Implications of Using Global Digital Elevation Models for Flood Risk Analysis in Cities

Fergus McClean¹, Richard Dawson¹, and Chris Kilsby²,³

¹School of Engineering, Newcastle University, Newcastle upon Tyne, UK, ²Willis Research Network, London, UK

Abstract. As urban populations grow, it is increasingly important to accurately characterize flood risk in cities and built up areas. Global digital elevation models (GDEMs) have recently enabled flood risk analysis at broad scale and worldwide, but their accuracy and its impact on modeled flood risk in cities has not been fully investigated. We compare flood extents, hydrographs, depths, and impacts between hydrodynamic simulations, using five spaceborne GDEM products and an airborne LIDAR product. Benchmark observations of a historical flood event in Carlisle (UK) were used to assess the accuracy of each simulation. GDEM simulations are shown to perform significantly less accurately than the airborne LIDAR-based simulations. No DEM outperforms the others across all metrics; the MERIT DEM is the best predictor of flood extent, but TanDEM-X performs best for discharge. However, the impacts of flooding from GDEM simulations are consistently overestimated, 2 to 3 times higher than those from LIDAR simulations. Until a high resolution, accurate, global DEM is available, multiple products should be used concurrently to enable the full uncertainty range to be quantified and communicated, to ensure flood risk management decisions are not misinformed.

1. Introduction

By 2050, approximately 68% of people will live in cities (United Nations Department of Economic and Social Affairs, 2019). Many of these cities are situated in low-lying, flood-prone areas. Consequently, a large and growing proportion of global flood risk is located in cities, and therefore accurate estimation of both the magnitude of the flood hazard and its impact in cities is crucial. The role of the DEM is key here in defining the pathway of the flood in both the upstream catchment (river channel/floodplain) and the city itself (complex urban topography of buildings and streets). The combination of complex topography and high population makes flood risk in cities both critical to understand and challenging to estimate. Global digital elevation models (GDEMs) are increasingly used in the production of continental and global flood hazard maps (Dottori et al., 2016; Sampson et al., 2015; Ward et al., 2013). The comprehensive coverage of GDEMs means that flood hazard can now be assessed in areas where low data availability had previously prevented this. However, the lower spatial resolution and vertical accuracy of GDEMs means that their use leads to increased uncertainty. The magnitude of this uncertainty and the differences between GDEM products are only beginning to be explored. When estimated flood hazard zones are used to analyze risks to infrastructure, this uncertainty propagates to damage and loss estimates. This propagation is of particular concern when pricing insurance premiums in areas were GDEMs have been used to assess flood risk.

GDEMs are usually digital surface models (DSMs) rather than digital terrain models (DTMs) as they include surface features such as trees and buildings. Some sensors, such as synthetic aperture radar, are able to at least partially penetrate vegetation but not buildings, leading to products which contain some terrain and some surface, such as TanDEM-X (Rizzoli et al., 2017). It can therefore be difficult to distinctly classify GDEMs as either DSMs or DTMs. Attempts have been made at converting global DSMs such as the Shuttle Radar Topography Mission (SRTM) to DTM products, such as Multi-Error Removed Improved Terrain (MERIT) (Yamazaki et al., 2017) and Hydrological data and maps based on SHuttle Elevation Derivatives at multiple Scales (HydroSHEDS) (Lehner et al., 2008). Hawker et al. (2018) provide a recent review of free and commercially available GDEMs, highlighting MERIT as “the most comprehensive error removal from SRTM to date.” The other widely used corrected SRTM product, HydroSHEDS, has large river channels burned in (subtracted from the DEM surface) based on limited hydrography information. This is detrimental for hydrodynamic flood modeling as channel geometry controls the bank-full depth. DTMs are preferable.
to DSMs in flood modeling applications as surface features may artificially impede flow pathways if a com-
mensurate grid resolution is not used or if the features are elevated. If the spatial resolution of the grid is
not high enough, surface features may be represented as larger than they really are. If they are elevated, as
is the case with bridge decks, water that should be allowed to pass through will be blocked by the simpli-
fied impermeable version of the feature. Therefore, products such as MERIT are expected to produce more
accurate estimates of flood extent than uncorrected global DSMs. The effect of using a DSM rather than a
DTM on the estimated spatial extent of flooding depends on the region but can be substantial. For example,
Courty et al. (2019) found that the accuracy, relative to using a LIDAR-based DTM, of flood depth and extent
when using GDEMs was highly variable.

Many large-scale flood risk assessments are based on a single GDEM (Dottori et al., 2016; Sampson et al.,
2015; Ward et al., 2013). This is because running ensemble simulations with many combinations of inputs
can become computationally expensive and unfeasible. However, selecting a single GDEM means that other
possible estimates of the spatial extent of flooding caused by a given event are excluded. The growing number
of available GDEMs is increasing the number of possible flood hazard zones that can be produced (Farr et al.,
2007; Rizzoli et al., 2017; Tachikawa et al., 2011; Tadono et al., 2014; Yamazaki et al., 2017). In cases where the
local accuracy of GDEMs can be established, it might be sufficient to only include the most accurate GDEM.
However, the fact that a GDEM is required may mean that there is no reference data available to assess
accuracy locally. Furthermore, accuracy assessments are not necessarily transferable between areas. For
these reasons, validation of flood risk estimates produced using GDEMs is an ongoing research challenge.

Flood extent validation is difficult to carry out robustly. The data sets required are often unavailable and can
be lacking in resolution, coverage, and accuracy. These issues are particularly acute in urban areas where
spatial variability in flood hazard zones has the greatest effect on damage estimates due to the high den-
sity of assets. Variation in extent accuracy is also not necessarily linked to variation in feature inundation.
Therefore, the model with the most accurate extent may not provide the most accurate estimate of feature
inundation. The estimation of building damage from floods is problematic and depends on many factors.
Depth-damage functions are used as a standard method for calculating losses from floods. However, the
building-level detail required to enable accurate functions to be designed is usually unavailable.

This study assesses the effect of using GDEMs on both inundation model performance and estimates of
impacts. A range of GDEMs are tested along with a national LIDAR product. By better understanding
the potential range of performance and impacts resulting from different GDEM products, more informed
decisions can be made about the reliability of model outputs.

2. Methods

2.1. Modeling System

The City Catchment Analysis Tool (CityCAT), developed by Glenis et al. (2018) at Newcastle University, was
designed primarily to simulate flooding in cities and solves the full shallow water equations (SWEs) numer-
ically using a Generalised Osher-Solomon Riemann solver. CityCAT has a variable timestep to maintain
stability and allows for cell wetting and drying along with hydraulic jumps. The numerical methods used
have been validated against analytical solutions and experimental data (Glenis et al., 2018). This means the
largest source of uncertainty is from the input data and not any approximation in the numerical methods, as
can be the case when using simplified versions of the SWEs to increase simulation speed. The lack of numer-
ical approximations is the main reason CityCAT was chosen for this study as it puts a focus on input data
rather than the modeling system. In this study, the model domain has been represented using only a DEM.
The computational mesh has uniform square cells and is topographically equivalent to the DEM itself. City-
CAT can include the effect of green areas, buildings, and sewer networks; however they are excluded here
due to the incommensurate resolution of the simulations.

2.2. Study Location

Carlisle was chosen as an example city due to its long history of flooding and dense urbanization with a
total of 15,030 buildings within 34 km², equating to 4.4 buildings per hectare. For reference, London has
22.5 dwellings per hectare, while England has 1.8 (Ministry of Housing, Communities & Local Government
(MHCLG), 2020). The urban area, located in the lower reaches of the Eden basin (Figure 1), was delineated
using the Urban Area Town Settlement Boundary published by Carlisle City Council (2015). Parks, golf
courses, and sports fields surround the River Eden as it passes the city centre. The southern tributaries Caldew and Petteril have smaller buffer zones and pass through the majority of the city.

Carlisle has flood defences, most of which were constructed following the severe flooding in January 2005. The two major schemes are the Eden and Petteril Flood Alleviation Scheme in the east and Caldew and Carlisle City Flood Alleviation Scheme in the west of the city. The schemes primarily consist of embankments along the Eden and floodwalls along the Caldew. Further defences are currently under construction in Carlisle along the Petteril. This study ignores the effects of flood defences as they were overtopped during the event modeled here, and many were not constructed until afterward. Some artifacts of defences which were present at the time of data collection may be present in GDEM products, but it is expected that these will have no significant effect on flood propagation due to the relatively low resolution of the grid.

### 2.3. Storm Event

Widespread flooding occurred in Carlisle from both surface water and fluvial sources in early January 2005. Approximately 1,934 properties were directly flooded, and three people died in the incident (Convery & Bailey, 2008). It took place prior to any major flood defences being constructed. The event provides a useful benchmark for this study as 263 measurements of water surface elevation and flood extent were recorded by Neal et al. (2009) on 24–26 January.

A range of previous studies have modeled the 2005 Carlisle floods. Horritt et al. (2010) tested a simple finite volume model using a DTM based on Environment Agency LIDAR and OS MasterMap data. They obtained a RMSE of 0.365 m after calibrating friction coefficients. Meanwhile, Smith et al. (2014) achieved 0.32 m RMSE using a finite-volume scheme and highlight the 0.5 m error in wrack mark measurements. Liu and Pender (2012) also used the event to compare a cellular automata model against a 2-D hydrodynamic model.
Figure 2. MIDAS rainfall observations interpolated to 1 km using inverse distance weighting for the Eden basin during the storm event. Panel (a) shows the locations of gauges used and the spatial distribution of the interpolated values, while panel (b) shows the temporal dynamics of the event.

The Met Office Integrated Data Archive System (MIDAS) Open data set (Met Office, 2019) provides quality controlled hourly time series of rainfall data recorded at weather stations during the event (Figure 2). MIDAS combines data from synoptic weather stations in the UK Land Surface Observing Network and supplementary rainfall stations, which may be either automated or manually operated. The data are quality controlled and published annually. Eight MIDAS are located within 50 km of the Eden basin, with a mean intergauge distance of 30 km. Hourly observations from these gauges were used to generate interpolated inputs for the simulations in this study. Based on interpolated values, a total of 109 mm fell between 7 January and midday on 8 January over the Eden. Rainfall totals were higher toward the southern, upstream end of the basin. It should be noted that these rainfall amounts are likely to be underestimates as winds over 25 m s\(^{-1}\) occurred during the storm period (Environment Agency, 2006) causing wind undercatch at the rain gauge by as much as 20% (Pollock et al., 2018).

2.4. DEM Data Sets

Figure 3 shows the six DEMs used in this study and their key characteristics are summarised in Table 1. All data sets apart from OS Terrain 50 (OST50) have global or near-global availability, and all six are free to access. OST50 is included to enable a comparison of DEMs based on data collected from satellites versus airborne LIDAR. LIDAR is a gold-standard for DEM data collection, however requires more time and resources to collect and therefore is unavailable for most of the globe. The five near-global products tested here provide viable alternatives where LIDAR is not available. No additional corrections were applied as part of this study as we aim to identify inherent issues with the original products, rather than test the viability of different correction algorithms. However, better performance may be achieved by applying such corrections to GDEMs (Baugh et al., 2013; Jarihani et al., 2015).
2.4.1. OS Terrain 50
The Ordnance Survey (OS) provides freely available elevation data for Great Britain at 50 m resolution, based on aerial LIDAR observations (Ordnance Survey, 2017). OS flies regular imaging missions in light aircraft to collect new data and updates their “Terrain” products yearly. The final regular gridded data set is derived from a triangular irregular network (TIN) elevation model which is used as it is able to capture edges of features more accurately. The conversion process may result in the loss of some topographic features, which is normal when generating a uniform grid from an irregular mesh. All vegetation, buildings, and supported structures such as bridges are removed to create a “bare earth” surface. However, permanent surface structures such as dams, bridge revetments, and earthworks are left unmodified. Water bodies are leveled to the height of the lowest surrounding elevation value. To assess accuracy, the DEM was compared with GPS points and found to have a vertical RMSE of 4 m.

2.4.2. AW3D30
JAXA’s Advanced Land Observing Satellite (ALOS) used its onboard Panchromatic Remote-sensing Instrument for Stereo Mapping (PRISM) to observe earth’s surface between 2006 and 2011. These data were processed and used to generate a global DEM at 0.15 arcsecond resolution with a vertical accuracy of 5 m (Tadono et al., 2014).

<table>
<thead>
<tr>
<th>DEM product</th>
<th>Spatial resolution</th>
<th>Year(s) data collected</th>
<th>Vertical error</th>
<th>Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>OST50</td>
<td>50 m</td>
<td>Unavailable</td>
<td>4 m RMSE</td>
<td>Great Britain</td>
</tr>
<tr>
<td>SRTM</td>
<td>1 arcsecond</td>
<td>2000</td>
<td>90% below 9 m</td>
<td>Near Global</td>
</tr>
<tr>
<td>MERIT</td>
<td>3 arcseconds</td>
<td>2000</td>
<td>58% below 2 m</td>
<td>Near Global</td>
</tr>
<tr>
<td>ASTER</td>
<td>1 arcsecond</td>
<td>2000–2010</td>
<td>95% below 17 m</td>
<td>Near Global</td>
</tr>
<tr>
<td>AW3D30</td>
<td>1 arcsecond</td>
<td>2006–2011</td>
<td>5 m RMSE</td>
<td>Near Global</td>
</tr>
<tr>
<td>TanDEM-X</td>
<td>3 arcseconds</td>
<td>2010–2015</td>
<td>90% below 10 m</td>
<td>Near Global</td>
</tr>
</tbody>
</table>
2.4.3. TanDEM-X
TanDEM-X is a relatively new product from the German Aerospace Centre, DLR (2018). The 90 m version is fully open, while 30 and 12 m products are available on request. As this is a new product, it has not been fully processed to remove artifacts, outliers, noisy areas, or voids (Earth Observation Center, 2018). Initial efforts are being made to correct the original data (Archer et al., 2018); however producing a fully hydrologically conditioned product will require considerable work. Due to its relative nascence and lack of corrected products, TanDEM-X is yet to be widely used within the flood modeling community.

2.4.4. ASTER
The Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) (Tachikawa et al., 2011) on-board the scientific research satellite Terra has been used to create three versions of a global DEM. The latest release, made available in August 2019, adds additional stereo-pairs in a refined production algorithm to produce fewer artifacts and improved accuracy and resolution. ASTER and PRISM are both passive sensors; however ASTER uses infrared rather than panchromatic imaging. This may partially explain why ASTER topography appears to differ strongly from the others in Figure 3.

2.4.5. SRTM
NASA’s Shuttle Radar Topography Mission (SRTM) (Farr et al., 2007), flown in 2000, remains the highest quality set of freely and globally available land elevation observations available, covering almost the entire earth surface between 60 degrees north and south. A single satellite, carrying two radar antenna connected by a mast, was used to measure distances to earth from two fixed locations on the instrument, allowing elevation to be calculated using interferometry.

2.4.6. MERIT
The Multi-Error-Removed-Improved-Terrain (MERIT) DEM, produced by Yamazaki et al. (2017) seeks to further improve the accuracy of SRTM, particularly on flood plains. The project aimed to remove speckle noise from surface reflectance, stripe noise from motion errors of the sensor, absolute bias caused by limited ground control points, and tree height bias where the radar incorrectly classifies canopies as the land surface. However, there was no specific effort to remove man-made structures. The accuracy of MERIT was tested using ICEsat which can penetrate forest canopies. Positive bias in forested areas was found to be reduced compared with the original SRTM data; however there was little difference in mountainous regions due to large topographic variability within pixels. Fifty-eight percent of corrected pixels were within 2 m of the ICEsat elevations, compared with only 39% in SRTM (Yamazaki et al., 2017). Validation was also conducted across eastern England using the UK 1 m LIDAR data where the processed DEM again showed improved agreement compared to the original SRTM data.

2.5. Model Setup
While validation and impacts analyses are only carried out within Carlisle, the entire Eden basin (Figure 1b) was simulated in order to include discharge from upstream. To extract domains from each of the DEM products, the basin outline for the Eden was extracted from the HydroBASINS data set (Lehner & Grill, 2014). Hourly MIDAS rainfall data from gauges within a 50 km boundary of the basin was interpolated using inverse distance weighting to a resolution of 1 km. The interpolation approach used introduces further uncertainty but is required to produce a continuous rainfall field. The boundaries of the domain were treated as open and the friction coefficient, Manning’s $n$, was set at 0.03 everywhere. This value was chosen based on Chow (1959) and realistically assumes the majority of the Eden catchment to be pasture with short grass. As this study focuses on the relative effects of using different DEMs, the friction parameter is not a key factor here. However, it is acknowledged that better performance may be obtained by optimizing this parameter or including spatial variability. Hydraulic modelers generally use Manning’s coefficient as a calibration parameter, often selecting very low values to compensate for deficiencies in the numerical method. Our approach is to use standard values of Manning’s coefficient as CityCAT is not subject to such large errors, and we believe that much better understanding and transparency is obtained in this way. In this paper we take this one step further and use the same value throughout the catchment, so we can focus on the effects of the GDEM choice. The domain was treated as being impermeable to represent the saturated conditions preceding Storm Desmond. Each DEM was reprojected to the European Lambert Azimuthal Equal Area (ELAEA) projection and bilinearly resampled to 50 m resolution to make each model cell identical across data sets apart from its elevation. If different resolutions were used for each DEM then the performance would be affected by this factor and the results would not be comparable. Using a resolution finer
Table 2
Subset of EA Recorded Flood Outlines Data Set for Carlisle 2005

<table>
<thead>
<tr>
<th>Name</th>
<th>Boundary source</th>
<th>Flood source</th>
<th>Cause</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Crosby 08_01_2005</td>
<td>Visual</td>
<td>Main river</td>
<td>Channel capacity exceeded (no raised defences)</td>
</tr>
<tr>
<td>Dalston 08_01_2005</td>
<td>Visual</td>
<td>Main river</td>
<td>Channel capacity exceeded (no raised defences)</td>
</tr>
<tr>
<td>Harraby Green 08_01_2005</td>
<td>Visual</td>
<td>Main river</td>
<td>Overtopping of defences</td>
</tr>
<tr>
<td>Carlisle (drainage) 08_01_2005</td>
<td>Local authority</td>
<td>Drainage</td>
<td>Local drainage/surface water</td>
</tr>
<tr>
<td>2005 Carlisle Event</td>
<td>Visual</td>
<td>Main river</td>
<td>Overtopping of defences</td>
</tr>
</tbody>
</table>

than 50 m would not add any further detail to OST50 (the reference DEM) and lead to unnecessary computation times for the Eden basin. Flood depths were stored hourly and the maximum depth throughout the model extracted for each cell.

2.6. Validation

A combination of flood depth and extent measurements from the event were used to assess the accuracy of each simulation. Using a combination of both extent and depth provides more potential for detecting variability in performance between models.

2.6.1. Measured Wrack Marks and Water Levels

Neal et al. (2009) provide 263 measurements of wrack marks and water levels recorded after the event. The data are available as mean heights above sea level (MASL) rather depths above ground level. Therefore a DEM must be used to generate a depth value at each point. The approach used here was to, for each model, use the corresponding DEM to calculate a depth value from the observed elevation and then compare this depth to the modeled depth.

2.6.2. EA Recorded Flood Outlines

The UK Environment Agency (EA) provide outlines of peak flood extent based on photos, videos, and level measurements (Environment Agency, 2020). The five polygons corresponding to the January 2005 event in Carlisle were extracted from this data set and are shown in Table 2. The descriptions of how each polygon was derived are limited to either “visual” or “Local Authority.” There is no further information about what “visual” refers to. These polygons were merged, and the resulting maximum extent outline was compared to the maximum extent from each model.

To compare modeled extent with the recorded flood outlines, the Critical Success Index (CSI) (Donaldson et al., 1975) was used. CSI is the most appropriate indicator for assessing extent accuracy as it describes both the ability to flood flooded cells and keep nonflooded cells clear. Cells correctly identified as nonflooded are ignored. Using other measures such as hit rate can be misleading as if the entire domain is flooded by the model, this would result in a hit rate of 100%. CSI effectively combines hit rate with false alarm ratio to create a more holistic metric. To achieve a good CSI score, a combination of high hit rate and low false alarm ratio is required.

2.6.3. EA Stage Data

While wrack marks provide distributed and precise measurements of peak flood depths, they are unable to capture the dynamics of the flood wave. For this, continuous observations are required which are only available at pre-existing gauging stations; 15 min stage data were obtained from the Environment Agency via a freedom of information request.

2.7. Impact Analysis

The buildings layer from OS VectorMapLocal (VML) (Ordnance Survey, 2018) was used to describe the location and geometry of buildings within the study area. Only buildings over 20 m² are included in VML, and each feature may be an amalgamation of multiple buildings. Therefore, the actual count of inundated buildings may be higher than estimated using these data.

Thresholds of between 0.01 and 1 m were used to produce total counts of inundated buildings for each model within the study area. Any building with a maximum depth below 0.01 m was considered not flooded and excluded from the analysis. Ideally, each building would be associated with its own stage-damage function related to its vulnerability. However, vulnerability analysis is outside the scope of this study; therefore a uniform binary inundation classification has been used. The methodology can be adapted in the future to account for vulnerability.
3. Results

The results from comparisons of observed and simulated water levels, and flood extents are described below, followed by an analysis of the estimated magnitude of inundation from each simulation. Key statistics are summarized in Table 3.

3.1. Channel Water Level

Figure 4 shows observed and modeled water height at the most downstream EA gauge in the Eden catchment. Observed water heights were replicated better by models using the corrected LIDAR DEM than those using GDEM products. GDEM models greatly underestimated stage at the gauge throughout the period modeled. TanDEM-X produced the highest peak of the global products but still did not reach above a third of the observed level. However, all products apart from ASTER did capture the timing of the flood peak.

A possible explanation for the large difference in accuracy between using LIDAR and GDEM products is the existence of Eden Road bridge on the A7 upstream of the gauge and the West Coast Main Line crossing just downstream. These features may have been removed in OST50 but not in the GDEMs, resulting in more water moving onto the flood plain and not reaching the gauge at Sheepmount. The spatial pattern of the railway line can be seen in the modeled maximum depths from each GDEM in Figure 4, running northwest to southeast and passing the gauge just downstream. The pattern was less pronounced when using MERIT, indicating that the correction methodology of Yamazaki et al. (2017) did have a positive effect. Performance may be improved by manually or automatically removing these bridges; however that is outside the scope of this study as it would skew the results toward certain DEMs, depending on the algorithm used. We highlight the presence of bridges as a key limiting factor in the potential for GDEMs in hydrodynamic modeling and one that has so far not been addressed at the global scale.

Vegetation has been removed in OST50 (Ordnance Survey, 2017). This may play a role in helping to capture the topography around the river channel more accurately. Meanwhile, vegetation artificially raises the wooded banks of the Eden in the GDEM products. Efforts were made by Yamazaki et al. (2017) to remove vegetation from MERIT globally; however this can be challenging for smaller forested areas such as those in Carlisle, and the process may not have been effective here.

3.2. Floodplain Water Level

Figure 5 shows the correlation between observed and modeled maximum water depths on the floodplain. The benchmark data used here are one of the most comprehensive collections of ground-based flood measurements available in the world (Neal et al., 2009); 216 of the 263 available observations fell within the urban area boundary of Carlisle and were therefore included in the analysis. The number of points corresponding to each model cell ranged between 0 and 7; 74% of cells with any observations had only one, while 23% had 2 or 3. As expected, models using OST50 produced the most accurate depth estimates, with an RMSE of 2.04 m. The ASTER model showed a strong negative bias and resulted in the least accurate depths by a considerable margin. This bias is likely to be caused by poor flow pathway connectivity and therefore more even distribution of water on the floodplain. TanDEM-X depths had the lowest RMSE of all GDEMs (2.76 m) and were more correlated with the benchmark than OST50. MERIT depths were less accurate than SRTM in terms of RMSE but were more correlated with the benchmark.

There was a strong positive bias in depths from all GDEM models apart from ASTER. This might be explained by the presence of surface features in GDEMs artificially lowering the benchmark observations as each absolute water elevation was converted to a depth using the same DEM as the model. Ideally, depth
Figure 4. Comparison between modeled water height and observed river stage (a) for the Eden at Sheepmount (b) and maps of maximum modeled water depths (above 0.3 m) surrounding this gauge (c). The black box shown in (b) refers to the extent of plots in (c). The extent appears skewed as (b) and (c) use different spatial projections.

observations would be used here instead of absolute water elevations; however only absolute water elevations were available. The reduced capacity and connectivity of river channels in GDEM will also lead to more water moving onto the floodplain and increasing depths. A third possible explanation for the positive bias is that it is an artifact of the difference in vertical datum between the GDEMs (EGM96) and the GPS measurements (ODN).

A large number of observed water elevations were below the elevation of all DEMs. This was likely caused by the spatial resolution of the DEMs meaning that very localized topographic characteristics were not captured. For example, many depths were recorded between buildings, within gaps representing areas of lower elevation smaller than the grid size of the DEMs used here.

3.3. Flood Extent

Figure 6 shows how similar each modeled flood extent was to the EA recorded outlines for the event. OST50 was the most similar to the EA outlines with a CSI of 0.56, closely followed by MERIT with 0.54. ASTER
Figure 5. Comparison between modeled water height and wrack marks measured after the event (a). Points more than 2 m below the DEM surface are not shown in the plot but were included when calculating RMSE and $r^2$. The locations of included observations are shown in (b).

Figure 6. Comparison between modeled maximum flood extent above 0.1 m and EA recorded flood outlines.
Figure 7. Numbers of buildings inundated by each model. Gray indicates that no buildings were inundated.

performed substantially worse than the other satellite products with a CSI value below half of any other. The speckling in ASTER led to a higher proportion of both misses and false alarms which resulted in a low CSI value. Apart from ASTER, all uncorrected GDEM products scored 0.49. The correction in MERIT led to an improvement over the original SRTM product of 0.05.

There was a lower rate of false alarms in OST50 than other models, but the main extent boundary did not extend to the observed outline. This meant that the number of misses also increased, and therefore CSI was not substantially different to the value produced by MERIT.

The similarity between AW3D30, TanDEM-X, and SRTM was unexpected, given that they are based on data collected at different times and using different types of sensors. The original spatial resolution of TanDEM-X was also lower than AW3D30 and SRTM. There were noticeable differences in flood extent between the products; however resolution did not seem to have had a detrimental effect on performance here.

The observed flood outlines are subject to uncertainty, and sparse metadata means it is difficult to determine their accuracy. However, the EA outlines provide the only source of openly available recorded flood extent data in the UK.

3.4. Inundation

Figure 7 shows the number of buildings inundated by each model by threshold and location. Fewer buildings were inundated when using OST50 than any of the GDEMs. The use of AW3D30, MERIT, and SRTM all led to approximately 2,000 buildings above 0.3 m being inundated, while the use of ASTER led to substantially more being inundated. The MERIT counts were most similar to the OS results. More variation was evident at moderate inundation levels than at the extremes, particularly between OS and ASTER. Thus, the threshold depth used to determine whether a building is flooded is a key modeling assumption as it amplifies uncertainties in the choice of DEM on flood impacts.

The increased inundation resulting from use of the ASTER DEM can be explained by speckle noise, as discussed earlier, leading to poor representation of drainage pathways and therefore accumulation on the floodplain. The lower inundation levels in the OST50 results were partly caused by the increased smoothness of the DEM, which led to the opposite effect. The river bathymetry is also better represented in OST50, and buildings have been removed, along with elevated structures such as bridges to better reflect the bare earth surface. Surface features can impede flood water and lead to increased floodplain inundation.

Whether features such as bridges and buildings should be present in the surface when modeling floods is location and feature dependent. For example, high bridges which can accommodate extreme flows should be removed, while smaller bridges that are more likely to interrupt channel flow should be included. Most buildings are smaller than the size of the model grid used in this study, and therefore it could be argued that
they should be excluded here. This presents a fundamental problem with many GDEMs, as they tend to have grid sizes larger than typical buildings while potentially including building elevations in their surfaces. To avoid this issue, corrected DTMs such as OST50 are required.

4. Discussion

The accuracy of flood models in cities is shown to decrease when using GDEMs in place of a LIDAR-based DTM, in terms of discharge, flood depths, extents, and impacts. Extents were found to be less sensitive than depths but were still less accurate when using GDEMs. This has wide-reaching consequences for flooding and climate change impact assessments that use models based on a GDEM. In many circumstances, GDEMs can provide data where it was previously unavailable; however, they should not be seen as a viable replacement for LIDAR when investigating detailed flood impacts, especially in cities.

These findings initially appear to contradict Fleischmann et al. (2019), who report that a locally derived DEM did not lead to any significant improvement compared to SRTM in the Itajaí-Açu basin (Brazil). Differences in model setup partially explain this; however, their study compared CSI over the entire 15,000 km² catchment. However, when considering just the city of Rio do Sul the local DEM provided a 29% improvement over the SRTM. This points to a systematically increased sensitivity to DEM quality in urban areas.

Models using a GDEM calculated a higher number of buildings to be inundated. This implies that flood risk may be exaggerated when calculated using a GDEM. In this case study, using the average UK flooding threshold depth of 0.3 m (Environment Agency, 2019), the impacts are at least doubled by using any GDEM. This effect will vary between regions, but the results presented here provide evidence that it should not be ignored. The threshold depth was found to have a significant impact on the variability of inundation between models. The impact was less sensitive to the choice of DEM for either extremely high or extremely low threshold values. However, these would not be realistic to use in a flood risk analysis. As the hydraulic model simulates precipitation falling directly onto the DEM surface, a threshold depth of zero would lead to all buildings being inundated. A very high threshold depth would imply that all buildings are constructed on elevated foundations or stilts. As threshold depth is a proxy for vulnerability, the resilience of individual assets has an effect on the importance of DEM accuracy.

Variability was much greater between numbers of inundated buildings than between the accuracy of flood extents. This is because buildings and other assets at risk of flooding are only present in a relatively small proportion of the modeling domain. Moreover, when situated near the floodplain boundaries, as many often are, small changes in extent which may not alter CSI much can still cause a substantial difference in the number of buildings affected. CSI, therefore, hides variability in the magnitude of inundation between models, reducing its efficacy as a measure of validation for flood impacts and risk assessment. A modified version of CSI which only measures accuracy where assets are present, or puts higher weight on these areas, may provide a more appropriate indicator but would make the results even more location specific.

The flood depths, extents, and impacts were quite similar for the analysis using the GDEMs. Archer et al. (2018) found that TanDEM-X shows improved accuracy over SRTM but not MERIT, highlighting the importance of removing surface features from GDEMs. The analysis presented here goes beyond this and other GDEM model intercomparison studies by validating against observations from a historic flood event and also assessing the uncertainties on impact metrics. Based on the results presented here, choosing the optimum GDEM for flood modeling in Carlisle is dependent on whether depths, extents or impacts are seen as the most important factor. TanDEM-X resulted in the most accurate floodplain depths and river discharge, while MERIT provided a better CSI and had the closest estimate of the number of buildings inundated above 0.3 m (1,872) when compared to corrected LIDAR (926). Hawker et al. (2019) suggest that TanDEM-X should be used alongside MERIT in flood risk applications, and our findings reinforce this message. We go further and recommend that AW3D30 should also be included to quantify the full range of uncertainties in flood risk estimates calculated using GDEMs. There were no systematic effects of the original spatial resolution of GDEM products on the accuracy of flood depths and extents. The higher resolution SRTM and AW3D30 products did not lead to more accurate results than TanDEM-X, MERIT, or ASTER.

The relative importance of flood extent, floodplain depth and the stage hydrograph is ultimately dependent on the application. It could be argued that the hydrograph does not need to be accurate in impact studies if peak floodplain extent is accurate enough. In some cases, model accuracy will be similar in the channel and...
on the floodplain, in others, such as here, it may be quite different. Local characteristics, such as the railway bridge downstream of the gauge used in this study, may mean that channel flow is less accurately simulated but a well-constrained floodplain could result in the inundation of surrounding areas being unaffected. The relationship between flood depth and extent is defined by how constrained the floodplain is. Highly constrained floodplains may hold increasing depths of water with no effect on maximum flood extent.

The findings presented here were benchmarked against high quality observations of water depths and extents in the United Kingdom. However, in other areas around the world, these data are not so readily available and records can be either difficult to access, incomplete or low quality. One solution to obtain water level measurements in these areas is using remotely sensed observations of river widths to derive flows and depths (Gleason & Smith, 2014). Sources of benchmark extent data could also become more readily available globally as new satellite data processing methods are developed (Clement et al., 2017). However, the capability of satellites is very limited in urban areas where their sensors are blocked from reaching ground level.

A lack of observed data describing flood impacts to buildings and infrastructure during the 2005 event has limited the validation of flood impacts here. Such data would enable validation of the final calculation step of a typical flood risk assessment. Insurance claims are one avenue that has already been investigated (Zischg et al., 2018). However, exact timings and depths are often not included in the claims, and the data itself can be difficult or impossible to access.

The sensitivity of flood risk assessments to the choice of DEM are undoubtedly location specific. Without local, high quality validation, using historic observations, it is not yet feasible to benchmark the accuracy of GDEMs for flood modeling in all urban areas. There is therefore a need for further investigation of events in other cities which have been recorded with the same level of detail as Carlisle 2005. Scale may play a role as cities vary greatly in size and this could influence their sensitivity to topography. As a pragmatic solution, GDEMs can be corrected and improved using various techniques (Hawker et al., 2018; Kulp & Strauss, 2018; Mason et al., 2016). However, given the data that is currently available and the findings presented here, flood risk estimates in urban areas using GDEMs should be interpreted accordingly.

5. Conclusions

Flood risk assessments for cities produced using GDEMs should be interpreted with caution as they are likely to overpredict risks. We found variability in the accuracy of models using different GDEMs. The corrections applied to the MERIT DEM had a positive effect on flood extent accuracy, relative to other GDEMs, making it the most appropriate choice of GDEM if this is the primary measure of interest. However, TanDEM-X provided the best performance for river channel discharge, making it a more appropriate choice for applications where timing, channel level and flow are important. However, all resulted in substantially higher impacts than the DEM produced from aerial LIDAR survey—with GDEMs estimating the number of buildings flooded to be 2 to 3 times higher. This effect is pronounced and should be considered by both producers and users of flood risk estimates based on GDEMs.

As the world’s cities grow, and climate and land use changes increase flood hazard, the importance of accurately understanding current and future flood risk is increasing. GDEMs have enabled flood risk assessments to be undertaken universally, with a standardized methodology allowing easy intercomparisons. However, they do not negate the need for locally informed study and should not be interpreted as a replacement. Uncertainties in flood risk assessment using GDEMs need to be properly quantified and communicated to insurers, local and national authorities and communities, to ensure flood risk management decisions are not misinformed. We therefore repeat, and add greater urgency to, the previous calls for a higher resolution and more accurate global DEM for flood modeling (Sampson et al., 2016; Schumann & Bates, 2018). In the meantime, as our analysis does not identify a single best GDEM, we recommend that the available products should be used alongside each other in flood risk applications to quantify the full uncertainty range in impacts.

Data Availability Statement

The data and code used to produce the figures in this paper are available for download from Newcastle University (McClean et al., 2020). OST50 is available from the Ordnance Survey (https://ordnancesurvey.co.uk). ASTER and SRTM are available from NASA EarthData (https://search.earthdata.nasa.gov/). AW3D30
is available from JAXA (https://www.eorc.jaxa.jp/). MRIT is available from OKI Lab (https://hydro.is.u-tokyo.ac.jp/). TanDEM-X is available from DLR (https://tandemx-science.dlr.de/).

Acknowledgments
The authors were supported by the Natural Environmental Research Council (NERC) Centre for Doctoral Training for Data, Risk and Environmental Analytical Methods (DREAM) (NE/M009099/1), TWENTY 65: Tailored Water Solutions for Positive Impact (EP/N010124/1) and Willis Research Network. The point observations of flooding in Carlisle were provided by Dr. Jeffrey Neal. The City Catchment Analysis Tool (CityCAT) was developed and provided by Vassiliis Glenis.

References


