Bare-Earth DEM Generation in Urban Areas for Flood Inundation Simulation Using Global Digital Elevation Models

Yinxue Liu, Paul D. Bates, Jeffery C. Neal, and Dai Yamazaki

1School of Geographical Sciences, University of Bristol, Bristol, UK, 2Institute of Industrial Sciences, University of Tokyo, Tokyo, Japan

Abstract Accurate terrain representation is critical to estimating flood risk in urban areas. However, all current global elevation data sets can be regarded as digital surface models in urban areas as they contain building artifacts that cause artificial blocking of flow pathways. By taking surveyed terrain and LIDAR data as “truth,” the vertical error in three popular global DEMs (SRTM 1″, MERIT DEM, and TDM90) was analyzed in six European cities and an Asian city, with RMSE found to be 2.32–5.98 m. To increase the utility of global DEM data for flood modeling, a Random Forest model was developed to correct building artifacts in the MERIT DEM using factors from widely available public datasets, including satellite night-time lights, global population density, and OpenStreetMap buildings. The proposed correction reduced the vertical errors of MERIT by 15%–67%, despite not using data samples from the target city in training the model. When training data from the target city was included, error reduction improved by between 57 and 76 percentage points. The resulting Urban Corrected MERIT DEM improved simulated inundation depth by 18% over original MERIT in a hydrodynamic model of flooding in the UK city of Carlisle, although it did not outperform TDM90 at this site. We conclude that the proposed method has the potential to generate a bare-earth global DEM in urban areas with improved terrain representation, although in data scarce regions this requires more complete OpenStreetMap building information. In the future, the method should be applied to TDM90.

Plain Language Summary Terrain representation plays a vital role in flood mapping. For wide area flood simulation where the model grid is typically larger than individual buildings, topography data without ground objects, such as buildings and trees, are preferred as this can generate more accurate inundation simulations than when ground objects are included. However, current global topography data all contain building height artifacts to some extent. This is especially a problem in areas where buildings are densely packed. The resulting vertical biases in popular global topography data were found to be 2.32–5.98 m (root mean square error) in seven cities in Europe and Asia. To address this problem, this study describes an approach using publicly available datasets that can reduce this bias by a significant amount (15%–67%). With the bias reduced topography, simulated water levels were improved over the original topography in a case study of flooding in the UK city of Carlisle. The proposed correction model has the potential to generate a bare-earth topography in urban areas globally, which will be useful for improving flood mapping and risk assessment.

1. Introduction

In the context of climate change and urban development, urban flooding issues are becoming more prevalent (Ford et al., 2019), highlighting the need for accurate flood mapping in these areas. Precise representation of terrain is of great significance for estimating flood risk (Sampson et al., 2015; Schumann & Bates, 2018; Yamazaki, Sato, et al., 2014) and although LIDAR surveys can provide high accuracy DEM data with vertical error of a few tens of centimeters, publicly available LIDAR data are limited to a handful of developed countries. Free-to-access spaceborne global DEMs (GDEMs) based on radar interferometry and photogrammetry are still the only viable data source for flood inundation simulation in many urban regions of the world. However, all such data sets are digital surface models (DSMs) in urban areas (Gamba et al., 2002) due to the reflection of radar and optical signals from ground objects such as buildings. As a
result, spurious artifacts are created which can block or alter flow pathways in ways that lead to anomalous results when these data are used in hydrodynamic modeling (Neal, Bates et al., 2009).

The undesirability of these blockage effects is, in part, a function of the DEM horizontal resolution. GDEM data are typically made open access at 1″ or 3″ pixel sizes (~30 and ~90 m at the equator respectively) and at these scales it is largely impossible to resolve individual buildings. The only solution for hydrodynamic modeling given such data is to remove the building artifacts from the DSM to create a digital terrain model (DTM) and then simulate flows over these bare earth elevations. By contrast, airborne LIDAR data with <2 m horizontal resolution can distinguish building shapes and the topology of the street layout. Then, if the computational cost of high-resolution simulations can be afforded, building resolving hydrodynamic models can be built using the DSM (Fewtrell, Duncan et al., 2011). Unfortunately, terrain data of LIDAR-like quality for most global urban areas may not be available for some considerable time, even if we could afford to simulate all urban areas at ~ meter scale. Hydrodynamic modeling at regional scale or beyond will therefore continue to be based on bare-earth simulations and the available GDEMs for the foreseeable future.

Among GDEMs, the Shuttle Radar Topography Mission (SRTM) DEM has been the most popular data source with flood inundation modelers for nearly two decades. It is a spaceborne interferometry radar data product acquired in February 2000, covering lands between N60°and S56°. Several versions of SRTM (3″ nonvoid filled, 3″ void filled, and 1″ void filled with global coverage) have been released with improving completeness or quality. The most recent version is the 1″ (~30 m at equator) void-filled data set released in late 2015 (https://www2.jpl.nasa.gov/srtm/). Another spaceborne DEM, TanDEM-X, is attracting keen interest for its unprecedented resolution (Hawker et al., 2019; Rizzoli et al., 2017; Wecklich et al., 2018; Wessel et al., 2018). It was generated from satellite interferometric radar with fully global coverage at 0.4" (~12 m at the equator) resolution. A 3" (around 90 m at the equator) resampled version (i.e., TDM90) was recently released for free access (https://geoservice.dlr.de/web/dataguide/tdm90/).

Although corrections of SRTM have been produced in many studies (Chen et al., 2018; J. Gallant, 2011; Gallant & Read, 2009; Hutchinson et al., 2009; Pham et al., 2018; Stevenson et al., 2010; Su et al., 2015; Wendleder et al., 2016; Zhao et al., 2018), most research has focused on the correction of vegetation biases (e.g., Baugh et al., 2013; O’Loughlin et al., 2016). Almost uniquely, Yamazaki et al. (2017) identified and removed multiple error sources (speckle and stripe noise, tree height, and absolute bias) from multiple GDEMs (SRTM 3″ at below 60°N, AW3D-30 m at above 60°N, Viewfinder Panoramas DEM for voids in the previous data sets), generating the “Multi-Error-Removed Improved-Terrain DEM” (MERIT DEM). MERIT has been proven to outperform the original SRTM in many places (Archer et al., 2018; Hawker et al., 2019, 2018; Hirt, 2018; Yamazaki, Ikeshima, Sosa et al., 2019), however corrections to MERIT have yet to be developed for urban areas despite these locations being where most flood risk is concentrated. The complexity of urban structures makes spaceborne terrain measurement problematic, as built-up facades often lead to layover, shadow, and multipath reflection due to the side-looking character of the synthetic aperture radars used to produce GDEMs. In addition, the smooth surfaces of some building roofs, stadiums, and water features can reduce backscattering of the radar signals and the coarse resolution of GDEMs limits how man-made features can be resolved. Thus, the elevation recorded in GDEM data over built-up areas represents an ambiguous height that is lower than the top of buildings but above the bare-earth ground level (Farr et al., 2007; Rossi & Gernhardt, 2013). Overestimation of some ground elevations in areas without buildings and trees in the 0.4” TanDEM-X DEM has also been reported (Rossi & Gernhardt, 2013). Although there is emerging research investigating the vertical error of GDEMs (Uuemaa et al., 2020), thorough studies of GDEM error specific in urban areas are lacking. In a very recent case, the impact of GDEM error on flood simulation in urban areas was examined (McClean et al., 2020). Overall, methods to correct these DEM data to bare earth in urban areas, as would be necessary for modeling flood inundation at this resolution, are lacking.

Estimating the building height bias in GDEMs by modeling directly the interaction between the various building facades in urban areas and the changing radar incidence angle is hugely complicated. Alternatively, the building height bias can be estimated by taking advantage of other relatively rich data sources, where machine learning can maximize the benefit through its ability to deal with a large volume of data in nonlinear systems (Lary et al., 2016) and its advantage in avoiding over-fitting (Breiman, 2001). Machine learning methods have been widely used in many parts of earth system research and beyond (Cooper et al., 2019;
Sinha et al., 2019; Stevens et al., 2015). For example, Kulp and Strauss (2018) adopted a multiple layer perceptron artificial neural network method to correct the SRTM 1″ DEM in coastal areas. Recently, building information from a global scale volunteer community – OpenStreetMap – has become widely available, and this can provide detailed information about not only building footprints but also building heights. This building information is contributed by volunteers around the world (https://wiki.openstreetmap.org/wiki/Buildings). The building footprint is reported with accurate locations (within 4–6 m offset) and with variable completeness (e.g., between 23% and nearly 100% coverage in some German cities). The completeness of building height data is much lower at around 0.5%, but it is growing over time (Fan et al., 2014; Haklay, 2010; Hecht et al., 2013). In addition, satellite night-time lights data has long been recognized as an indicator of human settlement (Elvidge et al., 2007; Lu et al., 2008; Shi et al., 2020; Xie et al., 2016) and recently high-resolution gridded population data have become available (Stevens et al., 2015). These datasets, along with machine learning methods suggest a possible way to estimate and remove building artifacts in GDEM data. Although a similar approach has been demonstrated in coastal areas, Kulp’s correction examined only for areas with elevation of 0–20 m which left many urban areas uncovered, and it was implemented based on SRTM which still contains speckle and stripe noise, tree height, and absolute bias compared to MERIT. This research therefore aims to further advance the usage of machine learning methods in SRTM correction by building a Random Forest model with more comprehensive, urban-targeted variables to correct the MERIT DEM data in urban areas by reducing building biases. This new bare-earth DEM is expected to principally, but not exclusively, improve the DEM data used in flood inundation models. To develop the method, seven cities were selected as case studies. First, the vertical error of publicly available GDEMs (SRTM 1″, MERIT, TDM90) in these locations was analyzed. Second, the vertical errors in the MERIT DEM were estimated using a Random Forest regression method based on a variety of global open access datasets, including night-time lights data, world population density, building information from OpenStreetMap, slope and elevation values. The resulting Urban Corrected MERIT DEM (MERIT-UC) was evaluated against LIDAR and surveyed terrain data and by its ability to improve a simulation of observed inundation that occurred during a major flood in the UK city of Carlisle in 2005 (Neal et al., 2009).

2. Data and Methodology

In this research, central Beijing (633 km²), Berlin (881 km²), and five UK cities, London (1,243 km²), Manchester (97 km²), Bristol (112 km²), Cambridge (38 km²), and Carlisle (58 km²), were selected for analysis because of the availability of benchmark terrain data, as well as their varied city type and size. The initial part of this research investigates the vertical error of existing GDEMs as a precursor to the main aim of correcting building biases. In order to clearly delimit the building biases, it is helpful to focus correction efforts on a DEM that has already had other potential sources of error removed for these locations. For this reason, corrections were developed for MERIT, rather than for SRTM or TDM90 as these DEMs still suffer from a number of additional biases (vegetation artifacts, speckle and strip noise) over urban areas.

2.1. GDEM Error Analysis Using a Reference DTM

The vertical accuracy of the SRTM 1″ DEM (~30 m at the equator) void-filled version, TDM90 (~90 m at the equator), and the 3″ MERIT DEM (~90 m at the equator) was examined in this research. Error analysis of the SRTM 1″, MERIT, and TDM90 DEMs was conducted using cross-section profiles, error histograms, and statistical metrics. The ASTER DEM was not included in the analysis because of its significant systematic errors (Hirt et al., 2010). The selected GDEMs were analyzed by comparison with reference DTM (i.e., bare earth) data derived from airborne LIDAR or airborne very high-resolution camera surveying (Table 1).

These airborne data sets were taken as ground truth considering the small magnitude of their error. The reference DTM data in Beijing was collected in 2002 by the Beijing Institute of Surveying and Mapping, with very high-resolution airborne cameras. Initial data from Beijing was provided as a 1:10,000 contour map with 1 m contour intervals. The vertical error of this data is within 0.5 m. Contour lines were converted to a 5 m resolution raster (using the ArcGIS Topo to Raster tool). In the other six cities the reference DTMs were recent LIDAR DTM data at 1 m spatial resolution and with vertical RMSE less than 15 cm. The difference between GDEM elevation and reference DTM elevation will be referred as the GDEM error in this study. Error analysis was conducted in terms of the vertical reference system of the MERIT DEM (EGM96) at 3″ pixel
For the reference DTM data, offset values between the regional vertical reference system and the global geoid were applied. LIDAR DTM data in the United Kingdom was converted to EGM96 by implementing a 0.8 m offset (Ordnance Survey, 2018), 0.21 m was added to the DTM data in Beijing (Li et al., 2017), and 0.711 m was subtracted from LIDAR data in Berlin (Ihde et al., 1998). TDM90 was converted to the EGM96 vertical reference system using the National Oceanic and Atmospheric Administration VDatum software. For comparison, reference DTM data were resampled to the MERIT grid resolution by averaging, SRTM and TDM90 were resampled to match the same grid as MERIT using bilinear interpolation.

In this research, GDEM error analysis was conducted within city administration boundaries with water excluded using the TDM90 water indication mask (WAM). The TDM90 WAM was resampled to match the same grid as MERIT using the Nearest Neighbor method, and pixels identified as water coverage were excluded from the analysis. Apart from water, if nonurban land cover, such as forest, represented a significant percentage (>15%) of land within the city boundary, these parts were also excluded from the analysis. Land cover data were taken from Zhou et al. (2019) for Beijing and from the CORINE Urban Atlas 2012 database in the other cities.

### 2.2. Random Forest Method for Correcting MERIT DEM

Although most of the DEM error was clearly contributed by buildings in urban areas, the relation between buildings information and the DEM error is complex. This complexity is a result of the diverse interaction between the radar signal and the building objects. Benefiting from its bootstrap aggregating technique, Random Forest has the advantage of digesting information from numbers of potential factors to estimate the target without overfitting. It creates multiple diverse regression models by taking different samples of the original dataset, and then combines their outputs to estimate the target. Data left outside each tree's sample group is called the out-of-bag data (Breiman, 2001). Based on the out-of-bag data, Random Forest can measure the contribution of each factor as importance score. The importance score is defined as the total decrease in node impurity resulting from splitting the factor variables, averaged over all trees (Breiman, 2001), measured by the residual sum of squares. To calculate this, we used the Random Forest package of R. The intercomparison of importance score among factors indicates their relative significance for predicting the target.

In this research, the regression model was developed to estimate the vertical error in the MERIT DEM over built-up areas in the studied cities. For each city, the regression target – the difference between the MERIT DEM and the reference DTM – and the 13 regression factor layers (night-time lights, population density, building density, building height, MERIT DEM slope, and the 8 MERIT neighboring elevation values for each pixel) were computed to construct the dataset for the Random Forest model. Usually, the original dataset is divided into training and testing samples. The training samples are used to develop the regression
tree by feeding both the regression target and the regression factors into the model. With the obtained tree, predicted values of the target in the testing sample are generated using only the regression factors in the testing sample as input. The predicted values of the target in the testing sample are evaluated against the actual values of the target, which indicate the model’s ability in prediction. Following the above training and testing process, the complete dataset for each city was randomly divided into training (70%) and testing samples (30%). Two types of regression model were examined: single city mode and combined cities mode. These differed in the way the training sample datasets were constructed. In the single city mode, vertical errors in MERIT for the 30% testing sample for each city were predicted using a model trained using the 70% training sample from the same city. In the combined cities mode, error estimation for each target city (100% testing sample) was obtained using a model trained on the 70% training sample of other cities, but not the data from the target city itself. The second model aims to test the possibility of applying an urban bias correction for GDEM data where no ground truth data are available. The two modes are summarized in Table 2. In both modes, the predicted value was evaluated against the actual value using the coefficient of determination, $R^2$ (see Equations 1 and 2). This indicates the percentage of the variation in the predicted value that is explained by the regression factors.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_{i,\text{predicted}} - y_{i,\text{actual}})^2}{\sum_{i=1}^{n} (y_{i,\text{actual}} - y_{\text{mean}})^2}$$  \hspace{1cm} (1)$$

$$y_{\text{mean}} = \frac{\sum_{i=1}^{n} y_{i,\text{actual}}}{n}$$  \hspace{1cm} (2)$$

where $n$ is the number of samples, and $y_{i,\text{predicted}}$ and $y_{i,\text{actual}}$ are the predicted and actual MERIT errors, respectively.

Additionally, the importance score of the regression factors was analyzed to assess their contribution to the error estimation in the single city and combined city models. Finally, the predicted errors from the combined cities model were subtracted from the MERIT DEM to get a building corrected DEM of the target city which we termed MERIT-UC (MERIT-urban corrected).

### 2.3. Input Datasets for the Random Forest Regression

The vertical error in the MERIT DEM is termed the regression target and was computed by subtracting the reference DTM from the MERIT elevation values on a per-pixel basis. Globally available data sets selected as regression factors were night-time lights data, population data, building density, building height, and the MERIT DEM slope, elevation, and the eight neighboring MERIT elevation values for each pixel. Ideally, the acquisition time of the input factors should be the same as that of MERIT (i.e., the year 2000 when SRTM was acquired). However, at this time night-time lights data were only acquired at a much coarser resolution of 1 km (Baugh et al., 2010). For this reason, night-time lights (with 500 m resolution) acquired in 2015 by the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB) were used. As the night-time lights data was involved in the generation of population density, population density from 2015 was also used. Building footprints and height data represent the most recent available information at the time of data processing, which implicitly assumes that the studied cities did not experience widespread construction since 2000. It is also difficult to filter buildings built after 2000 due to the absence of construction-time information in OpenStreetMap data. Details of the data used in the regression are shown in Table 3 and the

| Mode            | Training sample                        | Testing sample                        |
|-----------------|----------------------------------------|---------------------------------------|
| Single city     | 70% Sample of the target city          | 30% Sample of the target city         |
| Combined cities | 70% Sample of other cities but not the target city | 100% Sample of the target city |
necessary preprocessing for these is discussed in supporting information Text S1. Note that the building information from OpenStreetMap has limited coverage and, moreover, the coverage varies markedly between cities (see Table S1). The regression model was only built over areas with both building density and building height values greater than 0.

### 2.4. DEM Validation

#### 2.4.1. MERIT-UC Error Analysis Using Reference DTM

Vertical error reduction was achieved by subtracting the estimated errors from the original MERIT DEM on a per-pixel basis. This was undertaken for the two modes of Random Forest model: single city mode and combined cities mode. As the final aim of the proposed Random Forest model is to estimate the error of MERIT where no reference DTM is available to train the model, the MERIT DEM corrected with predicted error from the combined cities model, that is, MERIT-UC, was further analyzed by cross-section profiles, error histograms, and the RMSE metric.

#### 2.4.2. Flooding Evaluation

An evaluation of the ability of the DEM data to predict flooding was undertaken for the city of Carlisle, a medium sized urban settlement in the United Kingdom. Severe flooding from the Rivers Eden, Caldew, and Petteril with an approximate return period of 1 in 150 years occurred in Carlisle over the January 6 and 7, 2005 and caused extensive urban inundation. Ground data collection in the immediate aftermath of the event yielded a set of 263 wrack and water mark elevations that have previously been used to validate urban inundation models (Neal et al., 2009). Vertical errors in the observed wrack and water mark data were suggested to be ~0.3–0.5 m (Fewtrell, Neal, et al., 2011; Horritt et al., 2010; Neal et al., 2009).

For this study, the 2005 flooding of Carlisle was modeled using the LISFLOOD-FP hydrodynamic model (Bates et al., 2010). Key model inputs for this hydrodynamic model include terrain, topography-derived bathymetric variables (channel width, bed elevation), boundary conditions, and surface friction (Bates et al., 2013). Four different hydrodynamic models were built using different terrain data sets. A benchmark model was constructed from a LIDAR DTM at 5 m resolution, along with three models built with coarser resolution GDEM data (MERIT, MERIT-UC, and TDM90). Because the urban bias in MERIT-UC was corrected only over built-up areas, a low pass 3-by-3 filter was used on this data set to smooth the transition between areas with and without the modification applied.

Considering the coarse resolution of the terrain input for the GDEM based models, a subgrid-scale representation of channelized flows was used in LISFLOOD-FP, which allows river channels with any width below that of the grid resolution to be simulated (Neal et al., 2012). The same channel was used in all four models to control for this important factor. Channel width was generated using the water surface top width derived from LIDAR survey, and bed elevation was generated from ground bathymetric survey (Horritt et al., 2010).

### Table 3

| Variable name                  | Generation technique/method                          | Data source                  | Source data resolution | Acquisition  |
|--------------------------------|------------------------------------------------------|------------------------------|------------------------|--------------|
| Night-time lights              | Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB) | VIIRS cloud free-outlier removed—night-time lights | 500 m                  | 2015         |
| Population density             | Random Forest method based on a variety of datasets  | WorldPop 2015                | 100 m                  | 2015         |
| Building density               | Building area from building footprint                | OpenStreetMap                | Polygon                | Update to 2019 |
| Building height                | Attribute table of building footprint                | OpenStreetMap                | Polygon                | Update to 2019 |
| Slope                          | ArcGIS slope calculate                               | MERIT                        | Same as MERIT          | Same as MERIT |
| Elevation and neighbor values  | DEM value and surrounding neighbors by 3 x 3         | MERIT                        | Same as MERIT          | Same as MERIT |
Flow boundary conditions based on gauging station data were taken from Neal, Bates, et al. (2009). To account for the sensitivity of the model to the unknown friction parameterization, a range of Manning’s friction parameters were used to calibrate the models. These were 0.01, 0.02, 0.03, 0.04, 0.05, 0.055, 0.06, 0.065, 0.07, and 0.08 for channel friction and 0.02–0.12 with 0.02 intervals for floodplain and urban areas friction. The ranges were made large enough to bracket the optimum model performance against the observed water mark data.

RMSE and mean error of water surface elevations were calculated within the range of frictions. The calculation was implemented by the method proposed by Neal et al. (2009) in which the water surface elevation of the nearest wet cell to the wrack or water mark is used when the simulated inundation extent does not overlap with the observations. The LIDAR based model was calibrated to find the smallest RMSE against the wrack and water mark data. The inundation extent given by this calibrated model was visually checked, and its coverage of most observed marks was confirmed. This is important because the nearest wet cell method can be less sensitive to the wrack and water marks at overly underestimated inundation extents. For the same reason, the inundation extent of the calibrated LIDAR model instead of the observed water mark data was then used to calibrate the three GDEMs models taking the highest Critical Success Index (Wing at al., 2017) to be the most accurate model parameterization. RMSE and mean error of the calibrated GDEM-based models were compared with that of MERIT-UC.

3. Results

3.1. GDEM Error Analysis in Urban Areas

Typically, cities include several types of land cover, including water, bare land, grassland, and trees. By superimposing the error map for each GDEM on top of land use and land cover data some common characteristics were found. Unsurprisingly, built-up areas mostly show positive errors (i.e., GDEM elevations are higher than the reference DTM). Figure 1 shows this information for Berlin as a typical example. Errors caused by buildings relate to both building height and building area; pixels with large building areas and high building heights tend to have larger positive errors. Some types of urban infrastructure with flat and smooth surfaces, such as stadiums and even some roads and building roofs, show negative errors in the GDEM. GDEM errors are mostly negative in areas of water coverage. Lastly, pixels with forestry coverage show positive errors where this bias has not been corrected (i.e., in SRTM and TDM90). This can be more than 10 m in some places, as shown in Berlin (Figure 1). As forest makes up more than 15% of the land area within the Berlin city boundary, it was excluded from the GDEM error histograms and statistical analysis for this city.

To represent the typical land cover of urban areas, transects crossing the downtown areas of each city were selected to show the profile of the three GDEMs and the reference DTM (Figure 2). These data show that SRTM, MERIT, and TDM90 are all generally higher than the reference DTM and that the GDEM error is up to 25 m. The GDEMs mostly have similar variance, although MERIT is visually smoother than the other GDEMs.

Error distribution analysis was conducted on SRTM, MERIT, and TDM90 with errors grouped into 1 m bins (Figure 3). In the studied cities, the majority of GDEM errors is positive and concentrates in the range of 0–10 m, except for Berlin where the SRTM error range is −5 to 5 m. MERIT has fewer negative errors than SRTM in all cities, likely because of its additional set of bias corrections. In Beijing, Manchester, and Cambridge, MERIT outperformed SRTM, mostly by reducing large errors (>5 m). In other cities, MERIT introduced more positive errors compared with SRTM (Figure 3), probably due to the process of absolute bias correction undertaken in the generation of MERIT. In Berlin, the absolute bias of SRTM was shown to be negative, which can cancel out part of the positive bias caused by buildings, so that the removal of this bias made MERIT more positively biased than SRTM. In London, Bristol, and Carlisle, absolute bias might be overestimated as the benchmark ICESAT data used to estimate the absolute bias in MERIT contains very limited points and these are unevenly distributed (Yamazaki et al., 2017). In the capital cities (Beijing, Berlin, London) TDM90’s most common error (the highest frequency error-peak) is similar to that of MERIT at 2–4 m, and is 1 m lower than that of MERIT in the other cities. This suggests that the TDM90 DEM shows advantages over SRTM/MERIT in less densely built-up areas. In addition, TDM90 has more small errors
(<2 m) than the other GDEMs except in Beijing and Berlin. However, for TDM90 those peak errors values are all (except for Carlisle) higher than the peak error values of urban areas in floodplains (<1 m) reported by Hawker et al. (2019).

Figure 3 also indicates the variability of DEM errors by showing the frequency of GDEM peak error across the studied cities, which is nearly 30% in most cities (given 1 m bins), lower in Berlin (~20%), and higher in Cambridge (~35%). This demonstrates that the DEM error in Cambridge is the most centralized around the mean error, while that of Berlin is the most dispersed for the cities studied. This relates to the fact that the building heights in Cambridge are more uniform within the city boundary than at other locations. GDEM error statistics are given in Table 4.

Table 4 shows that the RMSE of GDEM error in the studied cities ranges from 2.31 to 5.98 m. Generally, the RMSE is close to that reported by Hawker et al. (2019) for the same three DEMs for urban floodplains.
In terms of RMSE, the vertical accuracy of GDEMs varied among the cities examined here. Among UK cities, TDM90 has lower RMSE than MERIT, and is the DEM with the best vertical accuracy in all UK cities except Bristol and London, where SRTM has the lowest error. In Beijing and Berlin, MERIT is better than TDM90. However, the best accuracy in Berlin was achieved by SRTM in contrast to Beijing where MERIT outperformed SRTM. The lower RMSE of TDM90 seems to be a consequence of the greater proportion of small errors than in the other two GDEMs, which can be seen in Figure 3. In addition, there is a reduction in RMSE from capital cities to small cities in TDM90 and MERIT, but this pattern is not so clear for SRTM. The standard deviation of MERIT is generally lower than TDM90 and SRTM. Moreover, the standard deviation among cities evidenced the more centralized error distribution in Cambridge and more dispersed distribution in Berlin as shown in Figure 3.
In terms of mean, absolute mean, median, and absolute median errors, the intercompared rank of GDEM vertical accuracy in each city is the same as obtained when using RMSE, and these values for SRTM in Berlin, Carlisle, and Bristol are much lower than those of other cities. Additionally, the impact of the pixel size on the error analysis was checked (Uuemaa et al., 2020), and we found that would not alter our conclusions.

As this research focuses only on the correction of building bias SRTM is not suitable for correction as it is also subject to a number of other confounding errors (Yamazaki et al., 2017). The finer collection resolution of TDM90 (12 m) enables more radar signal to penetrate to the ground than MERIT, which contributes to the better performance of TDM90 compared to MERIT in some cities. However, this also caused a more random pattern of building bias. For a correction algorithm, a systematic pattern of building bias is preferable even when the absolute error magnitude is slightly higher. Random pattern bias might be better addressed at source in the 12 m TanDEM-X data (e.g., the morphological filter suggested by Geiß et al. [2019]) but this

Figure 3. Error frequency of SRTM, MERIT, and TDM90 in seven cities. Error was calculated by subtracting the DTM from each DEM. For clarity, errors exceeding the range −5 to 20 m were excluded in this figure as the overall frequency of these errors is very small (0.24% on average).
is a commercial product and therefore unavailable for this work. For these reasons, the MERIT DEM was selected for urban bias correction in this study.

3.2. MERIT Error Estimation

3.2.1. Error Estimation for Single City Regression Models

Figure 4 shows the predicted error against actual error for the 30% testing sample in each city. Error was predicted by the model trained on the 70% training sample for the same city. For example, the plot for Beijing shows the estimated error against the actual error for the 2,130 pixels in the testing sample, with the error estimated by a model trained on the other 70% of the Beijing dataset (4,797 pixels). The number of samples for each city can be found in Table S1.

The $R^2$ values vary in the range of 0.45–0.75 in the seven cities (Figure 4). This indicates that the input factors can, in all but two cases, explain most of the variance in the target. It can also be seen that there is a tendency for errors smaller than 2.5 m to be overestimated and for large errors (>5 m) to be underestimated. This is possibly caused by the fitting strategy of the regression tree method in which the estimated value was achieved by averaging results from all trees.
Factor importance analysis (Figure 5) showed that night-time lights data was the key factor for predicting MERIT error in all the studied cities, although its importance score varied among them. In Beijing and Berlin, night-time lights and elevation are the most significant factors. For most UK cities night-time lights and building density play big roles, although these two factors have almost identical importance in Cambridge, differentiating it from the other UK cities. This could be explained by the fact that buildings are the main land cover in Cambridge. This land cover character was reflected by the high ratio of the building density layer to the whole area studied (87%, see Table S1) together with the vertical error statistics (Table 4) in Cambridge (higher proportion of small errors (2–5 m), higher RMSE, mean and median values but much lower standard deviation compared to other cities). In Carlisle, night-time lights and building height are the main explanatory factors, but they are not significantly more important than other factors.

3.2.2. Error Estimation for Combined Cities Regression Models

In each city, MERIT error estimation was also conducted using a combined model trained on data from the other cities (i.e., excluding the training data from the target city itself). This analysis tests the performance of the building bias correction for situations where benchmark elevation data may not be available. In Random Forest modeling, the representativeness of the input factor data impacts the model’s transferability (Sinha et al., 2019). It follows that when the relationship between prediction factors and building bias is assumed as the same or similar across cities, then a wide span of factors and bias would be more transferable than a narrow span of data. Therefore, a very preliminary grouping strategy was used in these multiple city models. To ensure enough training data, all other cities (except the target city) were included in training the model unless the structure of the target city was significantly different from the rest of them. Because of Carlisle’s rather different city structure (as seen from the low RMSE in Table 4), it was differentiated from the other cities. Its trained model was therefore constructed using data only from Manchester, Bristol, and Cambridge, whose city structures are a relatively close match to that of Carlisle. Trained models for the other target cities were generated using the training samples from all other cities excluding the target and Carlisle. Figure 6 shows the predicted error against actual error for all samples in each city. Different from Figure 4, these errors were predicted by the model trained on 70% sample of training data from all other cities except Beijing and Carlisle. The factor importance of the trained model for each target city is shown in Figure 7.

In terms of R², these results showed that only Manchester obtained a positive value, while in the other cities the R² was negative, meaning that MERIT-UC failed to surpass the effectiveness of subtracting a uniform mean error for all pixels. Results deviate more from the perfect fit line than the regression models trained on single cities shown in Figure 4. In all cities, small errors (0–5 m) were overestimated, and large errors (>10 m) were underestimated. This was especially true in Berlin and London. Predicted error was concentrated between 0 and 10 m, though actual error had a large range up to 30 m in some cities. The averaging strategy at the heart of the regression tree method, means target biases in the middle of the bias range tend to be better estimated than those at the edge.

Figure 4. The predicted error by Random Forest regression in single city mode against actual error in seven cities. The black dots represent the 30% sample of that city (testing), and the y-axis represents the predicted error generated based on the model trained by the 70% sample of each city. A very small number of outliers (error <−10 m) was excluded in Beijing and Berlin (less than 0.1%) to improve clarity.
In the regression models built using training data from combinations of nontarget cities night-time lights, building density, population, and the elevation of the target pixel are all significant factors (score >10%). As the importance score was measured based on a subset of the training data (i.e., out-of-bag data), this resulted in more balanced importance scores than the single city scores. Due to the different size of each

Figure 5. Importance scores of input factors in seven cities computed by the Random Forest model in single city mode. Scores were produced based on the out-of-bag data of the 70% training data of each city. NTL, night-time lights; POP, population density; BD, building density; BH, building height; SL, slope; ELE, elevation; N1-N8, neighbor elevation values of target pixel in 3 × 3 windows.
city, larger cities with more samples involved in the combined models can leverage the scores toward the importance achieved in its single city training. For example, building density scores are larger than 10% in all combined models with London samples used in the training because building density is shown as an important variable in the single city model of London (Figure 5). The same can be seen for the elevation score for Berlin. This importance score could change when training cities are weighted by their sample size.

3.3. Ground Truth Validation

The RMSE of MERIT before and after subtracting the estimated building biases from the two types of Random Forest model are shown in Table 5. Values represent the MERIT RMSE of test samples in each case. There is a significant RMSE reduction for MERIT in most cities produced by the single city model. The RMSE of MERIT decreased by up to 76% compared to that before building correction and resulted in a much lower RMSE of 1.08–2.05 m. For regression models trained with data from combinations of other cities, the RMSE decrease for MERIT is 15%–67% compared to that before correction. These values are smaller than those achieved by the models trained using the target cities’ own data.

It should be clarified that the slight difference in the original RMSE for the two models in Table 5 is due to the different number of pixels involved in these two calculations. In the single city models, 30% of the data is used in testing, whilst in the combined models 100% of the target city data are used in the testing step. The values are both higher than the RMSEs given in Table 4 because only pixels with built-up coverage were included in the regression data and these tend to have larger errors.

We generated MERIT-UC by subtracting the estimated error determined from the combined cities model from the original MERIT DEM. The profiles of MERIT-UC in Figure 8 generally show lower values compared to the original MERIT DEM and move closer to the reference DTM. However, overestimations of built-up errors were shown in MERIT-UC in some cities, especially in Beijing. For Beijing, this was probably caused by the mismatch in acquisition time between prediction variables and the building biases. The former data were mostly collected within the last 5 years, while the DEM data was collected over 20 years ago. Overestimation is also relevant to the model’s transferability between training cities and the target city, and is a bigger issue in small sized cities, such as Carlisle, Bristol, and Cambridge than in London and Berlin.

Error histograms (Figure 9) show the lower elevation values in MERIT-UC compared with MERIT, with errors typically in the range -5–5 m. The RMSE of MERIT-UC is the lowest compared with the other GDEMs, though before this correction MERIT DEM has higher RMSE than TDM90 in most cities. The percentage of errors within 2 m was improved from 23% in MERIT to 82% in MERIT-UC, making most errors within 2 m. However, this correction also introduced some negative errors, especially in Beijing and Bristol, although the overall error has been reduced.

### 3.4. Flooding Evaluation for Carlisle, UK

RMSE and mean error were evaluated based on postevent surveyed water and wrack mark observations as discussed above. These observed data should be considered to have likely ±1 cm error caused by the

Figure 6. The predicted error by Random Forest regression in combined cities mode against actual error in seven cities. The black dots represent all the data of the target city (testing), and the y-axis represents the predicted error generated based on the model trained by the overall 70% data of other cities but not the target city. A very small number of outliers (error <-10 m) was excluded in Beijing and Berlin (~0.02%) to improve clarity.
GPS plus additional errors relating to the interpretation of the water and wrack marks in the field to give a total error of ±30–50 cm (Fewtrell, Neal, et al., 2011). The smallest RMSE achieved by the 5 m LIDAR DTM model was 0.25 m, while mean error was 0.05 m. The RMSE obtained with the 5 m LIDAR model is 0.07 and 0.03 m less respectively than that of the 25 m LIDAR DEM and DTM models of Neal et al. (2009). The LIDAR model therefore matches the observed maximum water level as well as can be expected given

Figure 7. Importance score of factors from combined models for each city. Scores were produced based on the out-of-bag data of the overall 70% data of other cities but not the target city. Subtitle shows the names of the cities whose data were used in training the model.
the observation errors. This result was obtained at a channel friction of 0.06 and floodplain friction of 0.06 (i.e., a calibrated model). The channel friction value is a little larger than might be expected and is probably compensating for errors in either the gauged discharge used as boundary condition information or in the estimated channel bathymetry. As the error was small, and broadly equivalent to the uncertainty of the ground measurement of maximum water level obtained immediately postevent, the flooding extent from this calibrated 5 m LIDAR model was considered as a benchmark.

For the MERIT, MERIT-UC, and TDM90 based models, the RMSE and mean error calculated from the water and wrack mark observations are shown in Figure 10. Although the TDM90 based model achieved the lowest RMSE and mean error, both metrics were reduced in the MERIT-UC flood model compared with the MERIT DEM based model, especially at large channel frictions (>0.04). Specifically, RMSE and mean error at best fit flood extent (i.e., optimally calibrated models) were examined. Flood extent and water depth value maps of the calibrated models can be seen in Figure 11. In the calibrated models, optimal channel friction is 0.08 for MERIT, and 0.065 for MERIT-UC and TDM90, and floodplain friction is 0.06 for MERIT, and 0.1 for MERIT-UC and TDM90, respectively. Calibrated friction parameters for coarse resolution DEM models are higher than those of fine-resolution building-resolving LIDAR DTM models because the friction term has to parameterize the effect of blockages by buildings and also account for the GDEM vertical biases discussed below. In this case, the MERIT-UC DEM based model can simulate a more accurate result than MERIT with RMSE reducing from 2.79 to 2.29 m, and mean error reducing from 2.74 to 2.04 m. The TDM90 based model achieved an even better result than MERIT-UC, with an RMSE of 1.71 m and a mean error of 1.65 m. These water surface errors are substantially larger than those of the LIDAR simulation. An open ground examination of GDEM vertical error (75 samples manually selected at flat, grass covered pixels, slope <2° in the flood modeling area) showed that both MERIT and TDM90 DEM are subject to vertical biases at this site. The mean vertical error of these samples is 1.75 m (with standard deviation at 0.69 m) for MERIT and 0.31 m (with standard deviation at 0.45 m) for TDM90. Some of these errors are likely due to errors in the reference system conversion between EGM96 and OS Datum, it is clear that MERIT is positively biased by a larger number than TDM90 in Carlisle. This bias is beyond the capability of the building bias correction model, and an almost equal bias to MERIT was shown of MERIT-UC at 1.67 m. In addition, the building bias correction was applied to only around 15% of the city due to limited coverage of building footprint data in Carlisle, and an even smaller part (9%) of the area covered by the flood model. Thus, it is unsurprising that the simulation of MERIT-UC was no better than TDM90 in this case.

With friction calibration, all three GDEM models (MERIT, MERIT-UC, TDM90) can simulate a similar inundation extent to the LIDAR based model, with CSIs all around 0.75. Some areas of incorrectly predicted inundation extent were shown in all three models. For example, underestimation is evident in areas nearby the B6264 road (red circle labels in Figure 11). Examination of the elevation of MERIT, MERIT-UC, and TDM90 in this area shows that they are all higher than the ground elevation of the LIDAR DTM.

| City        | rmse_ori (m) | rmse_new (m) | Reduced (%) | rmse_ori (m) | rmse_new (m) | Reduced (%) |
|-------------|--------------|--------------|-------------|--------------|--------------|-------------|
| Beijing     | dsb          | 1.51         | 72          | 5.25         | 3            | 43          |
| Berlin      |              | 5.81         | 72          | 5.8          | 3.87         | 33          |
| London      |              | 6.44         | 71          | 6.58         | 5.6          | 15          |
| Manchester  |              | 4.5          | 68          | 4.41         | 2.82         | 36          |
| Bristol     |              | 4.75         | 57          | 4.65         | 3.28         | 29          |
| Cambridge   |              | 4.52         | 76          | 4.43         | 1.48         | 67          |
| Carlisle    |              | 3.45         | 61          | 3.35         | 2.38         | 29          |

Note. rmse_ori is the RMSE of MERIT DEM, rmse_new is the RMSE of MERIT DEM after building bias correction. Values are from test data in each model.
Overestimation (dashed-orange label in Figure 11) is seen at the top of the flooded area. Misprediction here is likely caused by the calibrated friction parameters which achieve an optimum extent for the whole model at the expense of an over-extension of predicted inundated area at this location.

In built up areas in Carlisle, both the MERIT-UC and TDX90 based models performed better than the MERIT model (dashed-green labels in Figure 11). For example, around the river Caldew, inundation from the MERIT-UC model reached the extent of the LIDAR model, although it failed to resolve many of the inundation gaps. However, inundation extent in MERIT-UC is not better than MERIT everywhere. For example, overestimation of inundation in the MERIT-UC model can be seen at the corner of the A7 road (yellow circle label in Figure 11) where some buildings are situated. This relates to the fact that MERIT-UC corrected MERIT over only 9% of the area of the flood model, but the inundation results depend on not only the...
Some of the overestimated vertical error in the Random Forest model and the inconsistency between the corrected and uncorrected pixels might have made MERIT-UC slightly noisier compared to MERIT. This might impact the flooding simulation, although the low-pass filter applied to MERIT-UC attenuated the impact to some extent. Moreover, it should be noted that the subgrid channel representation in all models was taken from the LIDAR data and channel cross-section data in order to control for this important aspect in the experimental design. Where MERIT-UC alone is used for flood modeling, careful attention to channel characterization is therefore recommended.

Figure 9. Histogram and RMSEs for MERIT, MERIT-UC, and TDM90 in regression areas of seven cities. Regression areas are defined as areas with both building density and building height greater than 0. *Errors exceeding the range (−10 to 20 m) were excluded in this figure and from the RMSE calculations.

The elevation of the pixel itself but also that of the pixels surrounding it, meaning that uncorrected pixels can cause incorrect inundation in areas beyond their immediate vicinity.
4. Discussion

Among the cities studied, the GDEM with the best vertical accuracy varied considerably. TDM90 has the best vertical accuracy in three out of seven cities (Manchester, Cambridge, and Carlisle), which appears to be a result of little building construction happening between the time when SRTM data were collected (2000) and TDM90 data acquisition (2015). This was confirmed by checking the land cover time series for these cities. Whilst this may appear counter-intuitive, for applications of GDEMs requiring bare-earth data in urban areas it is better to have larger areas of building free land as then fewer building artifacts will need to be corrected. Given a similar area of building free land in 2000 and 2015, the higher native resolution of TDM90 tends to lead to a better result. However, for urban areas like Beijing, which have experienced significant construction activity since 2000, MERIT and MERIT-UC can be better than TDM90 because they automatically see more of predevelopment bare earth. Furthermore, TDM90 still suffers from significant vegetation biases, whilst MERIT successfully reduces this error (as seen in Berlin in Figure 1).

Figure 10. RMSE (left column) and mean error (right column) of predicted water surface elevation for: MERIT, MERIT-UC, and TDM90 based simulations at 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.065, 0.07, 0.08 channel frictions and 0.02–0.12 floodplain frictions with 0.02 interval.
Although multiple errors have been removed in MERIT, it does not always have better vertical accuracy than SRTM. This relates to the estimation of absolute error when generating MERIT, which was determined using nonforest ICESAT centroid elevations as a benchmark (Yamazaki et al., 2017). This means the absolute error might be influenced by building artifacts when the ICESAT data fall into built-up areas. In this study, overestimation of absolute error was shown in some cities, such as Carlisle. The limited number of ICESAT data at some locations might also cause uncertainty in the absolute error estimation. This might be reduced by excluding densely built-up areas from the absolute error estimation process and using a richer data set such as ICESAT-2. In other cases, if the absolute error of SRTM is already negative it can cancel out the positive bias caused by buildings. The removal of negative absolute errors in the creation of MERIT can turn the negative biased SRTM into a positive biased MERIT with larger RMSE (as shown in Berlin, Figure 3, Table 1). However, we would argue that the removal of absolute errors is a necessary step in the production of DEMs for flood inundation modeling.

Therefore, when there is a need for a free access DEM, the extent and type of vegetation and the history of urban construction activities might help in choosing the one with better vertical accuracy, if local validation data are unavailable. For global scale DEMs, both MERIT and TDM90 have the potential to produce a global consistent bare-earth DEM with building height biases being removed using the methods we have outlined. Although TDM90 has an advantage in terms of the native resolution of the underlying data, multiple, and as yet uncorrected, errors (e.g., vegetation bias, phase unwrapping errors) (Rizzoli et al., 2017) complicate the generation of an urban bare-earth DEM from this source. These errors have been removed/largely reduced in MERIT, making MERIT the obvious choice for urban bias correction at the present time. However, the urban bias correction of the TDM90 is certainly worthy of study in the future once other errors have been dealt with.

Figure 11. Maximum water depth for LIDAR and GDEM models at optimal calibrations (color scale is the same for all figures and is shown in the upper left panel).
The performance of the Random Forest regression was impacted by inconsistencies between the prediction variables and building biases. First, the resolution of some prediction variables is not consistent with the scale of the building biases. For example, the night-time lights variable is of a much coarser resolution (500 m) than the ~90 m of MERIT, which distorted the relationship between this variable and the building biases. This distortion especially affected the factor-bias relationship in areas with sparsely distributed buildings. This could partly explain the lower $R^2$ value of small cities, such as Carlisle, in the regression model. Second, the acquisition time between the prediction factors and the vertical errors can be inconsistent. SRTM was acquired over 20 years ago, whilst building height data has only become available much more recently. Over this time period, cities like Beijing have experienced rapid urban expansion. This can cause a misplaced correspondence between the input variables and the vertical error and devalues the importance of building information factors in the regression modeling, whereas these should, in reality, be of significance.

Clearly, the ability of the proposed model to estimate vertical errors caused by building artifacts weakened when samples from the target city were not used in model training. In such cases, the Random Forest model’s ability is less impacted when estimating a less dispersed target (i.e., the error of MERIT). We found that the standard deviation of MERIT error showed a clear negative impact on the model’s transferability (see Figure S2). For example, the standard deviation of MERIT error in Cambridge is the lowest at 1.43 m and its error was reduced by 67% using a model trained on data from other cities. This is the largest error reduction among the cities studied. The vertical error caused by building artifacts might be characterized by a given city’s elevation, size, history of development as well as other aspects like building character and street patterns, all of which display significant variability worldwide. Although the universally available night-time lights data play a significant role in estimating building bias, the negative $R^2$ value obtained when transferring a model built for one set of cities to other locations indicates that the above differences cannot be neglected in the modeling. The negative $R^2$ values might also indicate that, for now, simply subtracting a uniform mean value of the estimated building bias for each city may be a better way of correcting these errors. Among the regression variables of the Random Forest model, night-time lights, population density, building density, and building height are likely to have a positive relationship to the building biases, whereas that of slope, elevation and neighboring elevation values are less clear. Because the Random Forest method develops trees by taking subsets of predictors, we speculate that the elevation and elevation related factors (slope and neighboring elevations) have a location-sensitive, complex relationship with the vertical error of MERIT which cannot be replicated with training data from other locations. Grouping these factors by values and normalization might help to enhance the model’s prediction ability in the future. The model might benefit from removing less important scored factors. The possible outstanding absolute bias of MERIT mentioned above might also impact the transferability as well. From the perspective of city size, biases tend to be larger in big cities than in small cities, which makes the former more forgiving locations in terms of overestimation than small cities, where smaller biases may be widely spread. We speculate that bigger cities are more tolerant to the selection of data used in correction model training. For a global scale correction, more details about the sensitivity of the relationship between the predictors and the vertical height error are needed and should be examined in future work. Although a preliminary grouping strategy for choosing the locations of training samples was used, a quantified, widely applicable grouping strategy is not yet possible with the current studied cities. Including more diverse cities and adequate building height data when available are essential to proposing such a grouping strategy. By that grouping, the current transferability issue could possibly be ameliorated.

Despite the multiple aspects affecting the regression model and its relatively weak transferability between cities, the proposed method reduced building bias effectively, with MERIT-UC showing the lowest RMSE among all studied GDEMs in areas where the correction was applied. It should be noted that when the evaluation of MERIT-UC was extended beyond the correction applied areas (which cover only 7%–26% of the whole city) to the whole city, the RMSE of MERIT-UC is not always the lowest when compared to SRTM and TDM90. This can be ascribed to the limited amount of building information data (mainly building height) available from OpenStreetMap (see percentage data in Table S1). With 7%–26% of the city area corrected (i.e., corrections only in the built-up areas where building data are available from OpenStreetMap), the RMSE of MERIT-UC over the whole city reduced by about 4%–8%. This still generates the most accurate DEM over the whole city for Beijing, Berlin, and London. However, in Carlisle, the RMSE of MERIT-UC
over the whole city is 2.77 m, larger than that of TDM90 (2.31 m), even though the RMSE in corrected areas for Carlisle was reduced from 3.35 m in MERIT to 2.38 m in MERIT-UC (see Table S1 for overall RMSE of MERIT-UC at whole city scale). This explains the better performance of TDM90 over MERIT-UC for flood simulation in Carlisle. In the future, when more datasets containing building information become available, for example 3D building volume (Geiß et al., 2019), a global bare-earth DEM could possibly be improved by including these datasets into a regression model. Hence, better flood mapping globally can be expected in the future depending on the emergence of finer and more complete building footprint and height data.

5. Conclusions

This study analyzed the vertical error characteristics in urban areas of three popular GDEMs and inter-compared them in seven cities. To our knowledge, it provides the first general analysis of GDEM errors in urban areas and addresses a widespread need to reduce GDEM errors across different cities. Bias caused by urban infrastructure has impeded the application of GDEMs where bare-earth terrain is required, for example for flooding simulations at large domain scales where fine resolution, building resolving models are not yet computationally feasible. To generate a bare-earth DEM from currently available GDEMs, we built a regression model using the Random Forest method and used it to correct the MERIT DEM to get a bare-earth DEM in urban areas. The corrected MERIT-UC DEM was further evaluated in terms of its ability to simulate a major flooding event in the city of Carlisle, UK.

In the studied urban areas, the RMSE against benchmark LIDAR data of the uncorrected GDEMs (SRTM 1°, MERIT DEM 3°, TDM90) was in the range of 2.31–5.98 m. There is not a single GDEM that is better than the others for all cities. The uncorrected GDEM that aligned best with the reference DTMs was TDM90 in three UK cities (Manchester, Cambridge, and Carlisle), MERIT in Beijing, and SRTM in Berlin and Bristol. All uncorrected DEMs achieved similar accuracy in London. TDM90 was found to be worse than the MERIT DEM in Beijing, because of the significant construction activity experienced in that city since the year 2000, and in Berlin, which has a significant coverage of forest (>15%) within the city. The magnitude of the urban development since the SRTM data collection and the extent of woodland/forest cover within urban areas should therefore be considered when choosing an appropriate GDEM.

The proposed urban bias correction method is effective in removing building artifacts. The RMSE of MERIT decreased by more than 50% in all cities when samples from the target city were used to train the regression model. For generalized applications, the RMSE of MERIT over the considered urban areas can be reduced by 15%–67% even when the target city’s data is not included in training. Thus, the transferability of the regression model is substantially less effective than for the single city case. A preliminary analysis found that large sized urban settlements seemed to be less affected by the transferability issue. However, model transferability needs to be much more deeply explored in the future. Currently, the correction is limited by the area of the city with building information in OpenStreetMap, which is incomplete in many places and especially sparse in small sized cities. On the other hand, building information in OpenStreetMap is growing over time, and thus the data set is still a promising resource with which to correct building biases in GDEMs. Our proposed method might also be used in conjunction with other developing building information sources, such as global 3D building volumes, as these become available in the future.

Finally, a test of the ability of the different GDEMs to simulate flooding in central areas of Carlisle, a medium sized UK city, showed that the MERIT-UC DEM performed better than MERIT, with the RMSE of predicted inundation depths reducing from 2.79 to 2.29 m for calibrated models. However, due to the outstanding absolute bias of MERIT and the limited areas adjusted by the urban bias correction in the case of Carlisle, MERIT-UC did not exceed the flooding performance of TDM90 in this case.

In summary, this work has, for the first time, characterized the urban height biases in freely available GDEM data and found them to be significant. Regression models developed to estimate this bias using benchmark airborne LIDAR as training data and globally available factors including night-time lights, population density, terrain slope and building footprint, and height information were shown to be successful where the regression was trained on data from the target city. However, it proved more difficult to transfer these relationships between cities and predict the building bias using data only from other locations. More work therefore needs to be undertaken to better understand how the nature of building height biases varies...
with the nature and history of city development. City landscapes and architecture vary considerably from place to place, and this currently impacts our ability to fully generalize building bias correction methods for global application. Nevertheless, this work does suggest some promising directions for future research that could potentially begin to address these issues.

**Data Availability Statement**

The MERIT-UC and other data used can be downloaded at the University of Bristol data repository, at https://data.bris.ac.uk/data/dataset/m1pnui7m717tl2trjbcp7t7le. Code is available at https://github.com/YinxueLiu/MERIT-Urban-Correction-DEM. MERIT DEM is free access for research and education purposes at http://hydro.iis.u-tokyo.ac.jp/~yamadai/MERIT_DEM/. SRTM DEM 1" can be downloaded from https://earthexplorer.usgs.gov/. TanDEM-X DEM 3" can be downloaded from https://download.geoservice.dlr.de/TDM90/#etdatais. LIDAR data of UK cities are available from Digimap at https://digimap.edina.ac.uk/roam/download/lidar, of Berlin is available from https://fbinter.stadt-berlin.de/fb/index.jsp. Landcover data of Europe cities can be found at Copernicus Urban Atlas at https://land.copernicus.eu/local/urban-atlas/urban-atlas-2012, and of Beijing is referred from Zhou et al. (2019) https://doi.org/10.1016/j.ufug.2017.12.010. Night-time lights data is available from https://www.ngdc.noaa.gov/eog/viirs/download_download_dnb_composites.html. Building footprints and building heights are retrieved from OpenStreetMap at http://download.geofabrik.de/, https://overpass-turbo.eu/, respectively, whereas building height data for Carlisle is available at https://digimap.edina.ac.uk/roam/download/os. Population data is available at https://www.worldpop.org/geodata/listing/?id=29. LISFLOOD-FP model is available for noncommercial research purposes by contacting Paul Bates at University of Bristol (http://www.bristol.ac.uk/geography/research/hydrology/models/lisflood/). Water surface elevation and boundary conditions for Carlisle are available in Neal, Bates, et al. (2009) at https://doi.org/10.1016/j.jhydrol.2009.01.026. Datum conversion from WGS84 ellipsoid to EGM96 was done using VDatum software available at https://vdatum.noaa.gov/. The Random Forest model is implemented based on the “randomForest” package in R (https://cran.r-project.org/web/packages/randomForest/randomForest.pdf).

**Acknowledgments**

Yinxue Liu was supported by the China-Scholarship-Council (CSC) – University of Bristol Joint PhD Scholarships Program. Paul Bates was supported by a Royal Society Wolfson Research Merit award. Jeffrey Neal was supported by NERCNE/S006079/1. Dai Yamazaki was supported by JSPS Kakenhi 16H06291. The authors would like to thank the anonymous reviewers and the editors for offering comments to improve this study.

**References**

Archer, L., Neal, J. C., Bates, P. D., & House, J. I. (2018). Comparing TanDEM-X data with frequently used DEMs for flood inundation modeling. *Water Resources Research*, 54(12), 10–205. https://doi.org/10.1029/2018wr023688

Bates, P. D., Horritt, M. S., & Fewtrell, T. J. (2010). A simple inertial formulation of the shallow water equations for efficient two-dimensional flood inundation modelling. *Journal of Hydrology*, 387(1–2), 33–45. https://doi.org/10.1016/j.jhydrol.2010.03.027

Bates, P. D., Trigg, M., Neal, J., & Dabrowa, A. (2013). LISFLOOD-FP user manual: Version 2.0. Bristol: University of Bristol. Retrieved from https://www.bristol.ac.uk/media-library/sites/geography/migrated/documents/lisflood-manual-v5.9.6.pdf

Baugh, C. A., Bates, P. D., Schumann, G., & Trigg, M. A. (2013). SRTM vegetation removal and hydrodynamic modeling accuracy. *Water Resources Research*, 49(9), 5276–5289. https://doi.org/10.1002/wrcr.20412

Baugh, K., Elvidge, C. D., Ghosh, T., & Ziskin, D. (2010). Development of a 2009 stable lights product using DMSP-OLS data. *APAN Proceedings*, 39(0), 114. https://doi.org/10.7125/apan.30.17

Breiman, L. (2001). Random forests. *Machine Learning*, 4(5), 1–32. https://doi.org/10.1023/a:1010933404324

Chen, H., Liang, Q., Liu, Y., & Xie, S. (2018). Hydraulic correction method (HCM) to enhance the efficiency of SRTM DEM in flood modeling. *Journal of Hydrology*, 559, 56–70. https://doi.org/10.1016/j.jhydrol.2018.01.056

Cooper, H. M., Zhang, C., Davis, S. E., & Troxler, T. G. (2019). Object-based correction of LiDAR DEMs using RTK-GPS data and machine learning modeling in the coastal Everglades. *Environmental Modelling & Software*, 112, 179–191. https://doi.org/10.1016/j.envsoft.2018.11.003

Elvidge, C., Tuttle, B., Sutton, P., Baugh, K., Howard, A., Milesi, C., et al. (2007). Global distribution and density of constructed impervious surfaces. *Sensors*, 7(9), 1962–1979. https://doi.org/10.3390/s7091962

Fan, H., Zifp, A., Fu, Q., & Neis, P. (2014). Quality assessment for building footprints data on OpenStreetMap. *International Journal of Geographical Information Science*, 28(4), 700–719. https://doi.org/10.1080/13658816.2013.867495

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., et al. (2007). The shuttle radar topography mission. *Reviews of Geophysics*, 45(2), https://doi.org/10.1029/2005rg000813

Fewtrell, T. J., Duncan, A., Sampson, C. C., Neal, J. C., & Bates, P. D. (2011). Benchmarking urban flood models of varying complexity and scale using high resolution terrestrial LiDAR data. *Physics and Chemistry of the Earth, Parts A/B/C*, 36(7–8), 281–291. https://doi.org/10.1016/j.pce.2010.12.011

Fewtrell, T. J., Neal, J. C., Bates, P. D., & Harrison, P. J. (2011). Geometric and structural river channel complexity and the prediction of urban inundation. *Hydrological Processes*, 25, 1747–1758. https://doi.org/10.1002/hyp.8035

Ford, A., Barr, S., Dawson, R., Virgo, J., Batty, M., & Hall, J. (2019). A multi-scale urban integrated assessment framework for climate change studies: A flooding application. *Computers, Environment and Urban Systems*, 75, 225–243. https://doi.org/10.1016/j.compenvurbsys.2019.02.005

Gallant, J. (2011). Adaptive smoothing for noisy DEMs. *Geomorphometry*, 2011, 7–9.

Gallant, J. C., & Read, A. (2009). Enhancing the SRTM data for Australia. *Proceedings of Geomorphometry*, 31, 149–154.
Gamba, P., Dell Acqua, F., & Housham, B. (2002). SRTM data characterization in urban areas. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 34(3/B1), 55–58.

Geß, C., Leichtle, T., Wurmm, M., Pelizari, P. A., Standfuß, I., Zhu, X. X., et al. (2019). Large-area characterization of urban morphology—mapping of built-up height and density using TanDEM-X and Sentinel-2 data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12(8), 2912–2927. https://doi.org/10.1109/jstarse.2019.291259

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets. Environment and Planning B Planning and Design, 37(4), 682–703. https://doi.org/10.1068/b35097

Hawker, L., Neal, J., & Bates, P. (2019). Accuracy assessment of the TanDEM-X 90 Digital Elevation Model for selected floodplain sites. Remote Sensing of Environment, 232, 111319. https://doi.org/10.1016/j.rse.2019.111319

Hawker, L., Rougier, J., Neal, J., Bates, P., Archer, L., & Yamazaki, D. (2018). Implications of simulating global digital elevation models for flood inundation studies. Water Resources Research, 54(10), 7910–7928. https://doi.org/10.1002/2018wr023279

Hecht, R., Kunze, C., & Hahmann, S. (2013). Measuring completeness of building footprints in OpenStreetMap over space and time. ISPRS International Journal of Geo-Information, 2(4), 1066–1091. https://doi.org/10.3390/iji2041066

Hirt, C. (2018). Artefact detection in global digital elevation models (DEM): The maximum slope approach and its application for complete screening of the SRTM v4.1 and MERIT DEMs. Remote Sensing of Environment, 207, 27–41. https://doi.org/10.1016/j.rse.2017.12.037

Hirt, C., Filmer, M. S., & Featherstone, W. E. (2010). Comparison and validation of the recent freely available ASTER-GDEM ver1, SRTM ver4.1 and GEODATA DEM-95 ver3 digital elevation models over Australia. Australian Journal of Earth Sciences, 57(3), 337–347. https://doi.org/10.1080/08120091003677753

Horritt, M. S., Bates, P. D., Fewtrell, T. J., Mason, D. C., & Wilson, M. D. (June 2010). Modelling the hydrodynamics of the Carlisle 2005 flood event. Proceedings of the Institution of Civil Engineers-Water Management (163, pp. 273–281). Westminster: Thomas Telford Ltd. 6. https://doi.org/10.1680/wama.2010.163.6.273

Hutchinson, M. F., Stein, J. A., Stein, J. L., & Xu, T. (2009). Locally adaptive gridding of noisy high resolution topographic data. In Proceedings of the Second Continental Workshop on the Geoid in Europe (pp. 199–204). Budapest.

Kulp, S. A., & Strauss, B. H. (2018). Coastal DEM: A global coastal digital elevation model improved from SRTM using a neural network. Remote Sensing of Environment, 206, 231–239. https://doi.org/10.1016/j.rse.2017.12.026

Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L. (2016). Machine learning in geosciences and remote sensing. Geoscience Frontiers, 7(1), 3–10. https://doi.org/10.1016/j.gsf.2015.07.003.

Li, J., Chu, Y., & Xu, X. (2017). Determination of vertical datum offset between the regional and the global height datum. Journal of Geodesy and Geoinformation Science, 46(10), 1262. https://doi.org/10.1016/j.jagcgs.2017.1070538

Lu, D., Tian, H., Zhou, G., & Ge, H. (2008). Regional mapping of human settlements in southeast China with multisensor remotely sensed data. Remote Sensing of Environment, 112(9), 3668–3679. https://doi.org/10.1016/j.rse.2008.05.009

McClean, F., Dawson, R., & Kilbey, C. (2020). Implications of using global digital elevation models for flood risk analysis in cities. Water Resources Research, 56(10), e2020WR028241. https://doi.org/10.1029/2020wr028241

Neal, J. C., Bates, P. D., Fewtrell, T. J., Hunter, N. M., Wilson, M. D., & Horritt, M. S. (2009). Distributed whole city water level measurements from the Carlisle 2005 urban flood event and comparison with hydraulic model simulations. Journal of Hydrology, 368(1–4), 42–55. https://doi.org/10.1016/j.jhydrol.2009.01.026

Neal, J., Schumann, G., & Bates, P. (2012). A subgrid channel model for simulating river hydraulics and floodplain inundation over large and data sparse areas. Water Resources Research, 48(11). https://doi.org/10.1029/2012WR012514

O’Loughlin, F. E., Paiva, R. C. D., Durand, M., Alsdorf, D. E., & Bates, P. D. (2016). A multi-sensor approach towards a global vegetation corrected SRTM DEM product. Remote Sensing of Environment, 182, 49–59. https://doi.org/10.1016/j.rse.2016.04.018

Ordinance Survey. (2018). A guide to coordinate systems in Great Britain. V3.3. Retrieved from https://www.ordnancesurvey.co.uk/documents/resourc es/guide-coordinate-systems-great-britain.pdf

Pham, H. T., Marshall, L., Johnson, F., & Sharma, A. (2018). A method for combining SRTM DEM and ASTER GDEM2 to improve topography estimation in regions without reference data. Remote Sensing of Environment, 210, 229–241. https://doi.org/10.1016/j.rse.2018.03.026

Rizzoli, P., Martone, M., Gonzalez, C., Becklich, C., Borla Tridon, D., Bräutigam, B., et al. (2017). Generation and performance assessment of the global TanDEM-X digital elevation model. ISPRS Journal of Photogrammetry and Remote Sensing, 132, 119–139. https://doi.org/10.1016/j.isprsjprs.2017.08.008

Rosi, C., & Gernhardt, S. (2013). Urban DEM generation, analysis and enhancements using TanDEM-X. ISPRS Journal of Photogrammetry and Remote Sensing, 85, 120–131. https://doi.org/10.1016/j.isprsjprs.2013.08.006

Sampson, C. C., Smith, A. M., Bates, P. D., Neal, J. C., Allferi, L., & Freer, J. E. (2015). A high-resolution global flood hazard model. Water Resources Research, 51(9), 7358–7381. https://doi.org/10.1002/2015wr16954

Schumann, G. J. P., & Bates, P. D. (2018). The need for a high-accuracy, open-access global DEM. Frontiers in Earth Science, 6, 225. https://doi.org/10.3389/feart.2018.00225

Shi, L., Fooody, G. M., Boyd, D. S., Girindran, R., Wang, L., Du, Y., & Ling, F. (2020). Night-time lights are more strongly related to urban building volume than to urban area. Remote Sensing Letters, 11(1), 29–36. https://doi.org/10.1080/2150704X.2019.1682709

Sinha, P., Gaughan, A. E., Stevens, F. R., Nieves, J. J., Sorichetta, A., & Tatem, A. J. (2019). Assessing the spatial sensitivity of a Random Forest model: Application in gridded population modeling. Computers, Environment and Urban Systems, 75, 132–145. https://doi.org/10.1016/j.compenvurbs.2019.01.006

Steven, R. S., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). Disaggregating census data for population mapping using Random Forests with remotely-sensed and ancillary data. PLoS One, 10(2), e0107042. https://doi.org/10.1371/journal.pone.0107042

Stevenson, J. A., Sun, X., & Mitchell, N. C. (2010). Despeckling SRTM and other topographic data with a denoising algorithm. Geomorphology, 114(3), 238–252. https://doi.org/10.1016/j.geomorph.2009.07.006

Su, Y., Guo, Q., Ma, Q., & Li, W. (2015). SRTM DEM correction in vegetated mountain areas through the integration of spaceborne LiDAR, airborne LiDAR, and optical imagery. Remote Sensing, 7(9), 11202–11225. https://doi.org/10.3390/rs7111202

Uuemaa, E., Ahli, S., Montbelier, B., Murr, M., & Knock, A. (2020). Vertical accuracy of freely available global digital elevation models (ASTER, AW3D30, MERIT, TanDEM-X, SRTM, and NASADEME). Remote Sensing, 12(21), 3482. https://doi.org/10.3390/rs12213482

Wecklich, C., Gonzalez, C., & Rizzoli, P. (2018). Height accuracy and data coverage for the final global TanDEM-X DEM data. Geomorphometry, 2018. 5th. https://doi.org/10.29007/kxq5
Wendleder, A., Felbier, A., Wessel, B., Huber, M., & Roth, A. (2016). A method to estimate long-wave height errors of SRTM C-band DEM. *IEEE Geoscience and Remote Sensing Letters*, 13(5), 696–700. https://doi.org/10.1109/lgrs.2016.2538822

Wessel, B., Huber, M., Wohlfart, C., Marschalk, U., Kosmann, D., & Roth, A. (2018). Accuracy assessment of the global TanDEM-X Digital Elevation Model with GPS data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 139, 171–182. https://doi.org/10.1016/j.isprsjprs.2018.02.017

Wing, O. E. J., Bates, P. D., Sampson, C. C., Smith, A. M., Johnson, K. A., & Erickson, T. A. (2017). Validation of a 30 m resolution flood hazard model of the conterminous United States. *Water Resources Research*, 53(9), 7968–7986. https://doi.org/10.1002/2017wr020917

Xie, Y., & Weng, Q. (2016). Updating urban extents with nighttime light imagery by using an object-based thresholding method. *Remote Sensing of Environment*, 187, 1–13. https://doi.org/10.1016/j.rse.2016.10.002

Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P. D., Allen, G. H., & Pavelsky, T. M. (2019). MERIT hydro: A high-resolution global hydrography map based on latest topography dataset. *Water Resources Research*, 55(6), 5053–5073. https://doi.org/10.1029/2019wr024873

Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J. C., et al. (2017). A high-accuracy map of global terrain elevations. *Geophysical Research Letters*, 44(11), 5844–5853. https://doi.org/10.1002/2017gl072874

Yamazaki, D., Sato, T., Kanae, S., Hirabayashi, Y., & Bates, P. D. (2014). Regional flood dynamics in a bifurcating mega delta simulated in a global river model. *Geophysical Research Letters*, 41(9), 3127–3135. https://doi.org/10.1002/2014gl059744

Zhao, X., Su, Y., Hu, T., Chen, L., Gao, S., Wang, R., et al. (2018). A global corrected SRTM DEM product for vegetated areas. *Remote Sensing Letters*, 9(4), 393–402. https://doi.org/10.1080/2150704x.2018.1425560

Zhou, D., Liu, Y., Hu, S., Hu, D., Neto, S., & Zhang, Y. (2019). Assessing the hydrological behaviour of large-scale potential green roofs retrofitting scenarios in Beijing. *Urban Forestry & Urban Greening*, 40, 105–113. https://doi.org/10.1016/j.ufug.2017.12.010

References From the Supporting Information

Ordnance Survey. (2017). *Topography layer-building height attribute product guide contents v1.2 – 10/2018*. Retrieved from https://www.ordnancesurvey.co.uk/documents/osmm-topo-layer-bha-product-guide.pdf