Intercomparison of global reanalysis precipitation for flood risk modelling

Fergus McClean[1], Richard Dawson[1], Chris Kilsby[1,2]

[1]School of Engineering, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK

5 [2]Willis Research Network, 51 Lime St., London, EC3M 7DQ, UK

Correspondence to: Fergus McClean (fergus.mcclean@newcastle.ac.uk)

Abstract. Reanalysis datasets are increasingly used to drive flood models, especially for continental and global analysis, and in areas of data scarcity. However, the consequence of this for risk estimation has not been fully explored. We investigate the impact of using four reanalysis products (ERA-5, CFSR, MERRA-2 and JRA-55) on simulations of historic flood events in

10 Northern England. These results are compared to a benchmark national gauge-based product (CEH-GEAR1hr). The benchmark demonstrated better accuracy than reanalysis products when compared with observations of water depth and flood extent. All reanalysis products predicted fewer buildings would be inundated by the events than the national dataset. JRA-55 was the worst by a significant margin, underestimating by 40% compared with 14-18% for the other reanalysis products. CFSR estimated building inundation the most accurately, while ERA-5 demonstrated the lowest error in terms of

15 river stage (29.4%) and floodplain depth (28.6%). Accuracy varied geographically and no product performed the best across all basins. Global reanalysis products provide a useful resource for flood modelling where no other data is available, but they should be used with caution due to the underestimation of impact shown here. However, given the lack of observed building inundation data and limited spatial coverage, the conclusions that can be drawn by this study are limited. Until a more systematic international strategy for the collection of rainfall data ensures more complete global coverage of validation data,

20 multiple reanalysis products should be used concurrently to capture the range of uncertainties.

1 Introduction

The primary drivers of pluvial and fluvial flooding are precipitation events. The duration, intensity and spatial extent of these events can all affect the depth and extent of any flooding caused. Therefore, the choice of precipitation data when simulating floods is critical. Inaccurate precipitation will undoubtedly lead to a spurious and potentially misleading understanding of the

25 risk posed by a given event. This effect is further exacerbated when low-quality precipitation data is used to project risk into the future, with planning decisions being made based on the results. Unfortunately, understanding which source of precipitation is most appropriate is challenging. There is also spatial variation in the availability and quality of precipitation data. High-quality data is often collected by national or regional authorities but can be inaccessible or difficult to obtain, therefore continental or global precipitation datasets, such as reanalysis products, are a popular option despite their generally

30 lower resolution and accuracy.

Reanalysis products are created by driving numerical models with recorded weather observations to build a comprehensive historical picture of a wide range of climatic variables. These datasets usually have global coverage and span multiple decades. Reanalysis precipitation data has been widely used in continental- and global-scale ~~large scale~~ flood risk modelling (Winsemius et al., 2013; Alfieri et al., 2013; Andreadis et al., 2017; Pappenberger et al., 2012; Xu et al., 2016; Seyyedi et al., 2015; Schumann et al., 2013). The main advantages of reanalysis products are their ~~vast~~ extensive spatiotemporal coverage and ease of access. In areas with a limited number of rain gauges that can provide high-quality observations, reanalysis products are often the best or only source of precipitation inputs for flood simulations. However, there is no guarantee that they are able to accurately represent extreme events and subsequently characterise flood risk. As a range of reanalysis datasets are available, there is also the question of which is more suitable for the application.

The influence of reanalysis data on flood risk estimates has previously been explored in part. Sampson et al. (2014) found that the loss ratio decreased by 8.5 times when using a reanalysis product (ERA-Interim) instead of a satellite rainfall product (CMORPH) in their catastrophe risk model of Dublin. Andreadis et al. (2017) compared flood models driven using an ensemble of parameters from 20CRv2 to a benchmark using observed flow boundary conditions and found that, overall, 20CRv2 only captured 15.7% of the benchmark inundated area. Mahto et al. (2019) used ERA-5, ERA-Interim, CFSR, JRA-55 and MERRA-2 to drive ~~VIC~~ a macroscale hydrological model and simulate monsoon season runoff in India. CFSR and JRA-55 resulted in a strong positive bias, compared to a national precipitation dataset, while MERRA-2 strongly underestimated runoff. The two ERA products showed a much less prominent positive bias. While their study represents one of the first intercomparisons of different reanalysis precipitation products for runoff modelling, it does not go as far as looking at the consequences for flood impacts. Meanwhile, Chawla et al. (2020) demonstrate a strong negative bias in flood discharge when using CFSR in the Himalayas. This is indicative of the spatial variability in accuracy inherent in reanalysis datasets, driven largely by assimilation data availability. Winsemius et al. (2013) compared flood impacts from the GLOFRIS model cascade, which uses ERA-Interim, with The OFDA/CRED International Disaster Database (EM-DAT). 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. However, the effect of using a different source of precipitation was not assessed and therefore the impact of using reanalysis data on the cascade is unknown.

This paper extends previous studies by undertaking a systematic intercomparison of how modern reanalysis products compare when used to drive a hydrodynamic flood model. This provides important insights to inform the selection of data for flood modelling in data-sparse regions as well as a more general assessment of how well extreme rainfall events are captured in each product. To provide further context and identify the potential effects on flood risk assessments, the flood model outputs are subsequently used to estimate the number of buildings that would be inundated by each rainfall product. While this study provides an example of how varied results may be between products, the relative performance of each dataset may differ between areas and events and is not necessarily transferable.

2 Methodology

2.1 Study Area

To assess the performance of global reanalysis precipitation, more reliable gauge-based data is required as a baseline to validate against. However, the quantity and quality of gauge observations are limited across much of the globe, particularly in sparsely populated and poorer regions. Local gauge data may in fact be of lower accuracy than the large scale products if the rain gauges on the ground are of poor quality or have been influenced by human error. There is no way to check which is more correct by looking at precipitation alone and an independent source of data is required. River flow data has been used for this purpose in the past (Beck et al., 2017) and presents a viable option for assessing precipitation performance in the context of flood events. To fulfil the requirements of high-quality local precipitation and river flow observations, an area of northern England, encompassing the Tyne, Tees, Eden, Wear and Lune basins (Figure 1) was selected for this study. The relatively simple flood response of these steep, surface water dominated basins, and the occurrence of recent flood events, means they provide a suitable testbed for investigating the effects of using global reanalysis products for more localised flood risk modelling.

2.2 Model Setup

The City Catchment Analysis Tool (CityCAT) (Glenis et al., 2018), a hydrodynamic surface water flood model, was used to simulate flooding in this study. CityCAT represents spatial rainfall fields falling directly onto ~~uniformly gridded~~ elevation surfaces made up of uniformly sized square grid cells and propagating according to the shallow water equations (SWEs). The model uses the method of finite volumes and shock capturing schemes to solve the SWEs with a Generalised Osher-Solomon Riemann solver. The system is suitable for this study as it is able to directly capture the effects of rainfall on flood depths without requiring any intermediate steps. The domain grid of each simulation is directly generated from the DEM and has the same resolution, in this case 50m. Model domains within the study area were delineated using HydroBASINS (Lehner and Grill, 2014). The outer boundaries of the domain were treated as being open which allows water to exit the domain at basin outlets. River channel bathymetry was not incorporated, and the DEM was used in its original form. This is a reasonable approximation as 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 and Dey et al (2019) observed that "the choice of bathymetric model becomes irrelevant at high flows for predicting hydraulic outputs". Water depths were output every hour for each grid cell within the domain and then extracted at each gauge location (Figure 1). The Manning's coefficient for all domains was uniformly defined as 0.03 (Chow, 1959) and the land surface was assumed to be impermeable given the extreme nature of the selected events (described below). Once the ground is saturated during long-duration flood events, subsurface processes will cease to have a large impact on water levels, especially for catchments such as those addressed here with generally shallow soils and low base flow indices. For example, Hossain Anni (2020) found an

increase of only 0.02m in average flood depth when excluding infiltration from a 100-year flood model. Additionally, it is assumed that there are no artificial water abstraction measures or flood defences present. As this study primarily aims to

100   compare model results with one another, the absolute accuracy of the hydrodynamic model is not the focus and therefore the friction, bathymetry and infiltration configurations are not critical.

2.3 Rainfall

Four global reanalysis products (JRA-55, MERRA-2, ERA-5 and CFSR) have been selected and compared against CEH-GEAR1hr, used here as a benchmark. Each rainfall dataset is described below and key characteristics are shown in Table 1.

105   The reanalysis products were selected based on their high spatiotemporal resolution, open availability and suitable duration. Events between the start and end dates of CEH-GEAR1hr (1990-2014) were selected based on the peak stage at the most downstream river gauge within each basin (Table 2). This identified the most extreme rainfall events, independently of the rainfall data itself. The largest events were chosen as they have the greatest impact in terms of flood damages. Looking at a wider range of events may have provided a more comprehensive view of the performance of reanalysis products across

110   different magnitudes, however this was outside the scope of the study. Different events were selected for each basin as the largest extremes may have occurred at different times in different areas. Each identified event was only simulated in the basin in which it was observed, to enable river gauge records to be used for validation. Simulations were commenced two weeks before the discharge peaks and ran until one week after. This was to allow model spinup and characterisation of hydrograph recession. The sensitivity to run duration was not explicitly assessed here but the duration was sufficient in all

115   cases to ensure adequate accounting for antecedent rainfall and return to normal flow conditions. Antecedent rainfall is necessary to initiate normal flow in river channels. If no spin-up period is included, then flood magnitudes are underestimated.

The events, according to each dataset, are mapped in Figure 3Figure 2 and compared with time series of observations at

120   selected gauges in Figure 2. CEH-GEAR1hr contained, on average, higher rainfall totals than the reanalysis products and JRA-55 represented only approximately half as much precipitation as other reanalysis products. ERA-5 significantly over-estimated the gauged precipitation for the 2005 event in the Tyne basin but other than this, reanalysis products underestimated rainfall totals. The 1995 event in the Lune basin was the most under-represented across the reanalysis products. CEH-GEAR1hr was consistently very close to the gauge observations as they will have been used as part of its

125   creation.

Each rainfall value was converted into a rate (kg m-2 s-1) at the observed corresponding times and each point on the original reanalysis grid was converted into an area with a width and height equivalent to the horizontal and vertical resolution of the dataset. This resulted in differently sized rainfall polygons for each dataset, corresponding to the resolutions listed in Table

130 1. These areas were then re-projected into British National Grid as cartesian coordinates are required by CityCAT. The rainfall products are described below-, along with findings from previous studies which have assessed their performance.

The Centre for Ecology and Hydrology provide an hourly version of their Gridded Estimates of Areal Rainfall dataset (CEH-GEAR1hr) (Lewis et al., 2018). This hourly product is based on a daily product which interpolates data from rain gauges

135 using natural neighbour interpolation (Tanguy et al., 2019). CEH-GEAR1hr uses nearest neighbour interpolation to maintain more realistic weather patterns and unmoderated peak values. To ensure consistency between the hourly and daily versions, the daily totals were maintained in the hourly dataset by scaling the interpolated values accordingly (Lewis et al., 2018). Quality control procedures were applied to the hourly gauge data used to produce the gridded product. Each gauge was compared with CEH-GEAR daily and 92.9% matched well. Other flags were applied to suspicious values, such as those

140 which exceeded the 1- or 24-hour record values, values which were preceded by 23 hours of no rain and for tipping bucket gauges where the frequency of tips was unexpectedly high. Combinations of these flags were used to identify and exclude values where necessary. –This gauge-based dataset was used as a baseline to compare against the reanalysis products identified below. The approach of using a gauge-based product as a baseline is well established (Jiang et al., 2021; Sun and Barros, 2010; Lei et al., 2021). However, there are always errors present in any rainfall product and gauge-based products

145 are no exception. For example, wind-induced under-catch may lead to a negative bias in precipitation (Pollock et al., 2018). The network may also not be dense enough to capture given events effectively. Despite these limitations, CEH-GEAR1hr provides the best available gridded hourly data for the UK, based on quality-controlled data from a well-established network.

Japanese Meteorological Agency reanalysis 55 (JRA-55) replaces JRA-25, incorporating higher resolution and better data

150 assimilation, among other improvements (Kobayashi et al., 2015; Japan Meteorological Agency, Japan, 2013). Suzuki et al. (2018) were able to effectively simulate continental river discharge using JRA-55, however, they found large biases attributable to precipitation error in some regions. Chen et al (2014) found that JRA-55 has a diurnal difference of ~50% which was comparable to an observed satellite dataset from the Tropical Rainfall Monitoring Mission (TRMM). Hua et al (2019) found that JRA-55 over-estimated rainfall by ~15% around southern Sahel and western equatorial Africa. Over

155 eastern China, Chen et al (2019) found that rainfall was better represented by JRA-55 than ERA-Interim, CFSR and MERRA-1. Meanwhile in the Northern Great Plains, JRA-55 has been found to perform worse than other reanalysis products such as ERA-5 and MERRA-2, demonstrating a strong wet bias (Xu et al., 2019). The variability in performance across these studies illustrates that the accuracy of JRA-55 is not consistent between regions.

160 Modern-Era Retrospective Analysis for Research and Applications 2 (MERRA-2) (Global Modeling and Assimilation Office, 2015) builds upon its predecessor, MERRA (Rienecker et al., 2011) with reduced biases in aspects of the water cycle, among other improvements (Gelaro et al., 2017). MERRA-2 uses observed precipitation products to correct the forecasts and provide better estimates (Reichle et al., 2017). Hua et al. (2019) found that MERRA-2 was better at representing rainfall

climatology over Central Equatorial Africa than ERA-Interim and JRA-55, among others, with a mean bias of only 0.01 mm/day. Hamal et al. (2020) found that MERRA-2 was able to accurately capture the seasonal precipitation in Nepal when compared to gauge observations (R > 0.95). Over Pakistan, MERRA-2 precipitation has been shown to have an RMSE of 1.68 mm and performed better than JRA-55 (2.2 mm) but not as well as ERA-5 (1.53 mm) (Arshad et al., 2021). Liu et al. (2021) found that MERRA-2 precipitation was more similar to satellite observations during summer in the Sichuan Basin, with a linear correlation coefficient of 0.9 compared to other parts of the year with 0.57. In India, MERRA-2 has negative bias of 10% in extreme rainfall, compared with 33% from ERA-Interim (Mahto and Mishra, 2019).

The European Centre for Medium-Range Weather Forecasts Reanalysis 5 product (ERA-5) (Hersbach et al., 2020) replaces and improves on ERA-Interim (Dee et al., 2011), which stopped being produced in August 2019. It supports an increased spatial and temporal resolution, along with an updated modelling and data assimilation system, which has resulted in better representation of convective rainfall (3.8% vs -5% median bias in monsoon precipitation over India) (Mahto and Mishra, 2019). The land surface component is being used to calculate river discharge for the Global Flood Awareness System (Harrigan et al., 2020). Albergel et al. (2018) found that ERA-5 resulted in better estimates of river discharge than ERA-Interim when used to drive a land surface model of the US. It has also been shown to outperform a range of other reanalysis products as part of a hydrological model applied in two Indian basins, with an RMSE of 25.5% compared with 40.2%, 59.2% and 75.6% from CFSR, JRA-55 and MERRA-2 respectively (Mahto and Mishra, 2019). In Pakistan, ERA-5 has an RMSE of 1.53 mm compared with daily gauge data, again outperforming MERRA-2 and JRA-55.

The NCEP Climate Forecast System Reanalysis (CFSR) (Saha et al., 2010a) replaces the previous NCEP/NCAR reanalysis (Kalnay et al., 1996) and uses a very similar analysis system to MERRA-2 (Saha et al., 2010b). Zhu et al. (2016) demonstrated that CFSR was liable to overestimate high streamflow in two Chinese basins using SWAT and highlighted that performance varied between basins (19.15% - 31.47% bias). Nkiaka et al. (2017) found that using CFSR over ERA-Interim resulted in substantially improved representation of river flow in the Sudano-Sahel Region, with maximum Nash Sutcliffe Efficiencies of 0.43 and -0.56 respectively. In the Amazon basin, CFSR has been shown to underpredict winter precipitation with a bias of -0.60 and overpredict summer precipitation with a bias of 0.11 (Blacutt et al., 2015). During 2010-2014 in Bangladesh, CFSR overestimated precipitation relative to gauge observations with a bias of 1.18, this was greater than ERA-5 which had a bias of 0.80 (Islam and Cartwright, 2020). In the Johor River Basin, Malaysia, the daily RMSE of CFSR was found to be 17.70 mm when compared with gauge observations (Tan et al., 2017).

2.4 Digital Elevation Model

The terrain dataset used to represent the domain surface is a nationally and freely available Digital Elevation Model (DEM) product from the Ordnance Survey, known as OS Terrain 50 (Ordnance Survey, 2017). This has been shown to perform best for flood risk modelling in a comparison with other DEMs (McClean et al., 2020). The product is based on airborne LIDARa

combination of photogrammetry and topographical surveys and is corrected using a combination of automated and manual processes to create a bare earth surface with raised structures removed. Other, higher resolution DEMs are available, such as Environment Agency LIDAR. However, complete LIDAR coverage of the study area was not available. The DEM is therefore a source of uncertainty and is likely to cause over-estimation of inundation. For example, Yunus et al. (2016) found that using OS Terrain 50 resulted in 3-10% more inundation for London when compared to 1m LIDAR. The ~~dataset~~ DEM was clipped to the area of each basin and used directly within the models (Figure 1).

2.5 Validation Data

15-minute stage observations were obtained from the Environment Agency (EA) at the most downstream gauge in each basin via a Freedom of Information request. The most downstream gauge was used in each case as these are influenced by the largest areas of rainfall. The IDs, catchment areas and locations of each gauge are listed in Table 2. Observations of flood extent during the event in Carlisle were extracted from the EA Recorded Flood Outlines dataset (Environment Agency, 2019a). Distributed measurements of maximum water depth from the same event were provided by Neal et al (2009).

2.6 Exposure

Building outlines from OS VectorMapLocal (VML) (Ordnance Survey, 2018) were used to estimate numbers of buildings inundated by each model. VML only represents individual buildings with a floor area over 20 m2 and each polygon may represent multiple buildings. Therefore, not all buildings are included in the inundation totals. This is acceptable for this analysis which compares the relative magnitude of flooding, rather than the absolute totals. Buildings from VML were classified as flooded if they intersected any model cell above a typical property threshold of 0.3 m (Environment Agency, 2019b).

3 Results

The performance of each simulation was compared in terms of the magnitude and timing of the hydrograph peak, the flood depth and extent, and the number of buildings inundated (Table 3). ERA-5 outperformed other reanalysis datasets in terms of hydrograph peak error and floodplain depth, however, CFSR produced more similar inundation levels to CEH-GEAR1hr and demonstrated more accurate peak timing, flood extent and depth compared to point observations. JRA-55 performed significantly worse than other reanalysis products across all ~~of these aggregated~~ measures. The variability of each metric will now be assessed in more detail, including spatial variations in performance.

The maximum water depths according to models using each of the rainfall datasets are shown in Figure 4~~Figure 3~~. Overall, the spatial distribution of floodwater is similar, as the same DEM is used in all models. There are noticeably higher depths along main river channels in the CEH-GEAR1hr results. JRA-55 presents less clearly visible channels than the other models,

particularly in the Lune and Tees basins. The maps also illustrate that the MERRA-2 model produced lower depths in the Tyne basin than other reanalysis precipitation datasets. Across all basins, ERA-5 and CFSR produced similar distributions of error relative to the CEH-GEAR1hr results. The inter-quartile range of errors in MERRA-2 is narrower but the median error is slightly further below zero than ERA-5 and CFSR. JRA-55 water depths were significantly further below the other reanalysis datasets.

The flood extents from each model during the 2005 event in the Eden basin are compared against Environment Agency recorded flood outlines in Figure 5. This event was chosen because it provided the largest available flood extents in the observed dataset which coincide with a built-up area. CEH-GEAR1hr resulted in the highest Critical Success Index (CSI) (0.54). CFSR and ERA5 performed similarly to each other and less well than CEH-GEAR1hr. MERRA-2 caused further underestimation of extent, while JRA-55 had the lowest CSI by a significant margin. The recorded outlines data does not contain information about the total area that was surveyed so regions incorrectly identified as flooded in the model outputs may still have been flooded in reality. This is clearly the case in the downstream section of the Eden in the upper left of each plot, along with other water courses visible in the model outputs. This means that the CSI values are under-representative of accuracy, however they provide a useful metric for comparison between datasets.

Wrack and water marks recorded following the 2005 event in Carlisle (Neal et al., 2009) have been compared against maximum modelled water depths in Figure 6. Wrack marks are left by debris deposited at the flood edge (HR Wallingford, 2004), while water marks are left as stains on the side of structures within the flooded area. The DEM values from the computational grid were subtracted from the observed water elevations to produce flood depths for comparison with the model outputs. Any depths which were calculated as being below zero were assumed to be zero. Again, CEH-GEAR1hr is the closest to the observed data with an RMSE of 0.41m, followed by CFSR and ERA-5 with approximately twice the error. JRA-55 resulted in less than half the r^2 value of other datasets. The ranking of datasets remained consistent between the CSI analysis and comparison against observed depths.

Time series of water depths were extracted from the models at each river gauge location and compared with the observed values (Figure 7Figure 4). In all basins, apart from the Wear, CEH-GEAR1hr was closest to the observed peaks and predicted the highest maximum depth. In the Wear basin, where all models overestimated river stage, CEH-GEAR1hr was actually the least accurate. However, the observed values may be misleading here as flows go out of bank above 3 metres and so peaks are truncated in the observed series. This means that the actual stage is likely to be higher than the recorded values and therefore the magnitude of the reported errors overestimated. JRA-55 consistently severely underestimated river stage and only captured peaks in the Eden and Tyne basins. ERA-5 and CFSR display relatively similar performance across all basins. Meanwhile, MERRA-2 underestimated the peaks in the Eden and Tyne. All reanalysis products strongly underestimated the flood peak in the Lune basin.

The total numbers of inundated buildings for each model are shown in Figure 8Figure 5. As there is no observed building inundation data available, it cannot be concluded which is the most accurate. However, a reasonable assumption might be that since CEH-GEAR1hr is based on rainfall observations and is at a higher resolution, it is likely to produce the closest estimate to the truth. In four out of five basins, using CEH-GEAR1hr resulted in the highest number of inundated buildings. ERA-5 inundated the most buildings in the Tees basin despite not being consistently higher than the other reanalysis datasets in the other basins. JRA-55 inundated the lowest number of buildings by a large margin in all basins apart from the Tyne, where it exceeded both MERRA-2 and CFSR. CFSR never resulted in either the highest or lowest number of inundated buildings. There is general agreement between the rankings of modelled peak water depth as shown in Figure 7Figure 4 and the number of inundated buildings. Notable exceptions include the Tees, where the high inundation levels predicted by ERA-5 were not replicated in its depth peak, which was lower than CEH-GEAR1hr by a clear margin. Changing the inundation threshold had only very minor effects on the relative differences in inundation between datasets.

4 Discussion

The results presented above demonstrate a persistent bias towards underestimation of flood depths and impacts when using global reanalysis products in place of high-resolution gauge-based rainfall datasets. 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. The negative bias has been shown to exist in depths across the basins studied, at river gauging stations and specifically at the locations of buildings, which correspond to built-up areas exposed to flooding. This finding is in line with Sampson et al. (2014), who show ERA-Interim, an older product, underestimated flood risk. Our results, however, do not indicate such a stark bias, perhaps because the products used here are more modern and advanced than ERA-Interim. This is backed up by Towner et al. (2019) who have demonstrated improved performance of ERA-5 over ERA-Interim using hydrological models of the Amazon. Hirpa et al. (2016) also illustrate that ERA-Interim can underestimate flood risk, with spatial variability, which further reinforces our finding. In contrast, Andreadis et al. (2017) find flood extent to be overestimated (relative to a benchmark simulation) when using the 20CRv2 reanalysis product. However, they did find that outflow discharge was underestimated, which agrees with our results. Their assessment of flood extent did not include flood depths or effects on the inundation of exposed assets, as we have done here, which may explain the observed overestimation to some degree. We also did not replicate the underestimation of streamflow found by Zhu et al. (2016) when using CFSR. Though, it is difficult to draw direct comparisons given the major differences in methodology between studies.

We found that no precipitation product performed better in all models and each product performed differently depending on the basin. This implies that the optimum dataset to use depends on the location of the model. JRA-55 was very poor at capturing extreme rainfall and subsequently hydrograph peak and inundation magnitude in almost all cases. This may be slightly influenced by the lower temporal resolution, but it is unlikely that the small difference in observation frequency would result in such a strong negative effect on model performance. ERA-5 consistently performed better than other reanalysis datasets in terms of capturing the observed hydrograph peak, apart from in the Eden, where CFSR was more accurate and also demonstrated a higher CSI and lower RMSE relative to recorded outlines and wrack mark depths. ERA-5, CFSR and MERRA2 were more evenly matched in terms of floodplain water depth (Figure 4), inundation extent, and impacts (Figure 8). We find no cause to favour any of these three datasets and suggest that all three could be adopted in parallel by reanalysis-based flood models to capture the range of uncertainty.

Links between hydrograph performance and estimated numbers of inundated buildings are present but the relationship is not consistent. For example, in the Tyne basin, CFSR estimates a higher gauge peak than JRA-55 but, at the same time, inundates fewer buildings. Meanwhile, MERRA-2 only has the lowest hydrograph peak in the Tyne, where it estimates the lowest total building inundation compared to other models. CEH-GEAR1hr is also both generally higher in terms of both building inundation and hydrograph peak, but the occasions where this is not the case do not correspond to the same basin. These findings demonstrate that there is generally a positive relationship between peak hydrograph depth and numbers of inundated buildings, but increased river depth does not always lead to greater inundation. Therefore, hydrograph performance is not an entirely reliable metric for assessing the accuracy of flood risk estimated using global reanalysis products.

The underestimation of inundation magnitude caused by using global precipitation data is counter to the overestimation that results from using global DEM data, as demonstrated by McClean et al. (2020). The negative inundation bias caused by using reanalysis precipitation is, however, not as strong as the positive bias from global DEM products. This is because changes in rainfall input have a less significant impact on the spatial distribution of flooding than changes in DEM input. Therefore, it is anticipated that the combined effects of using both global DEM and global reanalysis precipitation would not cancel themselves out and are likely to produce a net positive bias.

Undoubtedly the effects shown here are specific to the study area and other locations may present different patterns. Each reanalysis product may behave differently across climatic regions, for example. Areas with highly constrained topography are unlikely to be strongly affected by the choice of precipitation data, in terms of flood extent and numbers of inundated assets. This is because increases in total rainfall volume will not greatly alter flood extent if there are no new available flow pathways. A key limitation to applying our methodology in new locations is the requirement for high-quality gauge-based

precipitation datasets and river ~~flow~~ stage observations to compare against. Despite the caveat of locality, our results do demonstrate the potential for underestimation of flood risk when reanalysis products are involved. This underestimation has been replicated by previous studies in other areas (Sampson et al., 2014) and users of models based on reanalysis data should
330 be aware of this effect.

5 Conclusions

Using precipitation from global reanalysis datasets results in an underestimation of flood risk by 14-18 % of inundated buildings (Table 3, excluding JRA-55 as it was far outside the range of other products) ~~and 29-44 % of median floodplain depth (Table 3, excluding JRA-55)~~ relative to CEH-GEAR1hr, which outperformed reanalysis products in terms of flood
335 depth and extent when compared to observations. The effect is location-specific, though, and this study found that no product performed best across all five of the catchments we studied. In some areas, the reanalysis data did result in similar levels of inundation to the national observed precipitation product ~~(CEH-GEAR1hr)~~. This is a positive message for the use of reanalysis data in flood risk modelling generally and future progress in forecast models will undoubtedly reduce this gap even further.

340

As climatic and land-use changes increase flood hazard, the importance of accurately understanding current and future flood risk is increasing. Reanalysis data has enabled flood risk assessments to be undertaken more widely. However, this analysis shows global or regional reanalysis data should not be considered as a replacement for local, high resolution, observations. Uncertainties in flood risk assessment using reanalysis data need to be properly quantified and communicated
345 to insurers, local and national authorities and communities, to ensure flood risk management decisions are not misinformed.

While reanalysis datasets do show promising and improving results (ERA-5 achieved a mean absolute hydrograph peak error of 29.4 %, equivalent to CEH-GEAR1hr and CFSR only inundated on average 14.4 % fewer buildings than CEH-GEAR1hr), caution should be used when interpreting outputs from any models based on them due to the underestimation of
350 inundated buildings demonstrated here. However, as no observed building inundation data were available, our findings are not definitive, and the reanalysis data may have produced more accurate inundation estimates than the national gauge-based product. We suggest that multiple products, such as ERA-5, CFSR and MERRA-2, should be used where possible to capture the full range of uncertainty. This is because each of these products has been shown to perform better in different areas or when using different performance measures. Based on the comparatively strong negative bias in inundation and flood peak
355 shown here for a limited set of events, JRA-55 ~~should not be used in flood risk modelling~~ may result in substantially lower risk estimates than other reanalysis products and users of model outputs based on it should take this into account. However, as highlighted, certain products may perform better in other areas and further research is needed to assess new and existing reanalysis products for flood modelling across a wider range of climatic regions. To enable this, a more systematic

international strategy for the collection of rainfall data is needed to ensure more complete global coverage of validation data,
360 building on efforts from Lewis et al. (2019). New reanalysis products continue to be developed which may improve on the findings presented here (Muñoz-Sabater et al., 2021). This will require ongoing validation efforts to identify possible advancements in terms of flood risk analysis.

Code/Data availability

The reanalysis products can be downloaded using the DOIs in Table 1. All model results and code used to generate figures
365 will be made available by final submission.

Competing interests

The authors declare that they have no conflict of interest.

Author contribution

FM processed the data and executed the simulations. FM prepared the manuscript with contributions from all co-authors.

370 ## Acknowledgements

The authors were supported by the Natural Environmental Research Council (NERC) Centre for Doctoral Training for Data, Risk and Environmental Analytical Methods (DREAM), Engineering & Physical Science Research Council TWENTY 65: Tailored Water Solutions for Positive Impact (EP/N010124/1), UK Research & Innovation Global Challenges Research Fund Water Security and Sustainable Development Hub (ES/S008179/1), and Willis Research Network. The City Catchment
375 Analysis Tool (CityCAT) was developed and provided by Vassilis Glenis. Chris Kilsby acknowledges the support of the Willis Research Network. The water depth observations were provided by Dr. Jeffrey Neal.

References

Albergel, C., Dutra, E., Munier, S., Calvet, J.-C., Munoz-Sabater, J., de Rosnay, P., and Balsamo, G.: ERA-5 and ERA-Interim driven ISBA land surface model simulations: which one performs better?, Hydrol. Earth Syst. Sci., 22, 3515–3532,
380 https://doi.org/10.5194/hess-22-3515-2018, 2018.
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, 1161–1175, https://doi.org/10.5194/hess-17-1161-2013, 2013.

Andreadis, K. M., Schumann, G. J.-P., Stampoulis, D., Bates, P. D., Brakenridge, G. R., and Kettner, A. J.: Can Atmospheric
Reanalysis Data Sets Be Used to Reproduce Flooding Over Large Scales?, Geophys. Res. Lett., 44, 10,369-10,377,
https://doi.org/10.1002/2017GL075502, 2017.

Arshad, M., Ma, X., Yin, J., Ullah, W., Liu, M., and Ullah, I.: Performance evaluation of ERA-5, JRA-55, MERRA-2, and
CFS-2 reanalysis datasets, over diverse climate regions of Pakistan, Weather Clim. Extrem., 33, 100373,
https://doi.org/https://doi.org/10.1016/j.wace.2021.100373, 2021.

Beck, H. E., Vergopolan, N., Pan, M., Levizzani, V., Van Dijk, A. I. J. M., Weedon, G. P., Brocca, L., Pappenberger, F.,
Huffman, G. J., and Wood, E. F.: Global-scale evaluation of 22 precipitation datasets using gauge observations and
hydrological modeling, Hydrol. Earth Syst. Sci., https://doi.org/10.5194/hess-21-6201-2017, 2017.

Blacutt, L. A., Herdies, D. L., de Gonçalves, L. G. G., Vila, D. A., and Andrade, M.: Precipitation comparison for the CFSR,
MERRA, TRMM3B42 and Combined Scheme datasets in Bolivia, Atmos. Res., 163, 117–131,
https://doi.org/10.1016/j.atmosres.2015.02.002, 2015.

Chawla, I. and Mujumdar, P. P.: Evaluating rainfall datasets to reconstruct floods in data-sparse Himalayan region, J.
Hydrol., 588, 125090, https://doi.org/10.1016/j.jhydrol.2020.125090, 2020.

Chen, G., Iwasaki, T., Qin, H., and Sha, W.: Evaluation of the Warm-Season Diurnal Variability over East Asia in Recent
Reanalyses JRA-55, ERA-Interim, NCEP CFSR, and NASA MERRA, J. Clim., 27, 5517–5537,
https://doi.org/10.1175/JCLI-D-14-00005.1, 2014.

Chen, S., Gan, T. Y., Tan, X., Shao, D., and Zhu, J.: Assessment of CFSR, ERA-Interim, JRA-55, MERRA-2, NCEP-2
reanalysis data for drought analysis over China, Clim. Dyn., 53, 737–757, https://doi.org/10.1007/s00382-018-04611-1,
2019.

Chow, V. Te: Open-channel hydraulics, McGraw-Hill New York, 1959.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo,
G., and Bauer, P.: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, Q. J. R.
Meteorol. Soc., 137, 553–597, 2011.

Dey, S., Saksena, S., and Merwade, V.: Assessing the effect of different bathymetric models on hydraulic simulation of
rivers in data sparse regions, J. Hydrol., 575, 838–851, https://doi.org/10.1016/j.jhydrol.2019.05.085, 2019.

Environment Agency: Recorded Flood Outlines, https://data.gov.uk/dataset/16e32c53-35a6-4d54-a111-
ca09031eaaaf/recorded-flood-outlines, 2019a.

Environment Agency: What is the risk of flooding from surface water map, Report Version 2.0 pp., 2019b.

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M.
G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., da Silva, A. M., Gu, W.,
Kim, G. K., Koster, R., Lucchesi, R., Merkova, D., Nielsen, J. E., Partyka, G., Pawson, S., Putman, W., Rienecker, M.,
Schubert, S. D., Sienkiewicz, M., and Zhao, B.: The modern-era retrospective analysis for research and applications, version
2 (MERRA-2), J. Clim., 30, 5419–5454, https://doi.org/10.1175/JCLI-D-16-0758.1, 2017.

Glenis, V., Kutija, V., and Kilsby, C. G.: A fully hydrodynamic urban flood modelling system representing buildings, green space and interventions, Environ. Model. Softw., 109, 272–292, https://doi.org/10.1016/j.envsoft.2018.07.018, 2018.

420 Global Modeling and Assimilation Office: MERRA-2 tavg1_2d_flx_Nx: 2d,1-Hourly,Time-Averaged,Single-Level,Assimilation,Surface Flux Diagnostics V5.12.4, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), https://doi.org/10.5067/7MCPBJ41Y0K6, 2015.

Hamal, K., Sharma, S., Khadka, N., Baniya, B., Ali, M., Shrestha, M. S., Xu, T., Shrestha, D., and Dawadi, B.: Evaluation of MERRA-2 Precipitation Products Using Gauge Observation in Nepal, https://doi.org/10.3390/hydrology7030040, 2020.

425 Harrigan, S., Zsoter, E., Alfieri, L., Prudhomme, C., Salamon, P., Wetterhall, F., Barnard, C., Cloke, H., and Pappenberger, F.: GloFAS-ERA5 operational global river discharge reanalysis 1979- present, 2020.

He, S., Yang, J., Bao, Q., Wang, L., and Wang, B.: Fidelity of the observational/reanalysis datasets and global climate models in representation of extreme precipitation in East China, J. Clim., 32, 195–212, https://doi.org/10.1175/JCLI-D-18-0104.1, 2019.

430 Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Q. J. R. Meteorol.
435 Soc., 146, 1999–2049, https://doi.org/https://doi.org/10.1002/qj.3803, 2020.

Hirpa, F. A., Salamon, P., Alfieri, L., Thielen-del Pozo, J., Zsoter, E., and Pappenberger, F.: The effect of reference climatology on global flood forecasting, J. Hydrometeorol., 17, 1131–1145, https://doi.org/10.1175/JHM-D-15-0044.1, 2016.

Hossain Anni, A., Cohen, S., and Praskievicz, S.: Sensitivity of urban flood simulations to stormwater infrastructure and soil
440 infiltration, J. Hydrol., 588, https://doi.org/10.1016/j.jhydrol.2020.125028, 2020.

HR Wallingford: Flooding in Boscastle and North Cornwall, August 2004, 104 pp., 2004.

Hua, W., Zhou, L., Nicholson, S. E., Chen, H., and Qin, M.: Assessing reanalysis data for understanding rainfall climatology and variability over Central Equatorial Africa, Clim. Dyn., 53, 651–669, https://doi.org/10.1007/s00382-018-04604-0, 2019.

Islam, M. A. and Cartwright, N.: Evaluation of climate reanalysis and space-borne precipitation products over Bangladesh,
445 Hydrol. Sci. J., 65, 1112–1128, https://doi.org/10.1080/02626667.2020.1730845, 2020.

Japan Meteorological Agency, Japan: JRA-55: Japanese 55-year Reanalysis, Daily 3-Hourly and 6-Hourly Data, https://doi.org/10.5065/D6HH6H41, 2013.

Jiang, Q., Li, W., Fan, Z., He, X., Sun, W., Chen, S., Wen, J., Gao, J., and Wang, J.: Evaluation of the ERA5 reanalysis precipitation dataset over Chinese Mainland, J. Hydrol., 595, 125660, https://doi.org/10.1016/j.jhydrol.2020.125660, 2021.

450 Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Leetmaa, A.,

Reynolds, R., Jenne, R., and Joseph, D.: The NCEP/NCAR 40-Year Reanalysis Project, Bull. Am. Meteorol. Soc., 77, 437–472, https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2, 1996.

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., Onogi, K., Kamahori, H., Kobayashi, C., Endo, H., Miyaoka, K., and Takahashi, K.: The JRA-55 Reanalysis: General Specifications and Basic Characteristics, J. Meteorol. Soc. Japan. Ser. II, 93, 5–48, https://doi.org/10.2151/jmsj.2015-001, 2015.

de Leeuw, J., Methven, J., and Blackburn, M.: Evaluation of ERA-Interim reanalysis precipitation products using England and Wales observations, Q. J. R. Meteorol. Soc., 141, 798–806, https://doi.org/10.1002/qj.2395, 2015.

Lehner, B. and Grill, G.: HydroBASINS: Global watershed boundaries and sub-basin delineations derived from HydroSHEDS data at 15 second resolution—Technical documentation version 1. c, 1–9 pp., 2014.

Lei, H., Li, H., Zhao, H., Ao, T., and Li, X.: Comprehensive evaluation of satellite and reanalysis precipitation products over the eastern Tibetan plateau characterized by a high diversity of topographies, Atmos. Res., 259, 105661, https://doi.org/10.1016/j.atmosres.2021.105661, 2021.

Lewis, E., Quinn, N., Blenkinsop, S., Fowler, H. J., Freer, J., Tanguy, M., Hitt, O., Coxon, G., Bates, P., and Woods, R.: A rule based quality control method for hourly rainfall data and a 1 km resolution gridded hourly rainfall dataset for Great Britain: CEH-GEAR1hr, J. Hydrol., 564, 930–943, https://doi.org/10.1016/j.jhydrol.2018.07.034, 2018.

Lewis, E., Fowler, H., Alexander, L., Dunn, R., McClean, F., Barbero, R., Guerreiro, S., Li, X.-F., and Blenkinsop, S.: GSDR: A Global Sub-Daily Rainfall Dataset, J. Clim., 32, 4715–4729, https://doi.org/10.1175/JCLI-D-18-0143.1, 2019.

Liu, H., Yan, R., and Yang, J.: Credibility and statistical characteristics of CAMSRA and MERRA-2 AOD reanalysis products over the Sichuan Basin during 2003–2018, Atmos. Environ., 244, 117980, https://doi.org/https://doi.org/10.1016/j.atmosenv.2020.117980, 2021.

Mahto, S. S. and Mishra, V.: Does ERA-5 Outperform Other Reanalysis Products for Hydrologic Applications in India?, J. Geophys. Res. Atmos., 124, 9423–9441, https://doi.org/10.1029/2019JD031155, 2019.

McClean, F., Dawson, R., and Kilsby, C.: Implications of Using Global Digital Elevation Models for Flood Risk Analysis in Cities, Water Resour. Res., 56, e2020WR028241, https://doi.org/https://doi.org/10.1029/2020WR028241, 2020.

Met Office: MIDAS Open: UK hourly rainfall data, v201908, https://doi.org/10.5285/a58b9c8a724e4ec795a40a74455462b7, 2019.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D., Piles, M., Rodríguez-Fernández, N., Zsoter, E., Buontempo, C., and Thépaut, J.-N.: ERA5-Land: A state-of-the-art global reanalysis dataset for land applications, Earth Syst. Sci. Data Discuss., 1–50, https://doi.org/10.5194/ESSD-2021-82, 2021.

Neal, J., Hawker, L., Savage, J., Durand, M., Bates, P., and Sampson, C.: Estimating River Channel Bathymetry in Large Scale Flood Inundation Models, Water Resour. Res., 57, 1–22, https://doi.org/10.1029/2020wr028301, 2021.

Neal, J. C., Bates, P. D., Fewtrell, T. J., Hunter, N. M., Wilson, M. D., and Horritt, M. S.: Distributed whole city water level measurements from the Carlisle 2005 urban flood event and comparison with hydraulic model simulations, J. Hydrol., 368,

42–55, https://doi.org/10.1016/j.jhydrol.2009.01.026, 2009.

Nkiaka, E., Nawaz, N. R., and Lovett, J. C.: Evaluating Global Reanalysis Datasets as Input for Hydrological Modelling in the Sudano-Sahel Region, https://doi.org/10.3390/hydrology4010013, 2017.

Ordnance Survey: OS Terrain 50 User Guide v1.3, 2017.

490   Ordnance Survey: OS VectorMap Local Technical specification, 2018.

Pappenberger, F., Dutra, E., Wetterhall, F., and Cloke, H. L.: Deriving global flood hazard maps of fluvial floods through a physical model cascade, Hydrol. Earth Syst. Sci., 16, 4143–4156, https://doi.org/10.5194/hess-16-4143-2012, 2012.

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., and O'Connell, P. E.: Quantifying and Mitigating Wind-Induced Undercatch in Rainfall

495   Measurements, Water Resour. Res., 54, 3863–3875, https://doi.org/10.1029/2017WR022421, 2018.

Reichle, R. H., Liu, Q., Koster, R. D., Draper, C. S., Mahanama, S. P. P., and Partyka, G. S.: Land Surface Precipitation in MERRA-2, J. Clim., 30, 1643–1664, https://doi.org/10.1175/JCLI-D-16-0570.1, 2017.

Rienecker, M. M., Suarez, M. J., Gelaro, R., Todling, R., Bacmeister, J., Liu, E., Bosilovich, M. G., Schubert, S. D., Takacs, L., Kim, G.-K., Bloom, S., Chen, J., Collins, D., Conaty, A., da Silva, A., Gu, W., Joiner, J., Koster, R. D., Lucchesi, R.,

500   Molod, A., Owens, T., Pawson, S., Pegion, P., Redder, C. R., Reichle, R., Robertson, F. R., Ruddick, A. G., Sienkiewicz, M., and Woollen, J.: MERRA: NASA's Modern-Era Retrospective Analysis for Research and Applications, J. Clim., 24, 3624–3648, https://doi.org/10.1175/JCLI-D-11-00015.1, 2011.

Saha, S., Moorthi, S., Pan, H.-L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., Behringer, D., Liu, H., Stokes, D., Grumbine, R., Gayno, G., Wang, J., Hou, Y.-T., Chuang, H.-Y., Juang, H.-M. H., Sela, J., Iredell, M., Treadon,

505   R., Kleist, D., Delst, P. Van, Keyser, D., Derber, J., Ek, M., Meng, J., Wei, H., Yang, R., Lord, S., van den Dool, H., Kumar, A., Wang, W., Long, C., Chelliah, M., Xue, Y., Huang, B., Schemm, J.-K., Ebisuzaki, W., Lin, R., Xie, P., Chen, M., Zhou, S., Higgins, W., Zou, C.-Z., Liu, Q., Chen, Y., Han, Y., Cucurull, L., Reynolds, R. W., Rutledge, G., and Goldberg, M.: NCEP Climate Forecast System Reanalysis (CFSR) Selected Hourly Time-Series Products, January 1979 to December 2010, https://doi.org/10.5065/D6513W89, 2010a.

510   Saha, S., Moorthi, S., Pan, H.-L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., Behringer, D., Liu, H., Stokes, D., Grumbine, R., Gayno, G., Wang, J., Hou, Y.-T., Chuang, H., Juang, H.-M. H., Sela, J., Iredell, M., Treadon, R., Kleist, D., Delst, P. Van, Keyser, D., Derber, J., Ek, M., Meng, J., Wei, H., Yang, R., Lord, S., Dool, H. van den, Kumar, A., Wang, W., Long, C., Chelliah, M., Xue, Y., Huang, B., Schemm, J.-K., Ebisuzaki, W., Lin, R., Xie, P., Chen, M., Zhou, S., Higgins, W., Zou, C.-Z., Liu, Q., Chen, Y., Han, Y., Cucurull, L., Reynolds, R. W., Rutledge, G., and Goldberg, M.: The

515   NCEP   Climate   Forecast   System   Reanalysis,   Bull.   Am.   Meteorol.   Soc.,   91,   1015–1058, https://doi.org/10.1175/2010BAMS3001.1, 2010b.

Sampson, C. C., Fewtrell, T. J., O'Loughlin, F., Pappenberger, F., Bates, P. B., Freer, J. E., and Cloke, H. L.: The impact of uncertain precipitation data on insurance loss estimates using a flood catastrophe model, Hydrol. Earth Syst. Sci., 18, 2305–2324, https://doi.org/10.5194/hess-18-2305-2014, 2014.

520 Schumann, G. J. P., Neal, J. C., Voisin, N., Andreadis, K. M., Pappenberger, F., Phanthuwongpakdee, N., Hall, A. C., and Bates, P. D.: A first large-scale flood inundation forecasting model, Water Resour. Res., 49, 6248–6257, https://doi.org/10.1002/wrcr.20521, 2013.

Seyyedi, H., Anagnostou, E. N., Beighley, E., and McCollum, J.: Hydrologic evaluation of satellite and reanalysis precipitation datasets over a mid-latitude basin, Atmos. Res., 164–165, 37–48, 525 https://doi.org/10.1016/j.atmosres.2015.03.019, 2015.

Sun, X. and Barros, A. P.: An evaluation of the statistics of rainfall extremes in rain gauge observations, and satellite-based and reanalysis products using universal multifractals, J. Hydrometeorol., 11, 388–404, https://doi.org/10.1175/2009JHM1142.1, 2010.

Suzuki, T., Yamazaki, D., Tsujino, H., Komuro, Y., Nakano, H., and Urakawa, S.: A dataset of continental river discharge 530 based on JRA-55 for use in a global ocean circulation model, J. Oceanogr., 74, 421–429, https://doi.org/10.1007/s10872-017-0458-5, 2018.

Tan, M. L., Gassman, P. W., and Cracknell, A. P.: Assessment of Three Long-Term Gridded Climate Products for Hydro-Climatic Simulations in Tropical River Basins, https://doi.org/10.3390/w9030229, 2017.

Tanguy, M., Dixon, H., Prosdocimi, I., Morris, D. G., and Keller, V. J. D.: Gridded estimates of daily and monthly areal 535 rainfall for the United Kingdom (1890-2017) [CEH-GEAR], https://doi.org/10.5285/ee9ab43d-a4fe-4e73-afd5-cd4fc4c82556, 2019.

Towner, J., Cloke, H. L., Zsoter, E., Flamig, Z., Hoch, J. M., Bazo, J., De Perez, E. C., and Stephens, E. M.: Assessing the performance of global hydrological models for capturing peak river flows in the Amazon basin, Hydrol. Earth Syst. Sci., 23, 3057–3080, https://doi.org/10.5194/hess-23-3057-2019, 2019.

540 Winsemius, H. C., Van Beek, L. P. H., Jongman, B., Ward, P. J., and Bouwman, A.: A framework for global river flood risk assessments, Hydrol. Earth Syst. Sci, 17, 1871–1892, https://doi.org/10.5194/hess-17-1871-2013, 2013.

Xu, H., Xu, C. Y., Chen, S., and Chen, H.: Similarity and difference of global reanalysis datasets (WFD and APHRODITE) in driving lumped and distributed hydrological models in a humid region of China, J. Hydrol., 542, 343–356, https://doi.org/10.1016/j.jhydrol.2016.09.011, 2016.

545 Xu, X., Frey, S. K., Boluwade, A., Erler, A. R., Khader, O., Lapen, D. R., and Sudicky, E.: Evaluation of variability among different precipitation products in the Northern Great Plains, J. Hydrol. Reg. Stud., 24, 100608, https://doi.org/https://doi.org/10.1016/j.ejrh.2019.100608, 2019.

Yunus, A. P., Avtar, R., Kraines, S., Yamamuro, M., Lindberg, F., and Grimmond, C. S. B.: Uncertainties in tidally adjusted estimates of sea level rise flooding (bathtub model) for the greater London, Remote Sens., 8, 550 https://doi.org/10.3390/rs8050366, 2016.

Zhu, Q., Xuan, W., Liu, L., and Xu, Y.-P.: Evaluation and hydrological application of precipitation estimates derived from PERSIANN-CDR, TRMM 3B42V7, and NCEP-CFSR over humid regions in China, Hydrol. Process., 30, 3061–3083, https://doi.org/10.1002/hyp.10846, 2016.
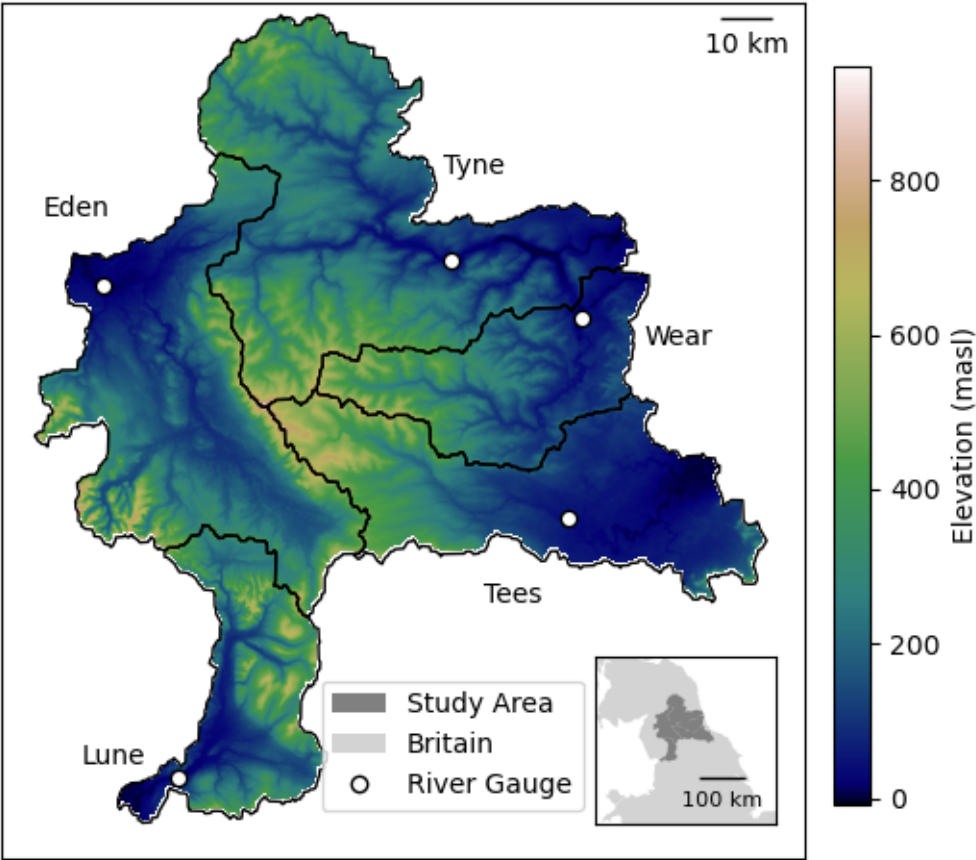
Figure 1 The location and topography of basins within the study area, illustrated using OS Terrain 50.
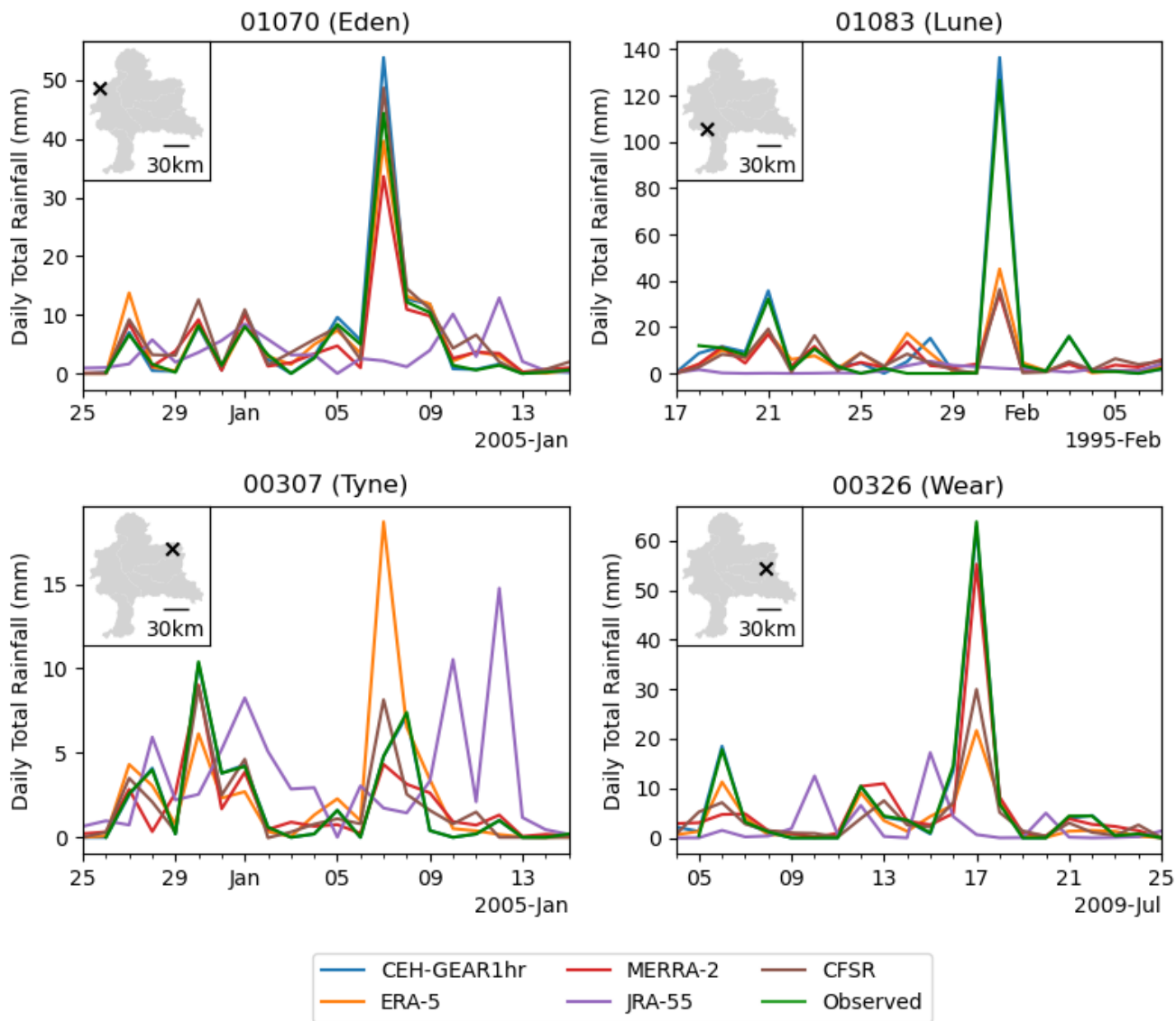
Figure 2 – Comparison of daily rainfall totals from each dataset with observed values at selected gauges. The observations were
560   **obtained from MIDAS Open** (Met Office, 2019)**. The MIDAS station ID of each gauge is shown in the title of each subplot. The series were converted from hourly to daily to improve clarity. CEH-GEAR1hr becomes obscured in places due to it precisely following the observed series.**
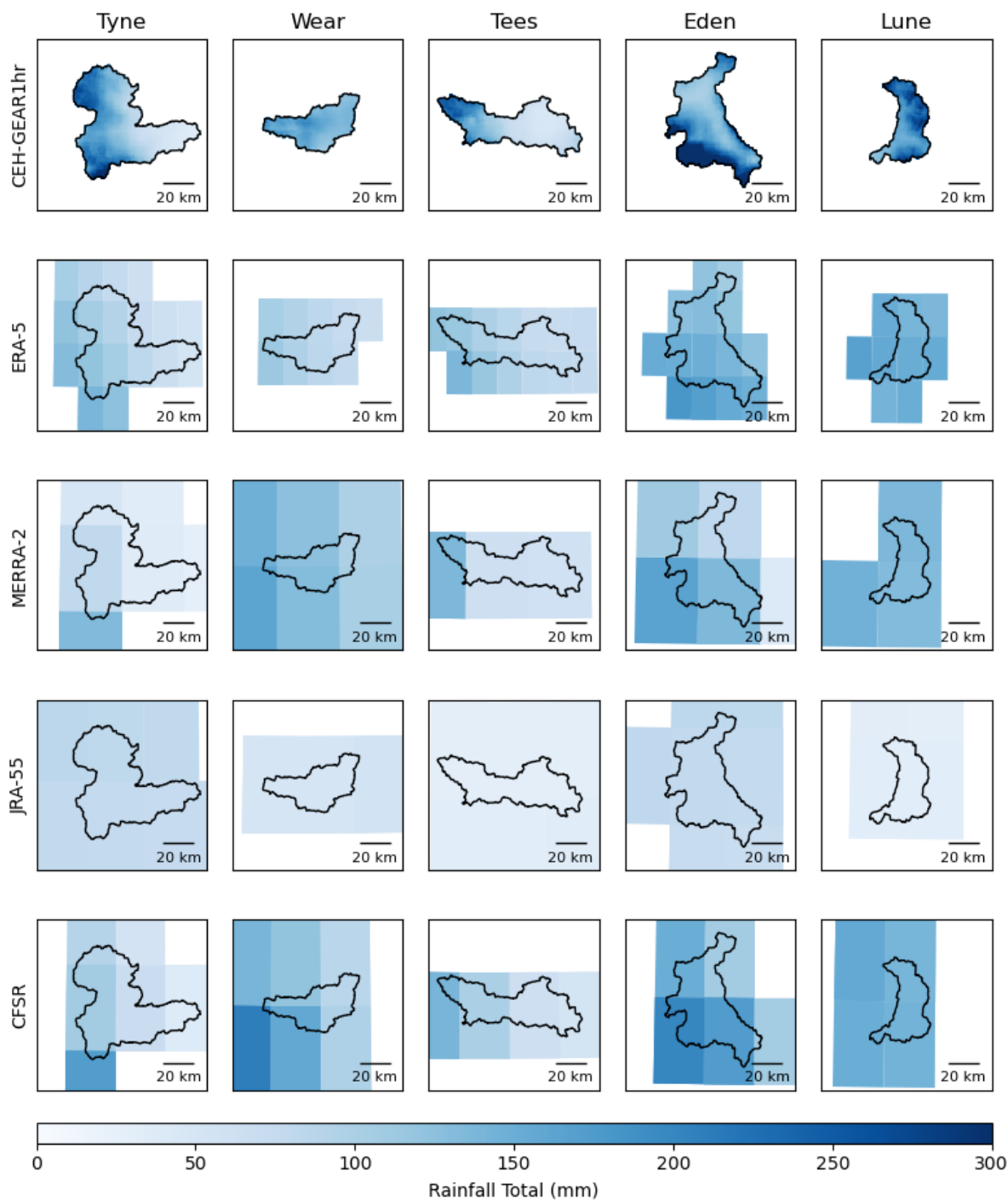
Figure 3 (A)-(E) show Rrainfall over the study area during the events in Table 2, according to each reanalysis dataset. (F) shows the total rainfall over the domain for all reanalysis datasets, The 1995 event is shown for the Tees and the Lune, the 2004 event is shown for the Tyne and the Eden, the 2009 event is shown for the Wear.
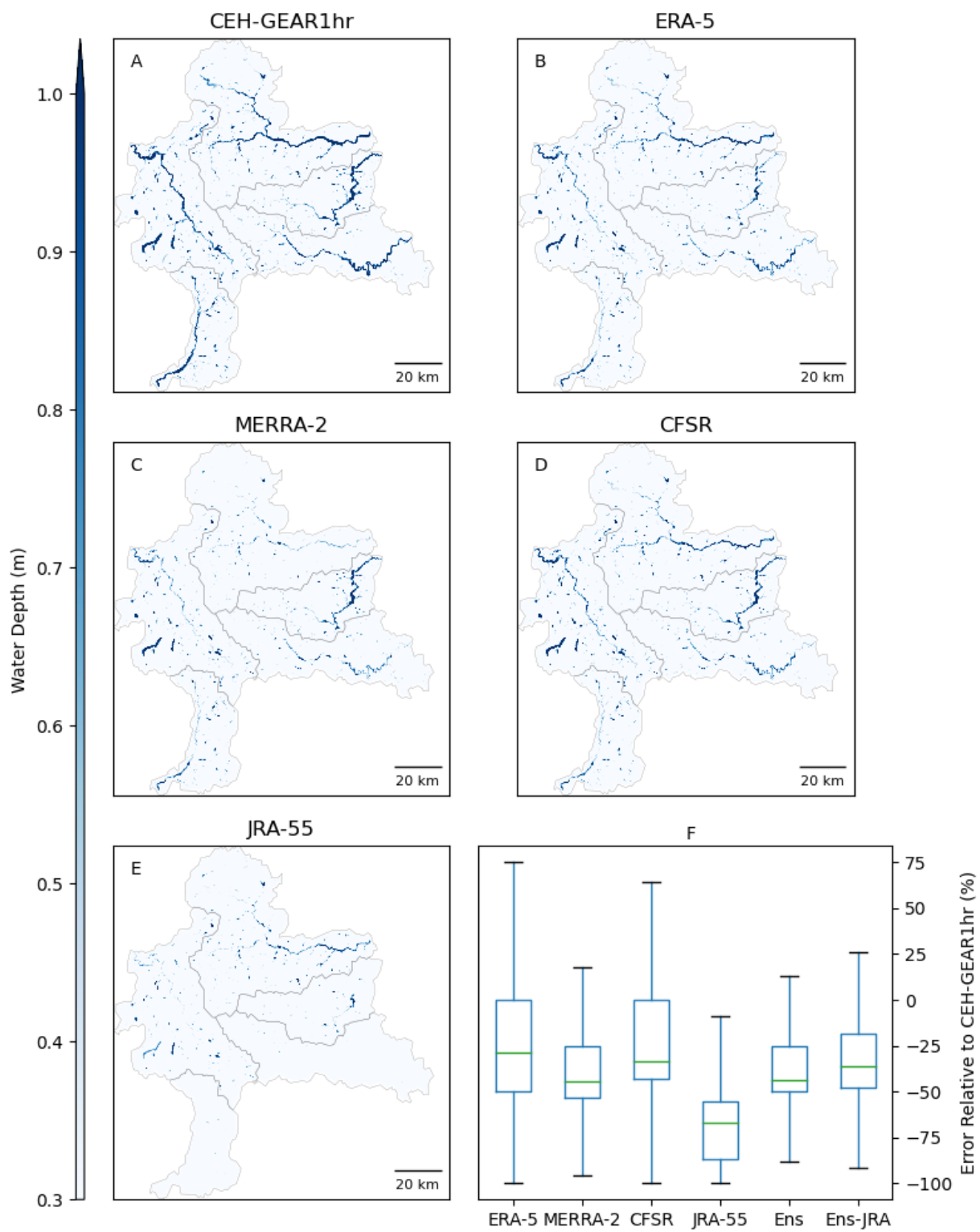
Figure 4 (A)-(E) show maximum water depth throughout the study area from models using each of the rainfall datasets. (F) shows
the depth error of the reanalysis datasets relative to CEH-GEAR-1hr across all cells, excluding outliers. "Ens" refers to the
ensemble mean of all reanalysis products. "Ens-JRA" refers to the ensemble mean of reanalysis product excluding JRA-55.
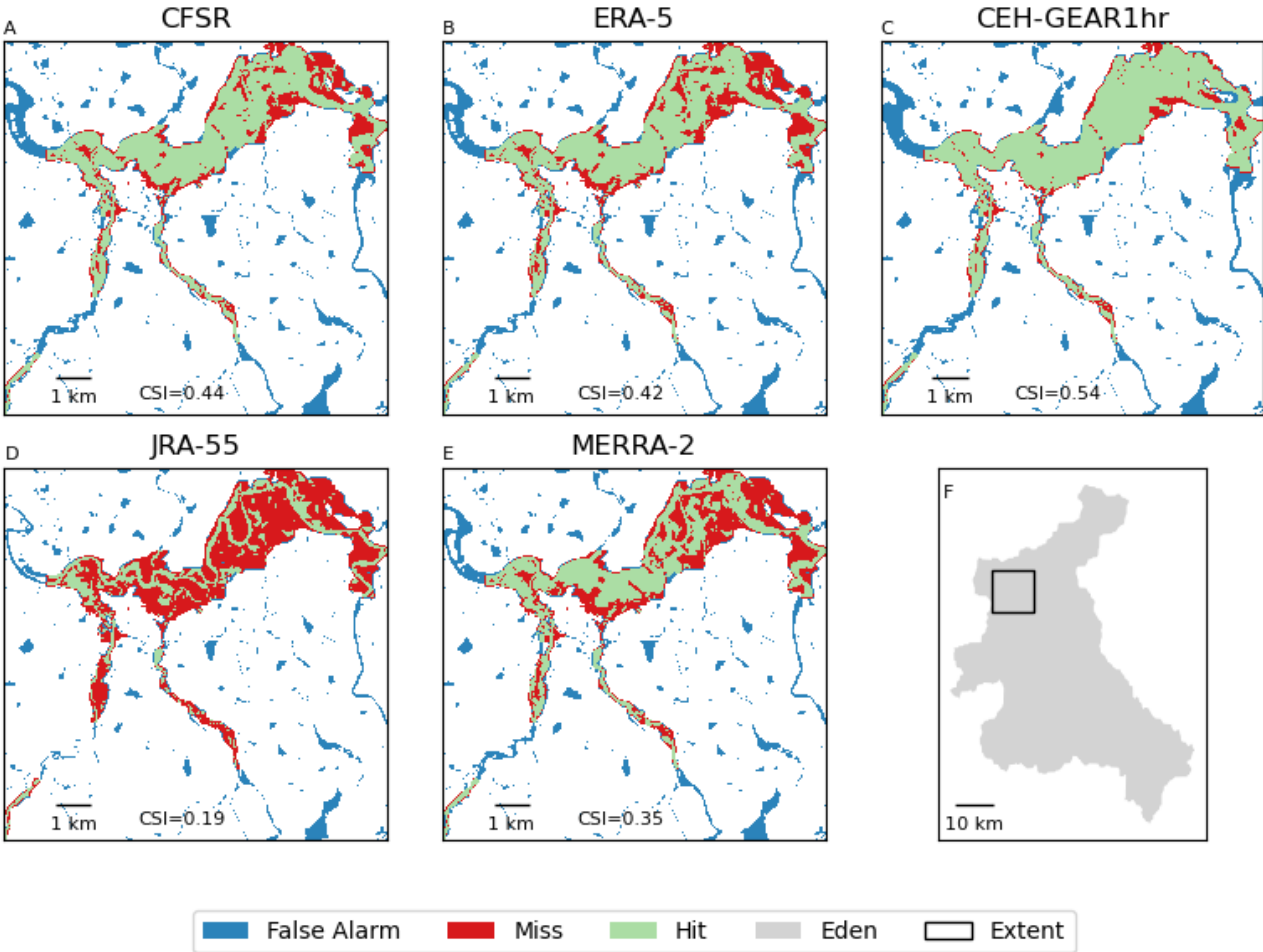
Figure 5 Comparison of flood extent based on a threshold of 0.3m with Environment Agency Recorded Flood Outlines for the 2005
event in Carlisle.

575
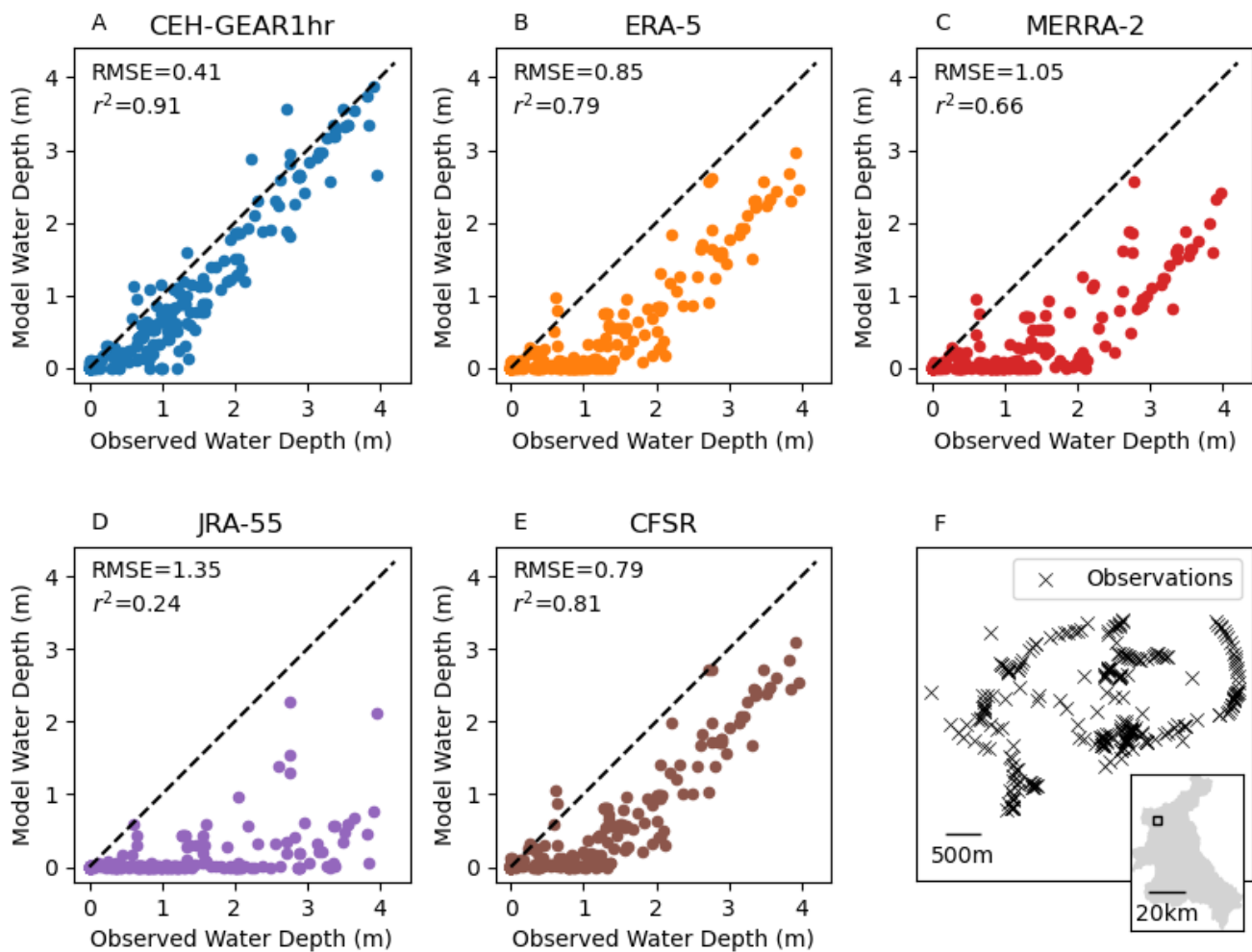
Figure 6 (A)-(E) show a comparison of modelled and observed water depths from wrack and water marks in Carlisle during the 2005 event (Neal et al., 2009)**. (F) shows the locations at which the water depths were measured.**
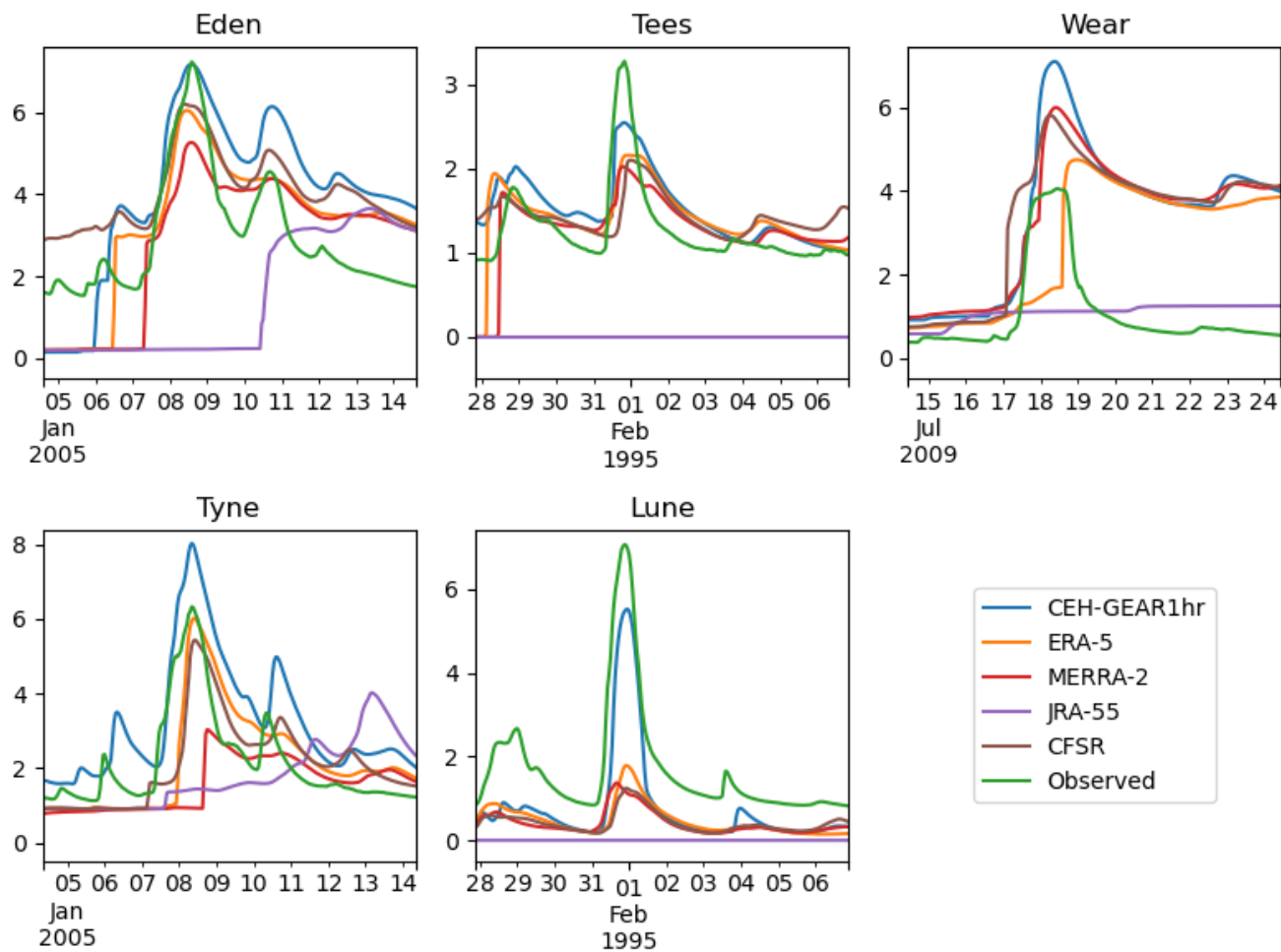
Figure 7 Stage hydrographs comparing water depths (m) from model results and observed values at each gauge.
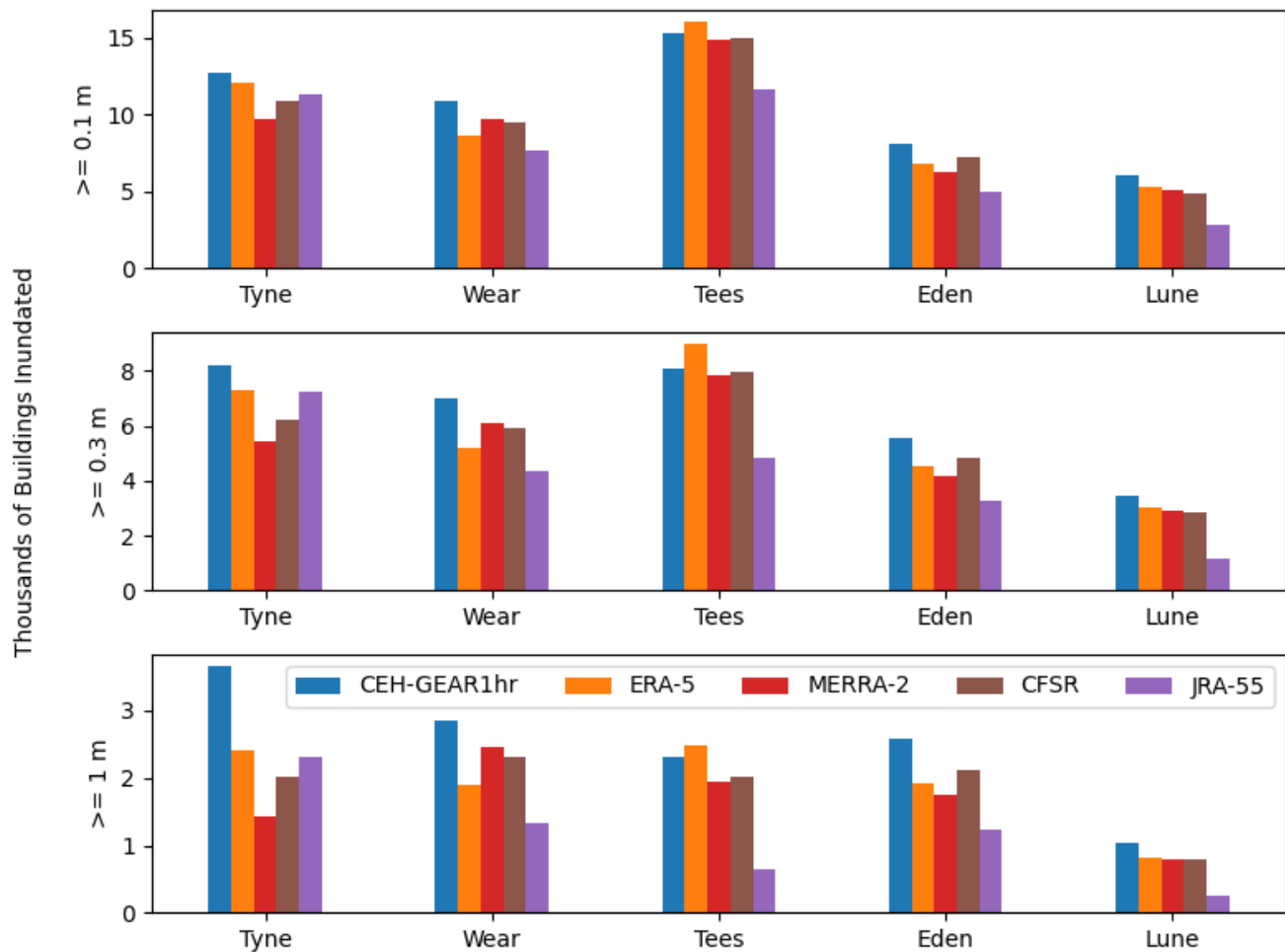
Figure 8 Number of buildings inundated above given thresholds~~a threshold (0.3 m)~~ per basin by each model.

Table 1 Precipitation products included in this study. Where the end of the period is not given, the product continues to be updated to the present day at the time of writing.

Dataset	DOI	Resolution	Coverage	Period	Frequency
CEH-GEAR1hr	10.5285/d4ddc781-25f3-423a-bba0-747cc82dc6fa	1 km	Great Britain	1990-2014	Hourly
ERA-5	10.24381/cds.adbb2d47	~30 km	Global	1979-	Hourly
MERRA-2	10.5067/7MCPBJ41Y0K6	~55 km	Global	1980-	Hourly
CFSR	10.5065/D6513W89	~35 km	Global	1979-2011	Hourly
JRA-55	10.5065/D6HH6H41	~60 km	Global	1958-	3 Hourly

585 **Table 2 Event start and end times for each basin based on the observed stage peak at the most downstream gauge. Start times are two weeks before, and end times two weeks after, the observed stage peak times to allow for model spin-up and inclusion of hydrograph. Rainfall totals are the mean rainfall between the start and end times across the gauged catchment from CEH-GEAR1hr.**

Basin	Gauge Location	Gauge ID	Catchment Area (km2)	Rainfall Total (mm)	Stage Peak (m)	Flow Peak (m3/s)	Peak Time	Start Time	End Time
Wear	Chester le Street	24009	1008.3	151.7	4.1	378	18/07/2009 11:00	04/07/2009 11:00	25/07/2009 11:00
Tyne	Bywell	23001	2175.6	169.2	6.3	1390	08/01/2005 08:00	25/12/2004 08:00	15/01/2005 08:00
Tees	Darlington Broken Scar	F3501	818.4	163	3.3	646	31/01/1995 20:15	17/01/1995 20:15	07/02/1995 20:15
Lune	Caton	724629	983	216.1	7.1	1400	31/01/1995 21:15	17/01/1995 21:15	07/02/1995 21:15
Eden	Sheepmount	765512	2286.5	216.5	7.2	1520	08/01/2005 14:30	25/12/2004 14:30	15/01/2005 14:30

590 **Table 3 Summary of performance statisticsmetrics for each model. Building inundation and floodplain depth errors are relative to CEH-GEAR1hr, therefore the CEG-GEAR1hr model has no values for these measures. CSI and water depth RMSE are reported for regions of the Eden basins corresponding to those shown in Figure 4 and Figure 5.**

		CEH-GEAR1hr	ERA-5	MERRA-2	JRA-55	CFSR
Stage peak error (%)	Wear	74.99	17.02	47.92	-68.55	43.29
	Tees	-22.28	-34.11	-38.16	-100	-35.9
	Eden	-0.9	-16.39	-27.17	-49.47	-14.24
	Tyne	26.88	-4.91	-52.04	-36.43	-14.09
	Lune	-21.81	-74.66	-80.57	-99.99	-82.37
Stage peak time error (hrs)	Wear	-2	13	-1	168	-5
	Tees	0	2	-1	-336	4
	Eden	-1	-3	0	114	-5
	Tyne	0	2	10	116	2
	Lune	2	1	-5	-79	1
Number of buildings inundated >= 0.3m	Tyne	8230	7287	5405	7240	6244
	Tees	8078	8956	7834	4801	7938
	Wear	6979	5187	6070	4357	5923
	Eden	5573	4515	4167	3262	4843
	Lune	3475	3005	2905	1177	2861
CSI		0.54	0.42	0.35	0.19	0.44
Water depth RMSE		0.41	0.85	1.05	1.35	0.79