Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., author comment AC2
https://doi.org/10.5194/hess-2021-153-AC2, 2021

Reply on RC2

Fergus McClean et al.

Author 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-AC2, 2021

RC2.1 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.

RC2.1 raises the issue of whether a gauge-based product is a reliable benchmark to compare reanalysis products against. Gauge-based products use observations of rainfall whereas reanalysis products use observations of wind, pressure, temperature, and humidity to drive atmospheric simulations and predict rainfall. Therefore, an assumption that gauge-based products are, at the point of measurement, closer to the 'truth' in terms of rainfall values is reasonable and the use of such "best-available" data sets is well established in the literature (Jiang et al., 2021; Sun & Barros, 2010; Lei et al., 2021). There are of course errors in any rain gauge observations, such as wind-induced under-catch (Pollock et al., 2018), and the network itself may not be dense enough to capture a

given storm event effectively. However, given the lack of alternatives, CEH-GEAR1hr clearly provides the best option here. We will clarify this, pointing towards relevant literature, in a revised version.

The paper seeks to provide a comparison rather than a validation of outputs and therefore the reliability of the benchmark simulation was not seen as being critical. The aim was to highlight the differences between reanalysis products and how they compared with a gauge-based product in the context of flood modelling. However, it is agreed that the paper would benefit from further absolute validation, therefore results will be compared with flood extent observations where available.

The average peak error metrics are currently skewed by the Wear level gauge, which does not capture high flows accurately ("Flows go out of bank at levels greater than 3m and so peaks are truncated above this level" https://nrfa.ceh.ac.uk/data/station/info/24009.html). To address this, Table 3 will be disaggregated to provide metrics for each individual gauge rather than averages across the gauges. This will provide a more conclusive demonstration that using CEH-GEAR1hr leads to more accurate estimates of stage peaks.

RC2.2 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.

A more in-depth statistical analysis of the rainfall products and comparison at specific rainfall gauges will be added to understand the spatio-temporal differences and elevation dependence and provide better context for the hydrodynamic model output.

RC2.3 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.

RC2.3 will be addressed by adding more information about the numerical scheme, grid resolution, river channels and boundary conditions . This issue was also raised by RC1.2.

R2.4 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.

RC2.4 requests further absolute validation and less aggregation of results. As mentioned above, the focus of the paper is a comparison rather than an absolute validation. However, we will add further analysis to assess extent performance relative to

observations in a revised paper. The most downstream gauges were used in each catchment as the upstream areas are largest, and therefore more of the reanalysis rainfall data is included in the modelling. The over-aggregation issue has also been discussed above in RC2.1 and will be addressed by reporting metrics for each basin individually.

RC2.5 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?

RC2.6 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?

RC2.5 and RC2.6 ask for more clarification about how the models can predict stage accurately given the lack of channel and subsurface representation. Channel representation is less important once flows are out of bank during large flood events as most of the water will not be within the channels . For example, Neal et al (2021) found that changing the bathymetry estimation method had less effect on flood volume and extent at more extreme return periods (100-year vs 5-year). Dey et al (2019) also observed that "the choice of bathymetric model becomes irrelevant at high flows for predicting hydraulic outputs". Cook & Merwade (2009) report a difference of only 0.35m in water surface elevation when including channel bathymetry in a 10m USGS DEM of the Brazos River using HEC-RAS. Therefore, the observed stage can still be captured despite the lack of channel bathymetry.

A similar explanation can be made about the subsurface. Once the ground is saturated, subsurface processes will cease to have a large impact on water levels. This means that during long duration flood events, the ground can be assumed to be impermeable, especially for catchments such as those addressed here with generally shallow soils and low base flow indices. Hossain Anni (2020) found an increase of only 0.02m in average flood depth when excluding infiltration from a model of the 100-year flood (University of Alabama campus). It is also possible that some compensating errors are present with numerical dispersion and underestimation of rainfall counteracting the effect of missing infiltration. Further clarification will be added to the text.

RC2.7 Line 28. For general readers it would be helpful to briefly explain what reanalysis products are and how they are constructed.

To address RC2.7, an explanation of what reanalysis products are and how they are constructed will be added.

RC2.8 Line 28. Please define large-scale.

To address RC2.8, more explanation will be provided about what is meant by large-scale in a revised version, i.e. continental and global analysis of flood risk.

RC2.9 Line 31. 'vast' is not typical scientific language. 'Extensive' would be better.

We agree, vast will be reworded to extensive on line 31 (RC2.9).

RC2.10 Line 42, Define and explain VIC. General readers will not know what this is.

VIC will be described in more detail (RC2.10).

RC2.11 Line 51. Please state what Winsemius et al found in their study.

The findings from Winsemius et al (2013) will be summarised (RC2.11), i.e. river flood risk maps and damage estimates produced using ERA40 and ERA-Interim were found to be in the same order of magnitude as estimates from EM-DAT and the World Bank.

RC2.12 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.

We don't feel this is picky and agree it is important to provide more appropriate qualification. The statement on line 56 (RC2.12) about transferability will be properly qualified, i.e. this study provides an example of how varied results may be between products, however the relative performance of each product may differ between areas and events.

RC2.13 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)?

"Uniformly gridded" on line 79 means a uniform square mesh equivalent to the DEM surface, this will be clarified (RC2.13).

RC2.14 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.

RC2.14 identifies that the HydroBASINS catchment boundaries may not exactly match the catchment boundaries in OS Terrain. This is true; however, they provide a reasonable estimate and any discrepancy will not have a significant effect on results given the 50m resolution of the model grid .

RC2.15 Line 87. There is lots of information about the model set up missing from this section. See comments above.

As noted in response to earlier comments (RC1.2, RC2.3) we will provide the additional model set up information.

RC2.16 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.

As stated in RC2.16, looking at a wider range of event types may have provided different results. This will be stated in the text and alternative designs considered. The largest available events for each basin were used to test how well extreme events are captured in

the reanalysis products.

RC2.17 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.

Antecedent rainfall is necessary to initiate normal flow in river channels. If no spin-up period is included, then flood magnitudes are underestimated . This will be clarified in the text to address RC2.17.

RC2.18 Table 2. Some more information on these events would be useful. Climatology, return period or rainfall and flow, dynamics etc.

More detail on climatology, return period etc. will be provided about the events in Table 2 (RC2.18)

RC2.19 Line 112. This needs a detailed discussion of likely errors in the benchmark. You have not conclusively established that it is fit for purpose.

Further discussion of errors in CEH-GEAR1hr will be added to address RC2.19, including a description of validation that was undertaken when producing the data (Lewis et al., 2018).

RC2.20 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.

More quantitative metrics about reanalysis errors will be added where available to address RC2.20.

RC2.21 Figure 2. The reanalysis ensemble mean would be interesting, and the ensemble of ERA-5, MERRA-2 and CFSR.

A mean of both all reanalysis products and ERA-5, MERRA-2 & CFSR will be added to address RC2.21.

RC2.22 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?

Figure 2 currently shows a different event for each basin which is admittedly confusing, as identified by RC2.22. This will be modified so that each event in each basin is shown within its own subplot (25 in total). Some plots may be moved to supplementary information if necessary.

RC2.23 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?

The OS Terrain 50 documentation is not very clear about the source of the data. It was originally thought that the product was based on LIDAR, but on further investigation, it has been identified that the DEM was created using photogrammetry and topographical surveys. This will be corrected in the text to address RC2.23. The EA LIDAR data was not

used as it does not provide full coverage of the basins and combining products may lead to artefacts. Absolute elevation is not as important as topography (i.e. relative elevation) when simulating flows, which may explain why the 4m vertical error does not prevent river stage being simulated. We will add discussion from other studies that have undertaken DEM uncertainty analysis that show OS Terrain 50 performs well (e.g. Yunus et al. (2016) shows a 3-10% larger inundation estimate in London when compared to 1m LIDAR).

RC2.24 Line 147. So, are the DEM resolution and the model resolution the same? What do you do about channels?

In answer to RC2.24, the DEM and model resolution are the same. Channels are not included in the DEM surface.

RC2.25 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?

More details about the observed hydrograph data will be added to address RC2.25. This will include their ID numbers, locations, catchment areas and time step, as identified in the response to RC1.1. The event return periods and flow hydrographs will also be incorporated.

RC2.26 Table 3. How did you calculate the inundation metrics? Particularly, how did you define the floodplain areas?

In answer to RC2.26, the inundation error is the relative difference in buildings inundated between each reanalysis product and CEH-GEAR1hr. The floodplain depth error is a comparison of each grid cell with the corresponding cell in the model using CEH-GEAR1hr. This is also plotted in Figure 3F. The meaning of each metric will be clarified in a revised manuscript.

RC2.27 Table 3. Why is the peak Q error so much bigger than the inundation error? Is this because these events are largely valley filling?

The explanation for RC2.27 (why peak stage error is larger than inundation error), given on L165, is that the same DEM is used in all models. As stated in the caption of Table 3, building inundation error is relative to the CEH-GEAR1hr simulation, whereas the peak error is relative to observed stage. As building inundation error is based on results from another model using the same DEM, it is expected to be lower than stage peak error. The fact that extreme events are likely to be valley-filling further contributes to a reduced variability of inundation error.

RC2.28 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.

The size of panels a-e in Figure 3 will be increased to address RC2.28.

RC2.29 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.

In answer to RC2.29, the out of bank flow is an issue for the river gauge measurements rather than the model, this will be made clearer in a revised manuscript.

RC2.30 Figure 3. The grid lines in panel f undermines the clarity of this diagram.

The grid lines will be removed from panel F of Figure 3 to address RC2.30.

RC2.31 Figure 4. Why do some panels have a zero on the y axis and others do not?

In answer to RC2.31, there is no reason for some plots having zeros and others not in Figure 4, this will be resolved in an updated figure.

RC2.32 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.

The reanalysis products (excluding JRA55) are 14-18% lower than CEH-GEAR1hr on average, which is seen as being significant. The units of the y axis are thousands of buildings (shown at the top left of the plot), this could be made clearer to the reader. Results comparing inundation across a range of flood depths will be reported in a revised manuscript to address RC2.32.

RC2.33 Line 192. It would be helpful to explain this underestimation bias in physical terms.

A physical explanation of the underestimation bias will be added to address RC2.33. The underestimation of extreme rainfall events by reanalysis products has previously been identified in the literature (Blacutt et al., 2015; He et al., 2019; de Leeuw et al., 2015). One contributing factor is that the model grid resolution of the global climate models (GCMs) used may not be high enough to capture the dynamics of extreme events. Seasonal and local characteristics may also not be captured by the GCMs. Any resulting negative bias in precipitation propagates into flood depths and impacts as less water enters the hydrodynamic model and accumulates on the floodplain.

RC2.34 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.

We will of course update the discussion and conclusions sections to reflect the material that we incorporate in a revised manuscript (RC2.34).

RC2.35 Line 142. You do not compare to river flow in this paper, so this statement seems out of place.

Thank you for spotting this, we will change this to refer to river stage, which we do compare, to address RC2.35.

Blacutt, L.A., Herdies, D.L., de Gonçalves, L.G.G., Vila, D.A. & Andrade, M. (2015) Precipitation comparison for the CFSR, MERRA, TRMM3B42 and Combined Scheme datasets in Bolivia. Atmospheric Research. 163117–131.

Cook, A. & Merwade, V. (2009) Effect of topographic data, geometric configuration and modeling approach on flood inundation mapping. Journal of Hydrology. 377 (1–2), 131–142.

Dey, S., Saksena, S. & Merwade, V. (2019) Assessing the effect of different bathymetric models on hydraulic simulation of rivers in data sparse regions. Journal of Hydrology. 575 (March), 838–851.

He, S., Yang, J., Bao, Q., Wang, L. & Wang, B. (2019) Fidelity of the

observational/reanalysis datasets and global climate models in representation of extreme precipitation in East China. Journal of Climate. 32 (1), 195–212.

Hossain Anni, A., Cohen, S. & Praskievicz, S. (2020) Sensitivity of urban flood simulations to stormwater infrastructure and soil infiltration. Journal of Hydrology. 588 (May), .

Jiang, Q., Li, W., Fan, Z., He, X., Sun, W., Chen, S., Wen, J., Gao, J. & Wang, J. (2021) Evaluation of the ERA5 reanalysis precipitation dataset over Chinese Mainland. Journal of Hydrology. 595 (October 2020), 125660.

de Leeuw, J., Methven, J. & Blackburn, M. (2015) Evaluation of ERA-Interim reanalysis precipitation products using England and Wales observations. Quarterly Journal of the Royal Meteorological Society. 141 (688), 798–806.

Lei, H., Li, H., Zhao, H., Ao, T. & Li, X. (2021) Comprehensive evaluation of satellite and reanalysis precipitation products over the eastern Tibetan plateau characterized by a high diversity of topographies. Atmospheric Research. 259 (January), 105661.

Lewis, E., Quinn, N., Blenkinsop, S., Fowler, H.J., Freer, J., Tanguy, M., Hitt, O., Coxon, G., Bates, P. & Woods, R. (2018) A rule based quality control method for hourly rainfall data and a 1 km resolution gridded hourly rainfall dataset for Great Britain: CEH-GEAR1hr. Journal of Hydrology. 564 (June), 930–943.

Neal, J., Hawker, L., Savage, J., Durand, M., Bates, P. & Sampson, C. (2021) Estimating River Channel Bathymetry in Large Scale Flood Inundation Models. Water Resources Research. 57 (5), 1–22.

Pollock, M.D., O'Donnell, G., Quinn, P., Dutton, M., Black, A., Wilkinson, M.E., Colli, M., Stagnaro, M., Lanza, L.G., Lewis, E., Kilsby, C.G. & O'Connell, P.E. (2018) Quantifying and Mitigating Wind-Induced Undercatch in Rainfall Measurements. Water Resources Research. 54 (6), 3863–3875.

Sun, X. & Barros, A.P. (2010) An evaluation of the statistics of rainfall extremes in rain gauge observations, and satellite-based and reanalysis products using universal multifractals. Journal of Hydrometeorology. 11 (2), 388–404.

Yunus, A.P., Avtar, R., Kraines, S., Yamamuro, M., Lindberg, F. & Grimmond, C.S.B. (2016) Uncertainties in tidally adjusted estimates of sea level rise flooding (bathtub model) for the greater London. Remote Sensing. 8 (5), .