanalysis data, and results in flood simulations which can be considered as synonymous with (or at least much closer to) 'truth' than the other model realisations. If you cannot make this assumption, then the analysis is reduced in value significantly. Whilst there may be some good reasons to believe local gauge-based products are better, these are never explicitly discussed or backed up with evidence. I think there thus needs to be a robust discussion of the likely quantitative errors in CEH-GEAR1hr. A particular worry in this respect is Figure 4, which compares downstream observed river stage hydrographs to the same quantity in

simulations driven by the reanalysis and benchmark precip data. However, it is not immediately clear to me from Figure 4 that the CEH-GEAR1hr simulation is significantly better than the simulations using ERA-5, MERRA-2 and CFSR. JRA-55 is clearly poor, but the other reanalysis products seem to do quite a good job given other errors in the modelling process. For the paper to be viable I would want to see more compelling evidence and arguments that the benchmark data really does provide a definitive point of reference. Figure 4 is the only absolute test of this in the paper and the results are not obviously conclusive. Properly quantifying the model performance shown in this figure with a basket of metrics including NSE and RMSE will be important and will perhaps show what I am missing just by eyeballing the plots. A wider range of other absolute measures of model performance (other gauge sites, flows as well as just stage, inundation observations if available) would also help convince the reader that the benchmark is robust.

- A second issue is that the analysis jumps straight to comparisons of hydrodynamic model output, and whilst this is interesting, I think the paper is missing a trick by not first simply analysing the differences between the various rainfall products. This should explain a lot about the differences in model performance that then follow. At present only Figure 2 really does this, but it is not a detailed enough dive into the differences between the precipitation data sets. As well as CEH-GEAR1hr I would also have liked to see data from the individual rain gauges across the study catchments and how the gridded products compare to these.
- I felt the paper was missing a lot of background information that I was expecting. Individually, each bit of missing information is minor, but taken as a whole I'm not able to really understand key aspects of how the analysis was undertaken. Just in terms of the model as one example, the paper is missing information on the numerical scheme, the grid resolution (I assume this is the same as the terrain data but you never say), how the river channels are handled (or not) and information about model boundary conditions etc. There are lots more comments like this below and they all need sorting out.
- I was just a little bit underwhelmed by the volume of analysis given the amount of effort that has obviously been spent wrangling the data into shape. The model simulations are inter-compared in terms of only a handful of metrics which are either aggregated over the whole area or over all events or are for just a single location in each catchment. I felt you could have exploited the hard work you have undertaken a lot more effectively and that this would have told a richer and more interesting story. There are many more gauge sites within each catchment for example, and many of these also record flows. In particular, more absolute validation is, I think, essential to increase confidence in this study.
- For Figure 4, I don't see how you can predict stage accurately when you don't seem to have river channels explicitly in your model. You do not mention bathymetry or channel data, and the model grid appears to 50m resolution which will not resolve the channels. I don't think the OS Terrain 50 data you use contains the channel geometry either. Maybe I am missing something, but if you do not have the channels explicitly represented then I don't see how you are able to simulate a reasonable stage-discharge curve at the gauge sites?
- On a similar note, the assumption that the subsurface hydrology is not important during these events is quite a big leap. This and point 5, would be (just about) fine if you were just doing a relative comparison, but to drive home the message in the paper you really do need to demonstrate that the benchmark is fundamentally better through an absolute validation. Given the lack of (i) channels and (ii) subsurface hydrology how does the model even get close to simulating stage correctly as shown in Figure 4? I completely accept that it does, but it seems counter-intuitive. Are there some compensating errors going on perhaps?

In addition, the following minor points also need attention:

- Line 28. For general readers it would be helpful to briefly explain what reanalysis products are and how they are constructed.
- Line 28. Please define large-scale.
- Line 31. 'vast' is not typical scientific language. 'Extensive' would be better.
- Line 42, Define and explain VIC. General readers will not know what this is.
- Line 51. Please state what Winsemius et al found in their study.
- Line 56. This is being a bit picky, but the claim here is that the results of this study are transferable to other areas bears closer examination. Is there any evidence that this is really going to be the case? The review in the paragraph above shows that reanalysis errors are complex in time and space, so this might indicate that the results of the present study are much less transferable than this statement supposes. You've only looked at five events in one particular part of the world, so it is not a very large sample.
- Line 79. I wasn't sure what you meant by 'uniformly gridded' here. Do you mean that a regular grid geometry is used, or do you mean that each grid cell has a single uniform elevation value (i.e. what would be a p0 discretization in a finite element model)?
- Line 82. To what extent do rivers in HydroBASINS line up with the valley structures in the OS DEM data. HydroBASINS was generated from SRTM so my working assumption would be that there are some areas of significant mismatch between the hydrography and the terrain. This is pretty much inevitable when you mix products from different terrain data sources.
- Line 87. There is lots of information about the model set up missing from this section. See comments above.
- Line 96. Is using the most extreme events the best research design? Would a mix of event types and magnitudes have been better? Extreme events tend to be valley filling which means some of your metrics may have reduced sensitivity.
- Line 101. But the land surface is assumed impermeable so how does antecedent rainfall affect the model? This statement seems at odds with the physics the model includes.
- Table 2. Some more information on these events would be useful. Climatology, return period or rainfall and flow, dynamics etc.
- Line 112. This needs a detailed discussion of likely errors in the benchmark. You have not conclusively established that it is fit for purpose.
- Line 118. Some more quantitative detail about what we already know about the reanalysis errors is needed here. There is likely to be a lot of this, so it needs to be summarised effectively. So far you just have qualitative statements.
- Figure 2. The reanalysis ensemble mean would be interesting, and the ensemble of ERA-5, MERRA-2 and CFSR.
- Figure 2. Is this the sum rainfall from all events? Might it not be better to pick a single event as an example, and have similar plots for the other events in SI?
- Line 145. The statement that the DEM is based on airborne LiDAR cannot be true for the whole area, can it? I did not think we yet had complete LiDAR coverage of upland areas. In your previous paper you say OS terrain 50 has vertical RMSE of 4m compared to ground control points, but LiDAR data are typically accurate to <10cm. How do you reconcile this if OS Terrain 50 is LiDAR-based? Why didn't you use the available bare earth Environment Agency LiDAR where available? Lastly, how can you predict stage (cf. Figure 4) well with DEM data that have 4m vertical error?
- Line 147. So, are the DEM resolution and the model resolution the same? What do you do about channels?
- Table 3. Some more information on the observed hydrograph data would be helpful. The circles in Figure 1 aren't really enough. What do the flow hydrographs look like at

different gauge sites in the catchments and what are the event return periods?

- Table 3. How did you calculate the inundation metrics? Particularly, how did you define the floodplain areas?
- Table 3. Why is the peak Q error so much bigger than the inundation error? Is this because these events are largely valley filling?
- Line 168. Panels a-e in Figure 3 are too small to be able to see this detail so the reader cannot verify these statements for themselves. Needs fixing.
- Line 174. Why is out of bank flow an issue? I did not think you have channels in the model so this is a bit odd. In fact, how the model predicts stage without channels explicitly represented is surprising. See comments above on this.
- Figure 3. The grid lines in panel f undermines the clarity of this diagram.
- Figure 4. Why do some panels have a zero on the y axis and others do not?
- Figure 5. These numbers are not that different apart from JRA-55. Is this because the events are valley filling? In which case number of buildings inundated may not be a great choice of metric. Loss might have been a better one as that has a depth dependency.
- Line 192. It would be helpful to explain this underestimation bias in physical terms.
- Discussion and conclusions. These sections will need careful editing once my comments have been dealt with as this might change the inferences that can be drawn from the work.
- Line 142. You do not compare to river flow in this paper, so this statement seems out of place.

In summary, this is an interesting paper but a bit more work is needed to make it acceptable for publication. Hopefully addressing the above comments will strengthen the work significantly.