



Hawker, L., Uhe, P., Paulo, L., Sosa, J., Savage, J., Sampson, C., & Neal, J. (2022). A 30m global map of elevation with forests and buildings removed. *Environmental Research Letters*, 17(2), [024016]. <https://doi.org/10.1088/1748-9326/ac4d4f>

Publisher's PDF, also known as Version of record

License (if available):
CC BY

Link to published version (if available):
[10.1088/1748-9326/ac4d4f](https://doi.org/10.1088/1748-9326/ac4d4f)

[Link to publication record in Explore Bristol Research](#)
PDF-document

This is the final published version of the article (version of record). It first appeared online via IOP Publishing at <https://doi.org/10.1088/1748-9326/ac4d4f> . Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: <http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/>

LETTER • OPEN ACCESS

A 30 m global map of elevation with forests and buildings removed

To cite this article: Laurence Hawker *et al* 2022 *Environ. Res. Lett.* **17** 024016

View the [article online](#) for updates and enhancements.

You may also like

- [Assessment of microcirculation in the diabetic foot with laser speckle contrast imaging](#)
O A Mennes, J J van Netten, J G van Baal et al.
- [Evaluating Greenland glacial isostatic adjustment corrections using GRACE, altimetry and surface mass balance data](#)
Tyler C Sutterley, Isabella Velicogna, Beata Csatho et al.
- [Analysis the Accuracy of Digital Elevation Model \(DEM\) for Flood Modelling on Lowland Area](#)
Ku Hasna Zainurin Ku Zainol Abidin, Mohd Adib Mohammad Razi and Saifullizan Mohd Bukari

ENVIRONMENTAL RESEARCH
LETTERS

LETTER

A 30 m global map of elevation with forests and buildings removed

OPEN ACCESS

RECEIVED
2 November 2021REVISED
17 January 2022ACCEPTED FOR PUBLICATION
20 January 2022PUBLISHED
3 February 2022

Original Content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](https://creativecommons.org/licenses/by/4.0/).

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Laurence Hawker^{1,2,4,*} , Peter Uhe^{1,2,3,4} , Luntadila Paulo³, Jeison Sosa³, James Savage³, Christopher Sampson³ and Jeffrey Neal^{1,2,3}¹ School of Geographical Sciences, University of Bristol, Bristol, United Kingdom² Cabot Institute for the Environment, University of Bristol, Bristol, United Kingdom³ Fathom, Square Works, 17-18 Berkeley Square, Bristol, United Kingdom⁴ These authors contributed equally to this work.

* Author to whom any correspondence should be addressed.

E-mail: laurence.hawker@bristol.ac.uk**Keywords:** digital elevation model, bare-earth, terrain, remote sensing, machine learningSupplementary material for this article is available [online](#)**Abstract**

Elevation data are fundamental to many applications, especially in geosciences. The latest global elevation data contains forest and building artifacts that limit its usefulness for applications that require precise terrain heights, in particular flood simulation. Here, we use machine learning to remove buildings and forests from the Copernicus Digital Elevation Model to produce, for the first time, a global map of elevation with buildings and forests removed at 1 arc second (~30 m) grid spacing. We train our correction algorithm on a unique set of reference elevation data from 12 countries, covering a wide range of climate zones and urban extents. Hence, this approach has much wider applicability compared to previous DEMs trained on data from a single country. Our method reduces mean absolute vertical error in built-up areas from 1.61 to 1.12 m, and in forests from 5.15 to 2.88 m. The new elevation map is more accurate than existing global elevation maps and will strengthen applications and models where high quality global terrain information is required.

1. Introduction

Topographic information, described here as digital elevation models (DEMs), are crucial inputs for a diverse range of applications in fields such as demography [1], ecology [2, 3], geomorphology [4, 5], glaciology [6, 7], hydrology [8–10], soil science [11] and volcanology [12, 13]. At the global scale, elevations are measured from space using near-infrared, radar and visible sensors, and then georectified and stored in a gridded storage structure. DEMs exist at the global scale, or near-global scale, at 3 arc second (~90 m), and more recently 1 arc second (~30 m) grid spacing. The global DEMs at 3 arc second grid spacing include the widely used Shuttle Radar Topography Mission (SRTM) [14], as well as MERIT [15], TanDEM-X 90 [16, 17] and Copernicus GLO-90. At 1 arc second, products include ASTER GDEM (v3) [18], ALOS World3D AW3D30 (v3.2) [19], SRTM

[14], NASADEM [20] and most recently Copernicus GLO-30 [21] (herewith referred to as COPDEM30).

The acquisition period, sensor type, processing, post-processing and geographical extent varies between DEMs [22]. As a result, accuracy, and therefore suitability, for a given application varies. COPDEM30, and the underlying TanDEM-X data, is the most recent and accurate global DEM [23–25]. COPDEM30 can accurately resolve most features, with Guth and Geoffroy [25] going as far to say COPDEM30 should become the ‘gold standard’ for global DEMs. Hence COPDEM30 was chosen as the basis of this work for producing a global bare-earth DEM.

DEMs can be broadly separated into digital surface models (DSM), which measures the upper surface of trees, buildings and other man-made features, and digital terrain models (DTMs) which measures the elevation of the ground surface, or the ‘bare-earth’

[26]. Each DEM has a different abstraction of the real surface. The presence of tree and building bias in DEMs is problematic in some applications, particularly in hydrology, where artifacts in the DEM can act as dams that change overland flow pathways. It is up to the user to decide which abstraction is best for their application, and they can be guided by the numerous accuracy assessment studies [27–30].

DEMs at the global scale fall towards the DSM end of the DSM-DTM spectrum. The MERIT DEM is an exception, as tree height bias has been removed, along with absolute bias, speckle noise and striping. Therefore, MERIT DEM is closer to a DTM than other global DEMs. Other examples of tree height bias being removed are in the quasi global CoastalDEM [31] and vegetation removal from SRTM of O’Loughlin *et al* [32], Zhao *et al* [33] or Magruder *et al* [34]. Although there have been promising recent developments [35], the removal of buildings from MERIT at the global scale has not yet occurred, and thus MERIT cannot be considered completely as a DTM. With the rich amount of auxiliary data related to global building and forest coverage now available, there is increasing potential to utilise machine learning techniques to remove trees [31], buildings [35, 36] or both [37–39], although to date these studies are limited to the local or quasi global scale. Machine learning allows us to ‘learn by example’ by building empirical models from the data alone and is particularly well suited to non-linear settings [40].

Using machine learning techniques, we remove building and tree height bias from COPDEM30 to create a new data-set called FABDEM (Forest And Buildings removed Copernicus DEM). The new data-set is available between 60° S and 80° N at 1 arc second (~30 m) grid spacing and is the first global DEM to remove both trees and buildings. We validate against reference elevation data and compare it to other global DEMs. Many applications where representation of the terrain is important will benefit from the improved resolution and accuracy of FABDEM compared to existing global DEMs. This is especially true of flood inundation modelling, where the terrain is a key determinant of water flows and hence locations of flooding.

2. Methods

2.1. Random forest regression

This study uses random forest regression models, which are popular in similar large-scale applications in remote sensing [41]. Random forest regression has recently been used for forest correction of SRTM [42], urban correction of MERIT [35] and correcting SRTM in New Zealand [39]. However to date, correction of DEMs using random forest models has not been utilised at the global scale.

In the preliminary stage of this study, we assessed generalized linear models (GLMs) and generalized

additive models (GAM). Neural networks have also been used successfully at the quasi-global scale for the CoastalDEM [31] and at the local scale [37–39]. We found random forests to be accurate, computationally efficient and robust, hence an appropriate choice for this application.

Random forests take a ‘divide and conquer’ approach by combining outcomes from a sequence of regression decision trees, and are particularly popular as they have few parameters to tune, are robust to over-fitting and are robust to small and large datasets [43, 44]. Accuracy is measured by ‘out of bag error’ which estimates error computed on reference data set aside prior to building the random forest model. In this study, 25% of the reference data are set aside for validation. We utilise a GPU accelerated random forest implementation in python which shows enhanced computational efficiency compared to CPU equivalents [45].

Once a machine learning algorithm is selected, the other key ingredients are a comprehensive reference data-set for training the model, and predictor variables. In machine learning, predictors or features are input variables used to predict the reference data or the target variable, hence predictors must be conceptually relatable to the application. Different predictors are used in the forest and buildings correction models, described further in sections 2.4 and 2.5.

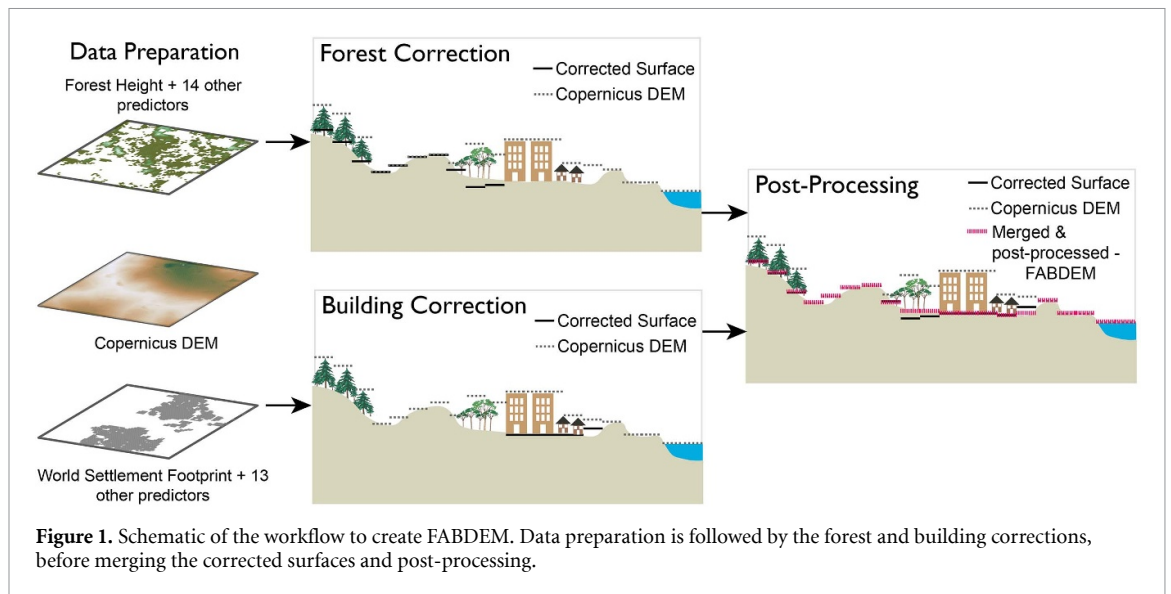
2.2. Correction workflow

The workflow of producing FABDEM is shown in figure 1. This consists of three stages: (1) data preparation, consisting of processing predictor data and reference DEMs; (2) Random forest correction, done separately for forests and buildings removal and (3) post-processing, to merge the corrected DEMs, fill unrealistic pits and apply smoothing filters. Each step of the workflow is described below, as well as the data contributing towards the machine learning based correction.

2.3. Data preparation

All data used in the FABDEM production were processed and regrided to the COPDEM30 grid (EPSG 4326 projection, at 1 arc second horizontal grid spacing). The predictor data and reference DEMs were then compiled into a tabular data format used by the random forest model.

Predictor data that were determined to be useful in estimating anomalies in the DSM associated with buildings and trees were chosen. The key datasets describing forest height, forest cover and building footprints have a similar resolution (10–100 m) to COPDEM30, allowing them to provide information at the grid-cell level. Predictor variables are described in the respective sections on Forests and Buildings removal (2.4 and 2.5) and the full list of predictors are listed in supplementary table S1 (available online at stacks.iop.org/ERL/17/024016/mmedia).



An extensive search of LiDAR DTM data was carried out to select reference data for training the random forest algorithm. We specified that LiDAR data should have an accuracy of <1 m which comfortably adheres to the recommendation that reference data should be at least three times more accurate than the DEM [46]. A visual inspection of the data was initially implemented as a first quality check. LiDAR DTMs from 12 countries were selected to use in training the random forest algorithm, see supplementary table S2 for details of each data set. This reference data covers a wide range of climatic zones and urban areas, thus allowing us to use our approach at a global scale. A map showing spatial distribution of these data sets is included in supplementary figure S1. The reference DEMs were reprojected to the COPDEM30 horizontal grid and to the COPDEM30 vertical coordinates (EGM2008), using geoid models specifying the vertical datum used in each reference DTM.

2.4. Forest removal

We use a random forest model to predict the difference between COPDEM30 elevations and terrain elevations from a reference data (LiDAR) for forested regions. Therefore, variables are needed to estimate the height of the forest and the canopy coverage of the forest. The characteristic of the forest is important as the penetration of the radar signal used in COPDEM30 penetrates the tree canopy differently depending on the characteristics of the tree and the radar characteristics [47, 48].

The predictor data for forest height is taken from the Global Forest Canopy Height 2019 data-set [49]. Global Ecosystem Dynamics Investigation (GEDI) LiDAR data onboard the International Space Station [50] has been integrated with Landsat GLAD ARD [51] data to create an estimate of forest canopy height at 30 m grid spacing. Coverage is limited to the extent of GEDI footprint-based measurements, which are

collected between 51.6° N and 51.6° S. A separate predictor variable, tree cover fraction from the Copernicus Global Land Service collection 3 epoch 2015 [52] was used to characterise canopy coverage of each 100 m cell globally. To decide where the regression model was applied, cells where both the forest height was greater than 3 m and tree cover was greater than 10% were considered forested for the vegetation correction algorithm below 52° N.

In regions outside the Global Forest Canopy Height 2019 data-set (above 52° N), we estimated canopy height using ICESat-2 L3A Land and Vegetation Height (version 4) ATL-08 data [53]. The ‘h_canopy’ and h_canopy_mean’ variables were used, representing the top of canopy and mean canopy height respectively [54]. The ATL-08 terrain measurements have a reported bias of -0.07 to 0.18 m [55, 56]. ICESat-2 records with non missing-values over land were used. Additionally, records were discarded if h_canopy was greater than 50 m and for h_canopy values between 30–50 m, if the layer_flag was equal to 1. Values between ICESat-2 canopy height measurements were interpolated to increase the coverage of canopy height estimates. Cells with canopy coverage $>10\%$ from the Copernicus Global Land Service Collection were considered forested and used in the forest correction algorithm above 52° N.

Additional predictor variables for the machine learning model were created by applying several image processing tools implemented in Whitebox-Tools [57]. These filters were applied to both the COPDEM30 elevations and the tree height data, to detect features such as edges, anomalies or variability that might be due to forests. For example, a sobel edge detector filter with a 3×3 window was applied, and an unsharp filter was applied to emphasize edges, while reducing noise. Filters computing the difference between gaussian filters of varying sizes were also computed. Lastly, a bilateral edge-preserving

smoothing filter and gaussian filters to emphasize long-range variability were applied to the forest height variable only.

An overview of the data-sets used to train the model and their filtered versions are in table S1 of the supplementary material, and the number of samples per country of LiDAR data are in table S2.

2.5. Buildings removal

A separate random forest model was built to predict the differences between COPDEM30 elevation and terrain elevation from LiDAR for urban areas. Multiple predictor data-sets were used for building removal to characterise factors relating to building height, for example building footprints, population and socio-economic indicators. In their urban correction of MERIT DEM, Liu *et al* [35] reported limitations of the transferability of their model due to the relatively small training data-set being unable to capture the variability of buildings in urban areas, leading to overestimation particularly in smaller cities. Hence in this study, we use a range of independent data-sets to increase the applicability of our model.

Predictor data-sets include information on population density (WorldPop [58]), travel times [59], night-time lights [60], urban building footprints (World Settlement Footprint [61]), built-up areas per capita, GDP and average greenness (GHS-UCDB R2019A [62]). Further detail on these predictor variables is given in the supplementary text S2 and table S1.

Data on building footprints and density from OpenStreetMap were not used due to the lack of global consistency [63] and the low importance of the variable(s) in other location specific random forest based DEM correction studies [35, 39]. Log transformations were applied to population, GDP per area and travel times. As for the forest correction, filters (difference of gaussians, sobel and unsharp) were applied, but in this case to detect edges of urban areas or taller buildings in COPDEM30. The number of samples per LiDAR country for building removal are detailed in table S2.

2.6. Post processing

Correction surfaces simulated through random forest models for both forest and buildings were subtracted from COPDEM30 to get the terrain elevations. Therefore, as illustrated in figure 1, two intermediate layers were produced—COPDEM30 Building removed and COPDEM30 Forest removed. The COPDEM30 Building removed and COPDEM30 Forest removed layers were subsequently merged, taking the minimum value at each pixel. Locations adjacent to the COPDEM30 water-body mask (WBM), were handled separately, without buildings or forest removal being applied. This was to ensure that coastline/riverbank pixels were not overly lowered, but were kept consistent with the surrounding terrain. Locations

underneath the WBM had no modifications made and are essentially identical to the COPDEM30 values.

Additional steps were applied to correct areas that had been corrected too much, or areas where pixels were noisy. Firstly, to remove artifacts outside building and forest corrected areas, pits up to 4 pixels in size were filled. In building and forest removed areas, large depressions were filled, but not higher than the original COPDEM30 (after filling small pits in the COPDEM30 of up to 100 pixels in size). Only pixels that had been adjusted (i.e. Building and Forested) were filled. Noise in the DEM was subsequently reduced by running an adaptive filter twice, and then a bilateral filter. An adaptive filter is effective at removing speckle noise [64], and was used in MERIT DEM [15]. A bilateral filter is an edge preserving smoothing filter [65] that preserves edges of features but reduces short-scale variation. These filters were implemented in WhiteBoxTools [57].

Further smoothing was done for pixels adjacent to the WBM, with a 5×5 pixel median filter applied to non-water pixels, with a subsequent single pit filling step. This step was necessary as we noted COPDEM30 contained incidences of pits. Finally, all smoothed sections of the DEM (values adjacent to WBM, WBM and all other land pixels) were combined to create FABDEM. Guidance of parameter selection for the filters are limited at the scale of global DEMs [66, 67], as studies predominantly focus on finer grid spacing LiDAR data [68–70]. Therefore, parameters were selected using an extensive trial and error approach.

3. Results and discussion

To demonstrate the utility of the FABDEM data-set we validate the DEM against reference elevation data, as well as comparing against other global DEMs. We use two sources of reference data. (1) LiDAR data from 12 countries (listed in supplementary table 2) and (2) ICESat best estimates of the ground terrain, randomly sampled from locations around the globe. Additionally, reference data over Florida, USA (not used in training) was obtained for a visual comparison.

We note here that the reference LiDAR data used for validation in section 3.2 are the same as the training data. A random sample of 1000 points per tile that contains LiDAR data were used for validation. These sampled points were taken from the whole reference LiDAR dataset. Similarly, the training data was randomly sampled from the reference dataset (see table S2 for details), so some of the validation points were also used for training. However, the ICESat elevation data are not used for training, and are evenly distributed across the globe, so give an independent estimate of errors across all terrain types.

Other global DEMs were also compared against FABDEM: the original COPDEM30,

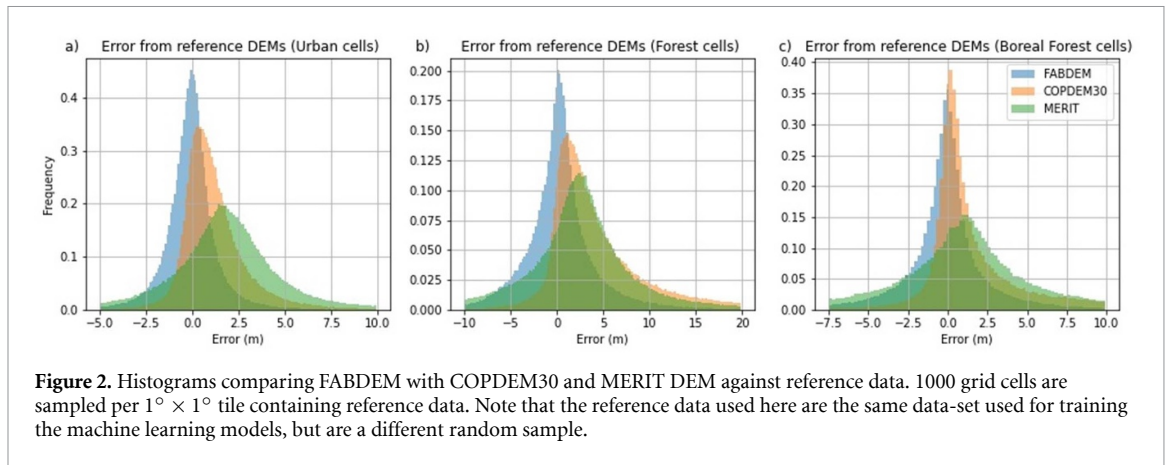


Table 1. Error statistics comparing COPDEM30, FABDEM and MERIT DEM against reference DEMs for each of the correction methods. Comparisons are only made for land cells where the FABDEM correction of the respective type was applied. ME is mean error, MAE is mean absolute error and RMSE is root mean squared error.

Area	DEM	ME	MAE	RMSE	90% Error	Error % <2 m
Urban	COPDEM30	0.86	1.61	2.87	3.54	74.77%
	FABDEM	-0.08	1.12	2.33	2.39	86.37%
	MERIT DEM	1.74	3.59	7.15	6.86	42.73%
Forest	COPDEM30	2.93	5.15	7.98	12.51	35.52%
	FABDEM	0.20	2.88	4.96	6.67	55.07%
	MERIT DEM	2.37	5.47	8.06	12.08	26.72%
Boreal Forest	COPDEM30	0.72	3.49	7.73	9.64	62.46%
	FABDEM	-0.11	2.55	6.66	5.51	68.70%
	MERIT DEM	1.14	9.38	19.62	24.24	33.58%

MERIT DEM [15], CoastalDEM [31] and NASA-DEM [20]. Particular focus is made of the comparison against MERIT DEM, as MERIT DEM has had errors from vegetation removed, and is thus most similar to FABDEM.

3.1. Split-sample validation

When building the random forest models, 75% of the data samples were used for training the models and 25% of the data samples were kept back and used for validation only. This approach is known as split-sample or out of bag validation, and is useful to ensure the model is not being over-fitted and has some applicability outside of the training set.

For the urban correction, the reference COPDEM30 data had a mean absolute error (MAE) of 1.72 m, which was reduced to 0.94 m for predictions on the training sample, while predictions on the validation sample had MAE of 1.35 m. For the forest model (south of 52° N), the split-sample validation results were: MAE of 7.2 m for the COPDEM30 data, MAE of 3.52 m in predictions on the training sample, and MAE of 6.55 m in predictions on the validation sample. Finally, for the boreal forest model (north of 52° N), the split-sample validation results were MAE of 3.77 m for COPDEM30 data, MAE of 1.72 m in predictions on the training sample, and MAE of 3.24 m in predictions on the validation sample.

In each case, the errors in the validation samples are greater than the training samples, but reduced compared to the COPDEM30 data. The results for the split-sample validations above are errors against reference data for the random forest model predictions before any post-processing is applied. So we validate separately on the final DEM after post-processing.

3.2. Global comparison

Figure 2 shows histograms of errors of FABDEM, COPDEM30 and MERIT DEM against reference data. Panels (a)–(c) show errors for urban, forest areas (south of 52° N) and boreal forest areas (north of 52° N) respectively. These consistently have lower errors in FABDEM compared to COPDEM30 and MERIT DEM, showing the benefit of both the newer COPDEM30 data-set compared to elevations based on SRTM, and the effectiveness of the forests and buildings correction applied for FABDEM. Table 1 presents some statistics corresponding to the histograms in figure 2. FABDEM has median errors close to 0 m, and the lowest errors in each of the statistics, showing a improvement on both COPDEM30 and MERIT DEM.

Figure 3 shows errors compared to ICESat-2 L3A Land and Vegetation Height (version 4) ATL-08 data [53]: for all land pixels (a) and over flood plains (b). Flood plains were delineated using the

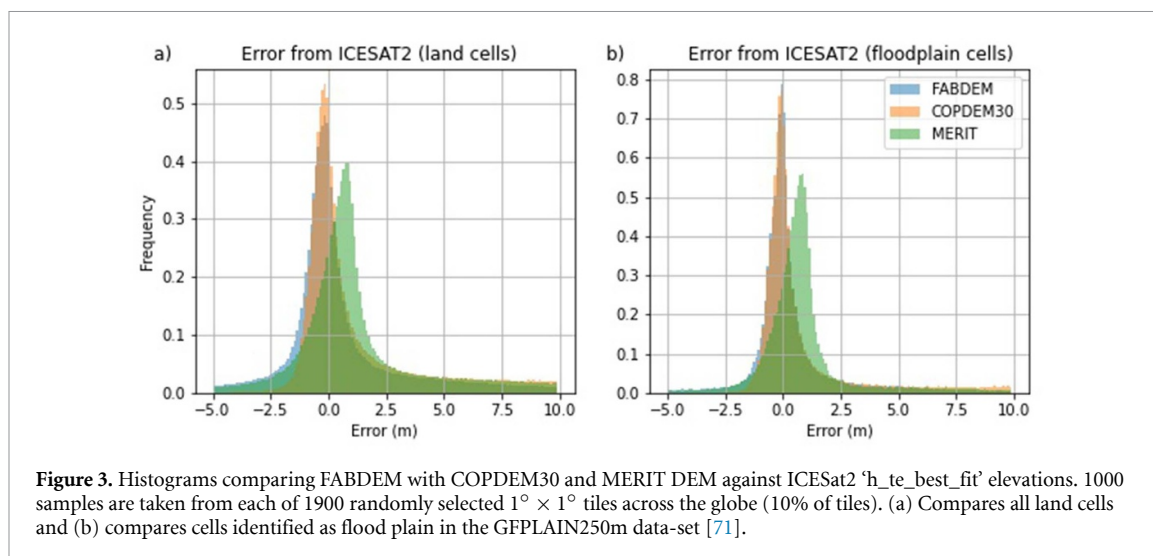


Figure 3. Histograms comparing FABDEM with COPDEM30 and MERIT DEM against ICESat2 'h_te_best_fit' elevations. 1000 samples are taken from each of 1900 randomly selected $1^\circ \times 1^\circ$ tiles across the globe (10% of tiles). (a) Compares all land cells and (b) compares cells identified as flood plain in the GFPLAIN250m data-set [71].

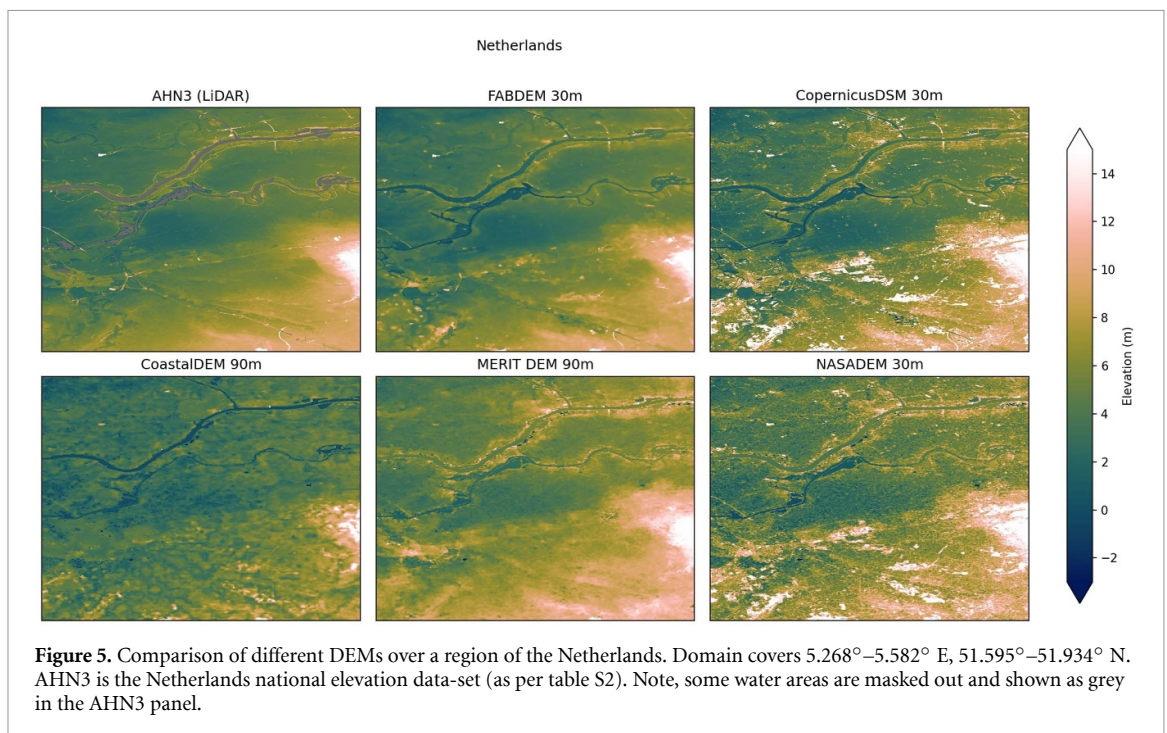
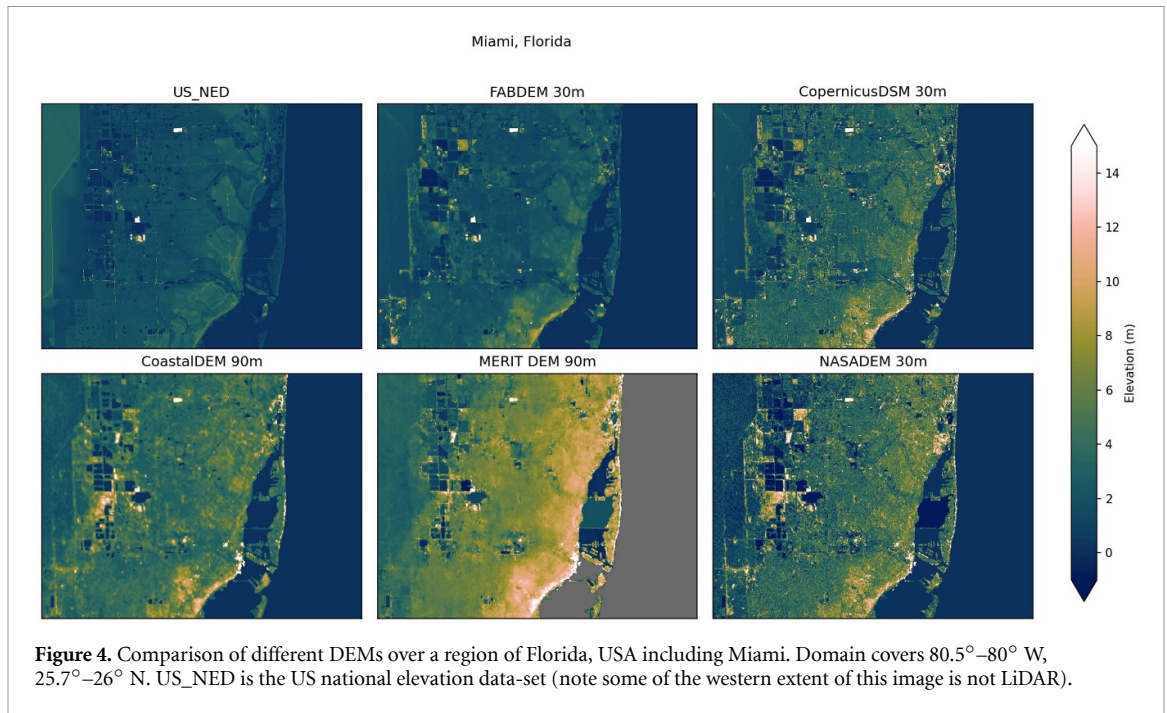
Table 2. Error statistics comparing COPDEM30, FABDEM and MERIT DEM against ICESat2 elevation data for land points, floodplain points (defined by the GFPLAIN250m data-set), canopy cover and population density. ME is mean error, MAE is mean absolute error and RMSE is root mean squared error.

Cells	DEM	ME	MAE	RMSE	90% error	Error % <2 m
Land points	COPDEM30	0.71	6.37	14.32	19.70	57.99%
	FABDEM	-0.07	3.41	11.13	8.85	67.50%
	MERIT DEM	0.81	4.81	13.13	12.55	60.90%
Floodplain	COPDEM30	0.17	4.04	10.21	14.29	71.14%
	FABDEM	-0.03	2.03	7.87	4.90	80.42%
	MERIT DEM	0.66	2.35	8.04	5.61	80.19%
Populated (<5000 km ²)	COPDEM30	0.98	4.22	12.95	11.16	62.42%
	FABDEM	0.35	3.50	12.46	8.88	70.32%
	MERIT DEM	0.75	4.14	12.90	10.23	62.96%
Populated (>5000 km ²)	COPDEM30	1.90	4.99	14.33	12.73	50.90%
	FABDEM	0.19	3.94	13.75	9.90	61.31%
	MERIT DEM	0.85	4.35	14.16	11.72	59.04%
Canopy cover: 10%–50%	COPDEM30	0.55	3.41	11.89	9.43	69.96%
	FABDEM	0.20	2.92	11.53	7.59	75.96%
	MERIT DEM	0.71	3.84	12.74	9.60	68.57%
Canopy cover: >50%	COPDEM30	12.09	14.03	26.70	26.55	13.33%
	FABDEM	0.45	6.10	22.17	12.76	35.29%
	MERIT DEM	2.95	8.14	23.60	17.72	30.81%

GFPLAIN250m data-set [71]. We use the ICESat-2 'h_te_best_fit' variable in this evaluation, which is the best estimate of terrain elevation [72, 73]. ICESat-2 ATL-08 canopy and ground surface data are processed in 100 m segments where a suitable number of photons are present, resulting in features not always being detected that are in the 1 arc second DEMs.

As both non-urban and non-forested pixels (which are not corrected) are considered, the histogram of FABDEM and COPDEM30 are very similar, with both showing errors closer to 0 m compared to MERIT DEM. Table 2 presents some statistics corresponding to the histograms in figure 3, and some additional categories based on population density and forest canopy cover. The COPDEM30 and FABDEM

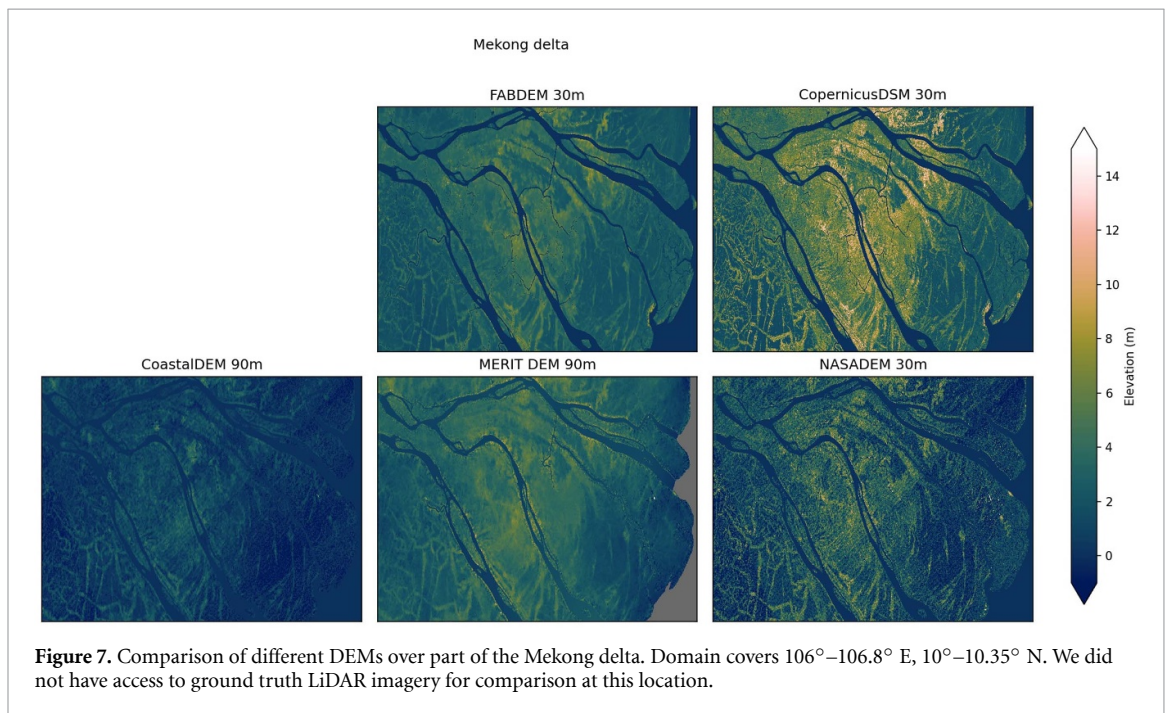
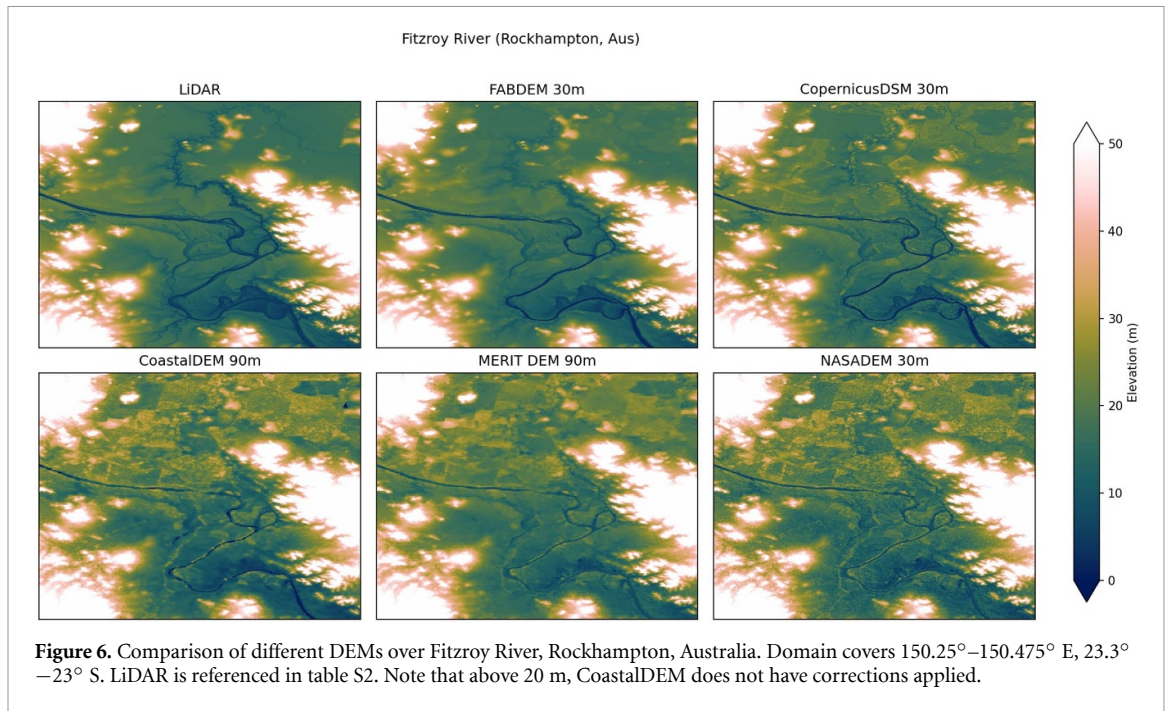
distributions are centered around 0 m error, but have a relatively long tail. The MERIT DEM distribution is centered around 1 m error, which is reflected by the worst median error for floodplains, at 0.66 m, compared to -0.03 m of FABDEM. There are fewer large errors in MERIT DEM and FABDEM compared to COPDEM30, as reflected by the lower RMSE and 90% error values and higher % error <2 m values. Indeed, over floodplains, FABDEM and MERIT DEM have very similar error <2 m at 80.42% and 80.19% respectively, with FABDEM having slightly better RMSE and 90% error values suggesting the tail of the error distribution is not quite as long as MERIT DEM. This is likely to be due to FABDEM having urban areas removed.



The spatial distribution of median biases in COPDEM30 and FABDEM compared to ICESat is shown in figure S2. This highlights the large scale biases in COPDEM which are removed in FABDEM. Regions do exist with biases remaining in FABDEM, including areas of steep terrain, however these areas also correspond to higher biases in ICESat [74].

In populated areas, FABDEM has the lowest median error, RMSE and 90% error, with the improvement over COPDEM30 most evident in the densest populated areas ($>5000 \text{ km}^2$) where the

median error of FABDEM is 0.19 m compared to 1.90 m in COPDEM30. The correction in forested areas is best for denser canopy coverage ($>50\%$), with a median error of 0.45 m for FABDEM compared to 2.95 m in MERIT DEM and 12.95 m for COPDEM30. In this case, The RMSE and error <2 m is still relatively high for FABDEM suggesting the tails of the distribution have not been completely corrected, but still much improved from COPDEM30 and MERIT DEM. Therefore, the statistical results suggest FABDEM is the most accurate global DEM for



all types of cell assessed, with median error ranging from -0.11 to 0.45 m and error statistics significantly better than other global DEMs.

3.3. Visual evaluation

To qualitatively establish the benefits of the corrections applied for FABDEM, we show spatial maps of reference data and five different global DEMs including FABDEM. Four locations are shown—one (Miami, USA), comparing against a reference DEM not included in the training of the random forest model; two (Netherlands and Australia), which were included in the training of the random forest models;

and one (Mekong delta), where the global DEMs are compared in the absence of a high quality local DEM.

In the first comparison, figure 4 shows part of Florida, USA, covering Miami. FABDEM produces the closest comparison to the US NED reference DEM. There are particular locations where artifacts from buildings remain in FABDEM, particularly along the coastal areas, however these are reduced compared to the COPDEM30. Similar artifacts are also present in the other global DEMs. Secondly, figure 5 shows a region of the Netherlands around the province of North Brabant, and sections of the Waal and Meuse rivers. Here again, the removal of

buildings and forests is evident in the comparison with COPDEM30. There are still some features not completely removed, particularly in the south-west of this domain, however, FABDEM again gives a more accurate comparison with the reference DEM compared to the other global DEMs. Thirdly, figure 6 shows a region in Australia, including part of the Fitzroy river near Rockhampton. This is a rural area where most of the correction seen between FABDEM and COPDEM30 is due to forests in the northern part of this domain. COPDEM30 and FABDEM additionally show a much clearer picture of the contours rivers and the floodplains in this domain compared to the other global DEMs. Finally, figure 7 shows part of the Mekong delta in Vietnam. FABDEM again has a substantial amount of correction compared to the COPDEM30, and is more consistent with the elevations of MERIT DEM, with FABDEM providing finer detail and better connectivity of the river channels.

4. Conclusion

This study presents the development of a new global DEM at 30 m grid-spacing, with artifacts from forests and buildings removed (FABDEM). The use of random forest machine learning models is a powerful tool to estimate anomalies in the terrain due to human settlements and forests. The median errors in FABDEM are close to zero and the absolute errors are reduced by up to half in our evaluation against reference DEMs.

Using the gold standard of global terrain data, the Copernicus GLO-30 DEM, gives this data-set an advantage over currently available global DEMs, many of which are derived from the 2000s era SRTM data-set. In comparison to other global DEMs, we find that FABDEM has smaller errors, and importantly resolves fine scale floodplain features which are not resolved by the SRTM based DEMs.

The correction procedure applied in the machine learning model relies on there being sufficient information in the predictor variables to characterise the artifacts we are aiming to correct (forests and building heights). In this study we sourced high quality public data-sets available at a global scale. However, these data-sets do have uncertainties and errors. This is evident in the visual evaluation, for example showing some artifacts from buildings which are not completely removed. Inclusion of more accurate predictor variables in the future will allow for improvements to the machine learning corrections in the future. The underlying Copernicus GLO-30 DEM was found to still have some random artifacts and pits, which were mostly removed by our processing but not completely. Seasonal vegetation anomalies are further difficult to predict. The user should note that the correction is applied in built and forested areas only, and should thus be mindful when using

the data in areas where errors are more likely to occur, particularly in areas of steep slopes.

In addition to the predictor variables, reference elevations for training data are key to producing accurate predictions in a machine learning model. A model using training data from one region of the world may be prone to over-fitting and can not necessarily be transferred to other regions. Hence for this study, we used a carefully selected set of reference DEMs from 12 countries around the world to increase the applicability of our models.

FABDEM has notable benefits compared to existing global DEMs, resulting from the use of the new Copernicus GLO-30 DEM and a machine learning correction of forests and buildings. This makes it preferable for many purposes where a bare-earth representation of terrain is needed, such as in hydrology and flood inundation modelling.

Data availability statement

The data that support the findings of this study are freely available for non-commercial use, at the following URL/DOI: <https://doi.org/10.5523/bris.25wfy0f9ukoge2gs7a5mqpq2j7>.

Acknowledgments

L H and J N were funded by a joint Natural Environment Research Council (NERC) and Vietnam National Foundation for Science and Technology Development (NAFOSTED) project under Grant No. NE/S3003061/1.

We acknowledge the developers of the produces contributing towards this data-set. Specifically, the Copernicus GLO-30 DEM developers, providers of the predictor data-sets (listed in supplementary table S1) and the reference LiDAR datasets (listed in supplementary table S2). We also acknowledge data.bris for hosting the FABDEM data-set.

ORCID iDs

Laurence Hawker  <https://orcid.org/0000-0002-8317-7084>

Peter Uhe  <https://orcid.org/0000-0003-4644-8559>

References

- [1] Zhang J, Wei X, Qin L and Tian Y 2018 Spatial distribution estimates of the urban population using DSM and DEM data in china *ISPRS Int. J. Geo-Inf.* **7** 435
- [2] Moore I D, Grayson R B and Ladson A R 1991 Digital terrain modelling: a review of hydrological, geomorphological and biological applications *Hydrol. Process.* **5** 3–30
- [3] Amatulli G, Domisch S, Tuanmu M-N, Parmentier B, Ranipeta A, Malczyk J and Jetz W 2018 A suite of global cross-scale topographic variables for environmental and biodiversity modeling *Sci. Data* **5** 180040

- [4] Farahmand A and AghaKouchak A 2013 A satellite-based global landslide model *Nat. Hazards Earth Syst. Sci.* **13** 1259–67
- [5] Zhang B, Fan Z, Du Z, Zheng J, Luo J, Wang N and Wang Q 2020 A geomorphological regionalization using the upscaled DEM: the Beijing-Tianjin-Hebei area, China case study *Sci. Rep.* **10** 10532
- [6] Rentsch H, Welsch W, Heipke C and Miller M M 1990 Digital terrain models as a tool for glacier studies *J. Glaciol.* **36** 273–8
- [7] Wang D and Kääb A 2015 Modeling glacier elevation change from DEM time series *Remote Sens.* **7** 10117–42
- [8] Sampson C C, Smith A M, Bates P D, Neal J C, Alfieri L and Freer J E 2015 A high-resolution global flood hazard model *Water Resour. Res.* **51** 7358–81
- [9] Yao F, Wang J, Yang K, Wang C, Walter B A and Crétaux J-F 2018 Lake storage variation on the endorheic Tibetan Plateau and its attribution to climate change since the new millennium *Environ. Res. Lett.* **13** 064011
- [10] Hirabayashi Y, Tanoue M, Sasaki O, Zhou X and Yamazaki D 2021 Global exposure to flooding from the new CMIP6 climate model projections *Sci. Rep.* **11** 3740
- [11] Park S J, McSweeney K and Lowery B 2001 Identification of the spatial distribution of soils using a process-based terrain characterization *Geoderma* **103** 249–72
- [12] Grosse P, van Wyk de Vries B, Euillades P A, Kervyn M and Petrinovic I A 2012 Systematic morphometric characterization of volcanic edifices using digital elevation models *Geomorphology* **136** 114–31
- [13] Kubanek J, Poland M P and Biggs J 2021 Applications of bistatic radar to volcano topography—a review of ten years of TanDEM-X *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **14** 3282–302
- [14] Farr T G et al 2007 The shuttle radar topography mission *Rev. Geophys.* **45** RG2004
- [15] Yamazaki D, Ikeshima D, Tawatari R, Yamaguchi T, O’Loughlin F, Neal J C, Sampson C C, Kanae S and Bates P D 2017 A high-accuracy map of global terrain elevations *Geophys. Res. Lett.* **44** 5844–53
- [16] Rizzoli P et al 2017 Generation and performance assessment of the global TanDEM-X digital elevation model *ISPRS J. Photogramm. Remote Sens.* **132** 119–39
- [17] Zink M et al 2021 TanDEM-X: 10 years of formation flying bistatic SAR interferometry *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **14** 3546–65
- [18] Abrams M, Crippen R and Fujisada H 2020 ASTER global digital elevation model (GDEM) and ASTER global water body dataset (ASTWBD) *Remote Sens.* **12** 1156
- [19] Tadono T, Nagai H, Ishida H, Oda F, Naito S, Minakawa K and Iwamoto H 2016 Generation of the 30 m-mesh global digital surface model by ALOS prism *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* **XLI-B4** 157–62
- [20] Crippen R et al 2016 NASADEM global elevation model: methods and progress *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* **XLI-B4** 125–8
- [21] AIRBUS 2020 Copernicus DEM: copernicus digital elevation model product handbook *Report AO/1-9422/18/I-LG* (European Space Agency) (Accessed 25 June 2020)
- [22] Guth P L et al 2021 Digital elevation models: terminology and definitions *Remote Sens.* **13** 3581
- [23] Gesch D B 2018 Best practices for elevation-based assessments of sea-level rise and coastal flooding exposure *Front. Earth Sci.* **6** 230
- [24] Grohmann C H 2018 Evaluation of TanDEM-X DEMs on selected Brazilian sites: comparison with SRTM, ASTER GDEM and ALOS AW3D30 *Remote Sens. Environ.* **212** 121–33
- [25] Guth P L and Geoffroy T M 2021 LiDAR point cloud and ICESat-2 evaluation of 1 second global digital elevation models: copernicus wins *Trans. GIS* **25** 2245–61
- [26] Polidori L and Mhamad E H 2020 Digital elevation model quality assessment methods: a critical review *Remote Sens.* **12** 3522
- [27] Hawker L, Neal J and Bates P 2019 Accuracy assessment of the TanDEM-X 90 digital elevation model for selected floodplain sites *Remote Sens. Environ.* **232** 111319
- [28] Gdulová K, Marešová J and Vítězslav M 2020 Accuracy assessment of the global TanDEM-X digital elevation model in a mountain environment *Remote Sens. Environ.* **241** 111724
- [29] McClean F, Dawson R and Kilsby C 2020 Implications of using global digital elevation models for flood risk analysis in cities *Water Resour. Res.* **56** e2020WR028241
- [30] Uuemaa E, Ahi S, Montibeller B, Muru M and Kmoch A 2020 Vertical accuracy of freely available global digital elevation models (ASTER, AW3D30, MERIT, TanDEM-X, SRTM, and NASADEM) *Remote Sens.* **12** 3482
- [31] Kulp S A and Strauss B H 2018 CoastalDEM: a global coastal digital elevation model improved from SRTM using a neural network *Remote Sens. Environ.* **206** 231–9
- [32] O’Loughlin F E, Paiva R C D, Durand M, Alsdorf D E and Bates P D 2016 A multi-sensor approach towards a global vegetation corrected SRTM DEM product *Remote Sens. Environ.* **182** 49–59
- [33] Zhao X, Yanjun S, Tianyu H, Chen L, Gao S, Wang R, Jin S and Guo Q 2018 A global corrected SRTM DEM product for vegetated areas *Remote Sens. Lett.* **9** 393–402
- [34] Magruder L, Neuenschwander A and Klotz B 2021 Digital terrain model elevation corrections using space-based imagery and ICESat-2 laser altimetry *Remote Sens. Environ.* **264** 112621
- [35] Liu Y, Bates P D, Neal J C and Yamazaki D 2021 Bare-earth DEM generation in urban areas for flood inundation simulation using global digital elevation models *Water Resour. Res.* **57** e2020WR028516
- [36] Olajubu V, Trigg M A, Berretta C, Sleigh A, Chini M, Matgen P, Mojere S and Mulligan J 2021 Urban correction of global DEMs using building density for Nairobi, Kenya *Earth Sci. Inform.* **14** 1383–98
- [37] Wendi D, Liong S-Y, Sun Y and Doan C D 2016 An innovative approach to improve SRTM DEM using multispectral imagery and artificial neural network *J. Adv. Model. Earth Syst.* **8** 691–702
- [38] Kim D E, Liu J, Liong S-Y, Gourbesville P and Strunz G 2021 Satellite DEM improvement using multispectral imagery and an artificial neural network *Water* **13** 1551
- [39] Meadows M and Wilson M 2021 A comparison of machine learning approaches to improve free topography data for flood modelling *Remote Sens.* **13** 275
- [40] Lary D J, Alavi A H, Gandomi A H and Walker A L 2016 Machine learning in geosciences and remote sensing *Geosci. Front.* **7** 3–10
- [41] Belgiu M and Drăguț L 2016 Random forest in remote sensing: a review of applications and future directions *ISPRS J. Photogramm. Remote Sens.* **114** 24–31
- [42] Yanjun S, Guo Q, Qin M and Wenkai L 2015 SRTM DEM correction in vegetated mountain areas through the integration of spaceborne LiDAR, airborne LiDAR and optical imagery *Remote Sens.* **7** 11202–25
- [43] Brieman L 2001 Random forests *Mach. Learn.* **45** 5–32
- [44] Biau G and Scornet E 2016 A random forest guided tour *Test* **25** 197–227
- [45] Raschka S, Patterson J and Nolet C 2020 Machine learning in python: main developments and technology trends in data science, machine learning and artificial intelligence *Information* **11** 193
- [46] Maune D F 2007 *Digital Elevation Model Technologies and Applications: The DEM Users Manual* (Maryland: ASPRS Publications)
- [47] Baltzer H 2001 Forest mapping and monitoring with interferometric synthetic aperture radar (InSAR) *Prog. Phys. Geogr.* **25** 159–77
- [48] Solberg S, Astrup R, Gobakken T, Næsset E and Weydahl D J 2010 Estimating spruce and pine biomass with interferometric X-band SAR *Remote Sens. Environ.* **114** 2353–60

- [49] Potapov P et al 2021 Mapping global forest canopy height through integration of GEDI and Landsat data *Remote Sens. Environ.* **253** 112165
- [50] Dubayah R et al 2020 The global ecosystem dynamics investigation: High-resolution laser ranging of the earth's forests and topography *Sci. Remote Sens.* **1** 100002
- [51] Potapov P, Hansen M C, Kommareddy I, Kommareddy A, Turubanova S, Pickens A, Adusei B, Tyukavina A and Ying Q 2020 Landsat analysis ready data for global land cover and land cover change mapping *Remote Sens.* **12** 426
- [52] Buchhorn M, Smets B, Bertels L, Roo D, Lesiv M and Tsendbazar N E 2020 Copernicus global land service: Land cover 100m: collection 3: epoch 2015. globe (version v3.0.1) *Zenodo* (available at: <https://doi.org/10.5281/zenodo.3939038>)
- [53] Neuenschwander A et al 2021 ATLAS/ICESat-2 L3A land and vegetation height, version 4 *Report* (NASA National Snow and Ice Data Center)
- [54] NSIDC 2021 ATL08 product data dictionary *Report* (NASA National Snow and Ice Data Center Distributed Active Archive Center) (Accessed 3 August 2021)
- [55] Neuenschwander A, Guenther E, White J C, Duncanson L and Montesano P 2020 Validation of ICESat-2 terrain and canopy heights in boreal forests *Remote Sens. Environ.* **251** 112110
- [56] Malambo L and Popescu S C 2021 Assessing the agreement of ICESat-2 terrain and canopy height with airborne lidar over us ecozones *Remote Sens. Environ.* **266** 112711
- [57] Lindsay J B 2016 Whitebox GAT: a case study in geomorphometric analysis *Comput. Geosci.* **95** 75–84
- [58] Bondarenko M, Kerr D, Sorichetta A and Tatem A (WorldPop) 2020 Census/projection-disaggregated gridded population datasets for 51 countries across sub-Saharan Africa in 2020 using building footprints (University of Southampton) (available at: <https://doi.org/10.5258/SOTON/WP00682>)
- [59] Nelson A, Weiss D J, Etten J van, Cattaneo A, McMenomy T S and Koo J 2019 A suite of global accessibility indicators *Sci. Data* **6** 266
- [60] NASA 2017 Earth at night: flat maps 2016 (available at: <https://earthobservatory.nasa.gov/features/NightLights/page3.php>)
- [61] Marconcini M et al 2020 Outlining where humans live, the world settlement footprint 2015 *Sci. Data* **7** 242
- [62] Florczyk A J et al 2019 Description of the GHS urban centre database 2015 *Report* (Luxembourg: Publications Office of the European Union)
- [63] Brovelli M and Zamboni G 2018 A new method for the assessment of spatial accuracy and completeness of openstreetmap building footprints *ISPRS Int. J. Geo-Inf.* **7** 289
- [64] Gallant J 2011 Adaptive smoothing for noisy DEMs *Geomorphometry* **2011** 7–9 (available at: https://gis-lab.info/docs/gallant2011_adaptive_smoothing_for_noisy_dems.pdf)
- [65] Tomasi C and Manduchi R 1998 Bilateral filtering for gray and color images *6th Int. Conf. on Computer Vision (IEEE Cat. No. 98CH36271)* (IEEE) pp 839–46
- [66] Reuter H I, Hengl T, Gessler P and Soille P 2009 Preparation of DEMs for geomorphometric analysis *Dev. Soil Sci.* **33** 87–120
- [67] Stevenson J A, Sun X and Mitchell N C 2010 Despeckling SRTM and other topographic data with a denoising algorithm *Geomorphology* **114** 238–52
- [68] Milledge D G, Lane S N and Warburton J 2009 The potential of digital filtering of generic topographic data for geomorphological research *Earth Surf. Process. Landf.* **34** 63–74
- [69] Lindsay J B, Francioni A and Cockburn J M H 2019 LiDAR DEM smoothing and the preservation of drainage features *Remote Sens.* **11** 1926
- [70] Van Nieuwenhuizen N, Lindsay J B and DeVries B 2021 Smoothing of digital elevation models and the alteration of overland flow path length distributions *Hydrol. Process.* **35** e14271
- [71] Nardi F, Annis A, Di Baldassarre G, Vivoni E R and Grimaldi S 2019 GFPLAIN250m, a global high-resolution dataset of earth's floodplains *Sci. Data* **6** 180309
- [72] Neuenschwander A and Pitts K 2019 The ATL08 land and vegetation product for the ICESat-2 mission *Remote Sens. Environ.* **221** 247–59
- [73] Tian X and Shan J 2021 Comprehensive evaluation of the ICESat-2 ATL08 terrain product *IEEE Trans. Geosci. Remote Sens.* **59** 8195–209
- [74] Liu A, Cheng X and Chen Z 2021 Performance evaluation of GEDI and ICESat-2 laser altimeter data for terrain and canopy height retrievals *Remote Sens. Environ.* **264** 112571