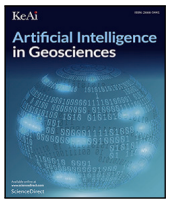
Artificial Intelligence in Geosciences 6 (2025) 100141

Contents lists available at ScienceDirect

Artificial Intelligence in Geosciences

journal homepage: www.keaipublishing.com/en/journals/artificial-intelligence-in-geosciences



Original research articles

Explaining machine learning models trained to predict Copernicus DEM errors in different land cover environments

Michael Meadows *, Karin Reinke, Simon Jones

Geospatial Science, RMIT University, Melbourne, 3000, VIC, Australia

ARTICLE INFO

Keywords:
Topography
Explainability
Interpretability
XAI
SHAP
Ensemble

ABSTRACT

Machine learning models are increasingly used to correct the vertical biases (mainly due to vegetation and buildings) in global Digital Elevation Models (DEMs), for downstream applications which need "bare earth" elevations. The predictive accuracy of these models has improved significantly as more flexible model architectures are developed and new explanatory datasets produced, leading to the recent release of three model-corrected DEMs (FABDEM, DiluviumDEM and FathomDEM). However, there has been relatively little focus so far on explaining or interrogating these models, especially important in this context given their downstream impact on many other applications (including natural hazard simulations). In this study we train five separate models (by land cover environment) to correct vertical biases in the Copernicus DEM and then explain them using SHapley Additive exPlanation (SHAP) values. Comparing the models, we find significant variation in terms of the specific input variables selected and their relative importance, suggesting that an ensemble of models (specialising by land cover) is likely preferable to a general model applied everywhere. Visualising the patterns learned by the models (using SHAP dependence plots) provides further insights, building confidence in some cases (where patterns are consistent with domain knowledge and past studies) and highlighting potentially problematic variables in others (such as proxy relationships which may not apply in new application sites). Our results have implications for future DEM error prediction studies, particularly in evaluating a very wide range of potential input variables (160 candidates) drawn from topographic, multispectral, Synthetic Aperture Radar, vegetation, climate and urbanisation datasets.

1. Introduction

Accurate topography datasets estimating "bare earth" ground elevations are crucial inputs to a wide range of geoscience applications, including natural hazard modelling (Sampson et al., 2015; Brock et al., 2020), above-ground biomass mapping (Dubayah et al., 2020) and ecological simulations (Moudrý et al., 2018). Airborne laser altimetry (LiDAR) surveys are ideal (Hancock et al., 2021) but coverage remains limited due to their high cost (Pronk et al., 2024). In many parts of the world, the only options are the free, global Digital Elevation Models (DEMs) derived from spaceborne sensors, such as the widely-used SRTM DEM (Farr et al., 2007) and the more recent Copernicus DEM (Fahrland et al., 2022). However, these suffer from well-known vertical accuracy issues relating to an inability to penetrate vegetation canopies (Brown et al., 2010; Martone et al., 2012) and resolve steep slopes (Liu et al., 2019; Li et al., 2022).

Given that these vertical errors are at least partially dependent on the local land cover and terrain conditions, there has been a sustained research effort to develop correction models (using relevant datasets such as canopy heights and slope) (Okolie et al., 2024a). This initially focused on vertical errors due to vegetation (Baugh et al., 2013; O'Loughlin et al., 2016) (taking advantage of early forest canopy height maps produced by Lefsky (2010) and Simard et al. (2011)), and sensor motion/alignment or data processing issues (Yamazaki et al., 2017). These early models tended to be relatively simple (e.g. linear regression and look-up tables), achieving only modest prediction performance but having the advantage of being easy to understand and interrogate.

More recently, a variety of machine learning algorithms have been trialled, including artificial neural networks (Wendi et al., 2016; Kulp and Strauss, 2018), random forests (Chen et al., 2020; Hawker et al., 2022), convolutional neural networks (Meadows and Wilson, 2021; Nguyen et al., 2022) and gradient tree boosting (Dusseau et al., 2023; Okolie et al., 2024a). These models are significantly more flexible (in some cases having millions of learnable parameters), meaning they are better able to represent non-linear patterns and take advantage of large data volumes (Qiu et al., 2022), and have proven to be more effective in predicting DEM errors.

* Corresponding author.
E-mail address: michael.meadows@student.rmit.edu.au (M. Meadows).

https://doi.org/10.1016/j.aiig.2025.100141
Received 11 February 2025; Received in revised form 9 May 2025; Accepted 29 June 2025
Available online 15 July 2025
2666-5441/© 2025 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

However, this increased predictive power has come at the expense of model explainability or interpretability. In other words, it is difficult to understand the patterns these models have learned in order to verify whether they make sense or may fail under certain edge cases (i.e. making predictions beyond the input data distributions encountered during training). This is an important consideration if model outputs are to be used in real-world applications potentially affecting people's lives or livelihoods (such as natural hazard modelling) — especially pertinent now, given the recent release of three machine learning model-corrected DEMs: FABDEM (random forest, global coverage) (Hawker et al., 2022), DiluviumDEM (gradient tree boosting, coastal zones only) (Dusseau et al., 2023), and FathomDEM (hybrid vision transformer, global coverage) (Uhe et al., 2025).

In past studies using machine learning to predict DEM errors, any discussion of model explainability has generally been limited to reporting input variable importance scores (Liu et al., 2021; Li et al., 2023a,b; Chen et al., 2024), providing insight into which variables the models relied on most but not their influence on predictions. Okolie et al. (2024a) recently went a step further, presenting partial dependence plots (Hastie et al., 2009) for the highest-ranked input variables, showing how each variable influenced the model's predictions. However, they considered only topographic input variables (i.e. nothing related to land cover) and limited their scope to agricultural land in Cape Town (South Africa). To the best of our knowledge, no study has yet rigorously explored model explainability for a wide range of potentially-relevant input variables (topographic, land cover, climate and urban) or assessed how this might vary for different land cover conditions.

In the research presented here, we use a diverse global dataset (65 sites) to train separate prediction models for different land cover groupings and then apply a state-of-the-art model explainability method – SHapley Additive exPlanation (SHAP) values (Lundberg and Lee, 2017) – to interrogate and compare them. Our primary objectives are to: (1) assess input variable importance rankings and how these vary by land cover, (2) where alternative datasets are available (e.g. we consider three global forest canopy height maps), identify which is most useful, and (3) demonstrate the value of evaluating the actual patterns learned (using SHAP dependence plots). Our results will directly inform future DEM error prediction studies by shortlisting and ranking input variables (from a very large candidate set), distinguishing between different training environments (by land cover) to provide extra nuance. More generally, we aim to promote more rigorous interrogation and explanation of the powerful but opaque machine learning models increasingly relied on in many geoscientific fields, as a complement to the standard performance evaluation.

2. Data and methods

2.1. Data processing

Data preparation for the modelling involved calculating maps of DEM errors (using higher-accuracy reference topography data), collating datasets representing potentially-relevant explanatory variables, resampling these to match the DEM grid resolution and alignment, and defining modelling subsets based on land cover and building footprints. We addressed potential issues relating to spatial autocorrelation (Legendre and Fortin, 1989; Ploton et al., 2020) by defining spatial blocks for each study site (Roberts et al., 2017; Valavi et al., 2019), with all subsequent sampling (for cross-validation and model evaluation) made at the block level (rather than individual grid cells). This overall workflow is summarised in Fig. 1 and each step is described in more detail in the following sections.

2.1.1. Target variable - DEM errors

Of the various global DEMs available, we selected the 1 arc-second (≈30 m) Copernicus DEM (GLO-30, DGED format), based on its currency (raw data collected in 2010–2015), independent accuracy assessments suggesting relatively narrow error distributions (Meadows et al., 2024; Bielski et al., 2024), and its use as the basis for the FABDEM product (providing a useful reference when evaluating our model performance). To determine the vertical error affecting each GLO-30 grid cell (taken here to mean its deviation from the bare earth ground elevation), higher-accuracy terrain elevation data are needed.

For this, we collated 65 Digital Terrain Models (DTMs) derived from airborne laser altimetry (LiDAR) surveys in diverse environments around the world (Fig. 2 and Table S1, Supplementary Material). We processed these to match the GLO-30 grid alignment and vertical datum (EGM2008), and then calculated grids of vertical error values by subtracting reference DTM values from GLO-30 (such that positive errors result where GLO-30 is above the true ground surface). These vertical error values are the prediction target for the machine learning models developed and explained in later stages.

2.1.2. Input variables - explanatory factors

Preparing input variable data involved three steps: (1) identifying potentially-relevant variables (based on a survey of the literature), (2) collating global datasets representing those variables, and (3) resampling all candidate datasets to align with the GLO-30 grid. In the first step, our literature review considered previous modelling studies (Olajubu et al., 2021; Hawker et al., 2022; Pimenova et al., 2022; Li et al., 2023a,b; Dusseau et al., 2023; Chen et al., 2024; Okolie et al., 2024a) as well as general assessments of DEM error distributions (Wessel et al., 2018; Hawker et al., 2019; Uuemaa et al., 2020; Li et al., 2022; Trevisani et al., 2023; Okolie et al., 2024c; Meadows et al., 2024; Bielski et al., 2024; Guth et al., 2024). This resulted in six broad groups of candidate variables: topography, vegetation, climate, urbanisation, multispectral imagery and Synthetic Aperture Radar (SAR). Note that we considered studies looking at the SRTM/NASADEM DEMs too, since they are based on the same remote sensing approach (Interferometric Synthetic Aperture Radar) as GLO-30 (albeit using a different radar wavelength band).

In the second step, we surveyed freely-available datasets representing each of the shortlisted variables, with three criteria in mind: global (or near-global) coverage, grid spacing (no coarser than 1 km, climate variables excepted) and temporal relevance (as close as possible to the GLO-30 data collection period: Dec 2010 to Jan 2015). Satisfying this third criterion often involved a trade-off: widening the temporal window increased data availability (allowing coverage across all sites) but made the resulting dataset less representative of conditions at the time of the GLO-30 data collection.

All candidate datasets eventually considered are briefly summarised in Table 1. More detailed descriptions of each dataset and its processing are provided in Table S2 and Text S1 (Supplementary Materials), including the temporal adjustment made to the canopy coverage map (Hansen et al., 2013) and the modified cloud mask used for the Landsat multispectral imagery. We note also that some of the specific datasets considered here have not yet been evaluated for DEM error prediction (to the best of our knowledge), including new global maps of forest canopy heights (Tolan et al., 2024), building heights (Pesaresi and Politis, 2023), and SAR backscatter intensity from PALSAR-2 (Shimada et al., 2014).

Clearly, many of these datasets – especially alternative products for the same variable – are likely to be highly correlated, with an obvious example being the three global DEMs (GLO-30, NASADEM and AW3D30). Our rationale for including all of these **initially** is that the machine learning algorithm used here – XGBoost (Chen and Guestrin, 2016), a gradient tree boosting model – is robust to multicollinearity in terms of predictive performance (Climent et al., 2019), and we speculate that alternative datasets (even if highly correlated) may

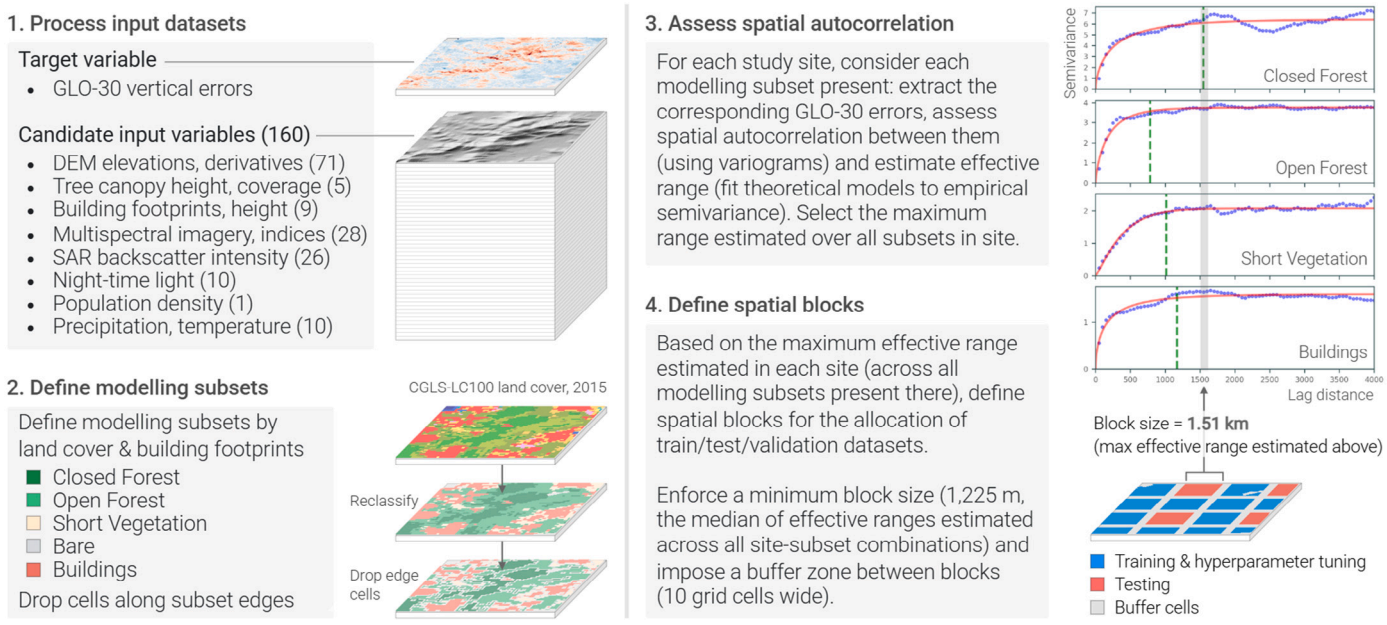


Fig. 1. Data processing workflow, with values in parentheses (under step 1) denoting the number of input variables processed in each category.

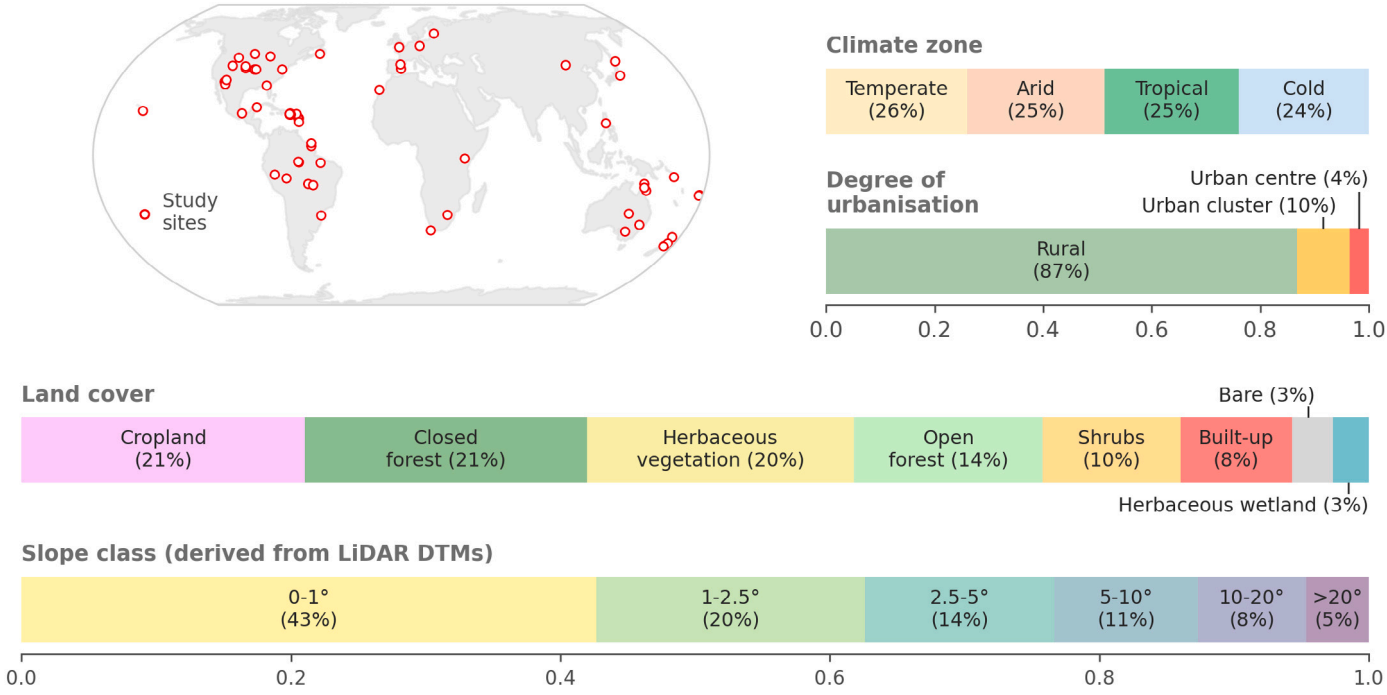


Fig. 2. Study site locations and coverage distributions according to climate zone (Beck et al., 2018), degree of urbanisation (Schiavina et al., 2023b), land cover class (Buchhorn et al., 2020), and slope class (derived from the reference DTMs).

provide some useful local nuance. These three global DEMs are derived from different sensing systems – X-band radar (GLO-30), C-band radar (NASADEM) and stereoscopic imagery (AW3D30) – and may each be preferred in specific contexts (e.g. NASADEM appears to resolve ground elevations under closed forest better than the other two; Meadows et al., 2024).

Correlated variables do pose challenges when it comes to model explainability though — potentially sharing variable importance credit amongst them, misrepresenting their actual importance (Basu and Maji, 2022) and complicating the calculation of Shapley values (Aas et al., 2021). We address this by developing a set of simplified models (with the most informationally-redundant input variables removed), supporting more clear explainability insights while still enabling an assessment of the potential value of using similar variables to provide local nuance.

The third step in data processing was to resample all input datasets to align with the 1 arc-second GLO-30 grid. Most datasets represented a single point/period in time, such that resampling was only required in the spatial dimension (taking the mean for higher-resolution datasets and using bilinear interpolation elsewhere). For datasets with multiple time-steps available (multispectral imagery, SAR backscatter intensity, climate variables and night-time light), we resampled along the temporal dimension too (taking the median value for datasets potentially

Table 1
Summary of global datasets from which our candidate input variables were derived.

Dataset	Ref.
Copernicus DEM, v2022 (elevations, quality layers[a], derivatives[b])	Fahrland et al. (2022)
NASADEM, v1 (elevations)	Crippen et al. (2016)
ALOS World 3D 30 m, AW3D30, v4.0 (elevations)	Tadono et al. (2016)
Landsat 5 ETM, 7 ETM+, 8 OLI/TIRS, merged (surface reflectance, surface temperature, spectral indices)	NA
Sentinel-1 Synthetic Aperture Radar (backscatter intensity, VV & VH)	NA
PALSAR-2 ScanSAR Level 2.2 (backscatter intensity, HH & HV)	NA
PALSAR-2 Yearly Mosaic, v2 (backscatter intensity, HH & HV)	Shimada et al. (2014)
Global Forest Canopy Height, 2019, GLAD (forest canopy heights)	Potapov et al. (2021)
High resolution canopy height model, v1, ETH (forest canopy heights)	Lang et al. (2023)
High Resolution Canopy Height Maps, Meta (forest canopy heights)	Tolan et al. (2024)
Global Forest Change, v1.10 (forest cover and gain/loss information)	Hansen et al. (2013)
Monthly DNB Composites, v1 (night-time light radiance)	Elvidge et al. (2021)
Black Marble, VNP46A2 (night-time light radiance)	Román et al. (2018)
Global Human Settlement Layer, R2023A: Building Height	Pesaresi and Politis (2023)
Global Human Settlement Layer, R2023A: Built-up Surface	Pesaresi (2023)
Global Human Settlement Layer, R2023A: Population	Schiavina et al. (2023a)
World Settlement Footprint 2015, v2 (human settlement mask)	Marconcini et al. (2020)
World Settlement Footprint 3D (building area, volume, height, fraction)	Esch et al. (2022)
ERA5-Land Monthly Aggregated (precipitation, temperature)	Muñoz-Sabater et al. (2021)
TerraClimate (precipitation, temperature)	Abatzoglou et al. (2018)

[a] Supplementary rasters provided with each Copernicus DEM tile: Editing Mask (EDM), Filling Mask (FLM), Height Error Mask (HEM), Water Body Mask (WBM).
[b] Evaluated using WhiteboxTools (Lindsay, 2016), v2.3.6: slope, aspect, minimum, maximum, percentile, Topographic Position Index (TPI), Terrain Ruggedness Index (TRI), Vector Ruggedness Measure (VRM), Average Normal Vector Angular Deviation (ANVAD), Total Curvature (TC).

subject to outliers, such as surface reflectance, and the mean otherwise). Finally, this stack of spatially-aligned grids was converted to a tabular format (each row representing a particular GLO-30 grid cell) for input to the machine learning models.

2.1.3. Defining modelling subsets based on land cover and building footprints

Our objective is a comparative explanation of error prediction models trained on different subsets of the available data, filtered on the basis of land cover. For this, we use the 100 m resolution land cover map from the Copernicus Global Land Service (version 3, 2015 epoch) (Buchhorn et al., 2020), based on its verified accuracy (Tsendbazar et al., 2021) and temporal relevance to GLO-30. Grouping individual land cover classes together, we define five subsets: Bare ("Bare/sparse vegetation"), Buildings ("Urban/built up" cells which intersect with building footprints; Marconcini et al., 2020), Short Vegetation ("Shrubs", "Herbaceous vegetation", "Cultivated and managed vegetation/agriculture" and "Herbaceous wetland"), Open Forest (all six "Open forest" classes) and Closed Forest (all six "Closed forest" classes). To avoid mixed land cover classes, we exclude grid cells along the edges of each modelling subset, as indicated in Fig. 1.

2.1.4. Accounting for spatial autocorrelation

At various stages of machine learning model development (variable selection, hyperparameter tuning and performance evaluation), the full dataset is split into training and validation/testing sets. A key assumption here is that these separate samples are independent of each other (Roberts et al., 2017). This requires careful consideration when working with geospatial data due to spatial autocorrelation, whereby locations near each other will tend to be more closely related than locations further apart (Tobler, 1970; Legendre and Fortin, 1989). If a purely random split of individual data points (best practice in many machine learning applications) is used with spatial data, this is likely to result in training and validation/testing points being spatially adjacent (and thus at least somewhat related) to each other (Karasiak et al., 2022).

If present and not accounted for, this spatial dependency between training and validation/test samples results in significant overestimates of the model's predictive performance (Schratz et al., 2019; Ploton et al., 2020) and compromises model inference (Le Rest et al., 2014; Roberts et al., 2017). Various methods to address this have been proposed (Telford and Birks, 2009; Brenning, 2012; Roberts et al., 2017), with the common objective of spatially separating train and test data. We adopt the spatial blocking approach outlined by Roberts et al. (2017) and illustrated in Fig. 1. For each study site (i.e. reference DTM extent), we define a grid of regular blocks and systematically allocate (Valavi et al., 2019) a quarter of these as test data (unseen during model training and used to evaluate final models later).

The spatial block size appropriate to each study site is estimated by calculating variograms for our target variable (GLO-30 errors), as recommended in the literature (e.g. Valavi et al., 2019; Ploton et al., 2020). Following Hawker et al. (2018), we calculate separate variograms for each modelling subset present in each study site: filtering the site's GLO-30 error grid by the selected subset, fitting a theoretical model to the empirical semivariance, and extracting the effective range (the distance at which spatial autocorrelation tapers off). Variograms are calculated using the `SciKit-GStat` Python package (version 1.0.16) (Mälicke, 2022), assuming stationarity and isotropy for simplicity, as done in many previous studies of DEM error patterns (e.g. Hawker et al., 2018, 2019). To ensure spatial blocks large enough to mitigate spatial autocorrelation (Roberts et al., 2017) in a given site, we specify block size based on the maximum range estimated across all subsets present there, as well as enforcing a minimum block size (1225 m, the median range across all site-subset combinations). Furthermore, we impose buffer zones (ten grid cells wide) between all generated blocks to increase the spatial separation of samples (Valavi et al., 2019).

2.2. Predictive model development

We used a specific gradient tree boosting algorithm – XGBoost (Chen and Guestrin, 2016) – for all predictive modelling, training for each land cover-based subset a "full" model (using most of the input variables found to be relevant) and a "simple" model (with the most redundant variables removed). The following sections explain each model development step in more detail.

2.2.1. Choice of machine learning algorithm

We selected XGBoost (Chen and Guestrin, 2016) for the following reasons: state-of-the-art performance on many tabular data problems, relatively fast compared with more complex models, robustness to overfitting, native handling of missing data, and capacity to capture non-linear patterns and variable interaction effects (Chen and Guestrin, 2016; Bentéjac et al., 2021; Qiu et al., 2022). Briefly, gradient tree boosting algorithms such as XGBoost use a sequential ensemble of shallow decision trees (so-called "weak learners"), in which each new tree learns to address the error residuals left by previous trees (Natekin and Knoll, 2013). For further details, interested readers are referred to Bentéjac et al. (2021) (for gradient boosted trees in general) and Chen and Guestrin (2016) (for XGBoost in particular).

2.2.2. Developing initial models using all relevant input variables

For each of the five modelling subsets defined above (based on land cover), we started by training an initial model (using all relevant input variables) in three steps: (1) filter available training data (as defined using spatial blocks) by the corresponding land cover group, (2) identify all input variables relevant to this subset using the `powershap` Python package (Verhaeghe et al., 2023) (version 0.0.11), and (3) tune XGBoost hyperparameters using the `hyperopt` Python package (Bergstra et al., 2013) (version 0.2.7). The latter two steps are explained in more detail below.

The `powershap` algorithm is an all-relevant variable selection approach built around the intuition that a relevant variable should have a larger influence on model predictions than a randomly-generated variable. Over a series of iterations (we used 500 for each subset), a new XGBoost model is trained on a different random sample drawn from the training block data, each time adding a new randomly-generated input variable. Looking at all iterations, the distribution of influence values for each genuine variable is then compared with the average influence of all randomly-generated variables, to judge their relevance. As the name suggests, `powershap` relies on SHapley Additive eXplanation (SHAP) values (Lundberg and Lee, 2017) to quantify influence, aligning with our model explainability approach (Section 2.4). We note that `powershap` selects **all** relevant variables, including correlated/redundant ones, which we address in the following section.

Selecting appropriate model hyperparameters (which are not learned during training and must be specified in advance) can have significant impacts on model performance (Schratz et al., 2019). We used `hyperopt` – a Bayesian optimisation algorithm – to search across predefined distributions for six XGBoost hyperparameters (`max_depth`, `subsample`, `colsample_bynode`, `gamma`, `lambda` and `eta`). This iterative process initially samples randomly from those distributions but as the process continues, past results and Bayesian methods increasingly inform the next values tested (Bergstra et al., 2013). In each iteration, we evaluate the sampled hyperparameter values using a 5-fold cross-validation (with fold allocations made on the basis of training blocks only).

2.2.3. Extracting simplified models by eliminating redundant variables

Starting with the initial models trained above (containing **all** relevant variables), we then identify and remove redundant variables to derive two models for each subset: a "full" model (only the most redundant variables are removed) and a smaller "simple" model (for which we remove even moderately redundant variables). Our main focus in this study is on the "simple" models (better suited to inference and explanation), with the "full" models used only to check the performance penalty incurred by using the smaller subset of input variables. To estimate variable redundancy, we use the novel "supervised distance" metric proposed by Qiu et al. (2022) and follow the recursive elimination process outlined in their study, in which the least informative of two redundant variables is dropped in each iteration.

The supervised distance metric quantifies how similar two variables are in terms of their contribution towards a specific prediction task (Qiu et al., 2022). As implemented in the `shap` Python package (Lundberg et al., 2020), this involves training a single-variable XGBoost model (using the first variable) to predict the target (GLO-30 errors, in our case) and then a second single-variable model (using the second variable) to predict the output of the first model. This is repeated for the converse case, after which the supervised distance is calculated (as a function of the coefficients of determination for each set of predictions), with equations and further details available in Qiu et al. (2022). The final output of this step is a matrix of pairwise supervised distances (for all variable pairs), which scale roughly from 0 (variables are perfectly redundant) to 1 (variables are completely independent).

Following Qiu et al. (2022), we then run a recursive variable elimination process, progressively reducing the number of input variables from all of those initially selected by `powershap` to one. In each iteration, the most redundant remaining variable pair is identified (using the supervised distance matrix calculated above) and the least influential of these is removed (judged by mean absolute SHAP values, recalculated after each iteration). Model performance is evaluated throughout using a withheld subset of training blocks (20%), to support later decisions about model simplification.

Given that the "simple" models use a much smaller subset of input variables than the initial models tuned above, we retune model hyperparameters (following the same process outlined in Section 2.2.2) to ensure that the simplified models' performance is not compromised by sub-optimal hyperparameters. Tuned hyperparameter values for each subset model are provided in Table S3 (Supplementary Material).

2.3. Evaluating model performance

To evaluate the performance of each model, we use its predictions to correct GLO-30 errors in the spatial blocks allocated for testing. The corrected error distributions can then be compared directly with error distributions for both GLO-30 (before correction) and FABDEM (itself based on modelled corrections to GLO-30, providing a relevant and independently validated reference standard; Bielski et al., 2024). These comparisons are made using histograms (showing the full error distributions) and appropriate error metrics.

For this, we chose three recommended in the literature: Root Mean Square Error (RMSE), 90th percentile of absolute error values (LE90), and Median Absolute Deviation (MAD), with equations given below. The commonly-used RMSE enables comparisons with past studies (Okolie et al., 2024c), LE90 focuses on the largest errors, and MAD indicates the distribution spread without making any assumptions as to its form (Höhle and Höhle, 2009).

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \Delta h_i^2} \tag{1}$$

$$\mathrm{LE90} = \mathrm{P}_{90}(|\Delta h_\mathrm{i}|) \tag{2}$$

$$\mathrm{MAD} = \mathrm{median}(|\Delta h_i - m_{\Delta h_i}|) \tag{3}$$

where Δh_i is the vertical error for a given grid cell i (GLO-30 elevation minus reference DTM elevation), $m_{\Delta h_i}$ is the median of these error values, and $P_{90}(|\Delta h_\mathrm{i}|)$ is a percentile of the absolute error values.

2.4. Explaining models using SHapley Additive exPlanation (SHAP) values

Derived from cooperative game theory (Shapley, 1953), SHapley Additive exPlanation (SHAP) (Lundberg and Lee, 2017) is a method for explaining model predictions for individual data points (GLO-30 grid cells, in this case), in terms of the contribution made to that prediction by each input variable. These local explanations can also be aggregated in different ways to provide fine-grained global summaries across the full range of input variable values. Those most relevant to our study are the SHAP summary plots (beeswarm visualisations showing the magnitude, frequency and direction of each variable's effects) and dependence plots (scatter plots of a given variable's values versus effects, optionally showing interaction effects for a second variable) (Lundberg et al., 2020).

We compute SHAP values using the TreeSHAP algorithm, optimised by Lundberg et al. (2020) for tree-based models and implemented in the `shap` Python package (version 0.45.1). As recommended in the literature (Janzing et al., 2020; Chen et al., 2023), we use the interventional (rather than path-dependent) method, which requires the provision of two datasets: background (used to compute marginal expectations) and foreground (specific predictions to be explained). To achieve reasonable computation times, we use samples for each: 5000 rows of the training data (background) and 10,000 rows of the test data (foreground), consistent with past studies (Qiu et al., 2022).

As noted above, we focus on explaining the "simple" models, given the explainability challenges involved when there are highly-correlated input variables present (Basu and Maji, 2022; Aas et al., 2021). By default, variable importance rankings are based on mean absolute SHAP values but Okeson et al. (2021) have shown that relying on a single statistic for ranking may mask other useful insights. For a better understanding of variable importance, they recommend considering additional statistics, two of which are adopted here. These are the "typical range" (difference between 5th and 95th percentiles), which highlights variables which are very important for at least some data points, and the frequency in the top three (proportion of data points for which a given variable is ranked in the top three).

3. Results

3.1. Simplifying models by eliminating redundant or spurious variables

The results of the recursive variable elimination are shown in Fig. 3. For each modelling subset, we begin with all input variables found to be relevant (left edge of plots) and then progressively remove redundant variables (moving from left to right). At each step, we track the minimum supervised distance between the remaining input variables (primary y-axis) and the simplified model's performance on a withheld set of training blocks (validation RMSE, secondary y-axis).

For each subset, we extract smaller sets of variables for "full" and "simple" models by min–max normalising the validation RMSE values across all iterations, such that they range between 0 (representing the lowest validation RMSE) and 1 (the highest validation RMSE, associated with the univariate model), and then selecting the smallest variable sets for which the normalised RMSE is less than 0.01 and 0.1 (respectively). As summarised by annotations in Fig. 3, this significantly reduces the number of input variables used (with "simple" models relying on only 12–36 variables), without substantially degrading model performance (less than 5% difference in validation RMSE reduction rates, across all subsets).

Based on an analysis of preliminary SHAP dependence plots, we also identified and removed nine potentially problematic variables from those shortlisted above. For all of these (either climate variables or counts of SAR images available by grid cell), variation between sites was significantly larger than within individual sites, meaning they could function as proxy variables identifying specific sites. While some of these variables appeared important (based on SHAP values), there was no discernible pattern to their dependence plots (see Figure S1 for examples). We speculate that they may have provided a "learning shortcut", allowing the model to memorise local biases for each study site (e.g. processing artefacts from vertical datum conversion) that would likely be spurious if applied to new sites. Removing these variables simplified the models further (11–32 variables).

3.2. Evaluating model performance (full and simple)

To evaluate model performance (on the withheld test blocks), Fig. 4 compares four distributions of vertical errors: GLO-30 before correction (magenta lines), after correction using predictions made by the "full" (light orange) and "simple" (dark orange) models, and FABDEM (teal). These vertical error distributions are summarised using histograms (left column) and by the relative change in error metrics, with reference to the original GLO-30 errors (right column). Our models performed well for all subsets (as shown by corrected error distributions more tightly centred around zero and significant reductions in error metrics), with only minor differences observed between the "full" (42–111 variables) and "simple" (11–32 variables) models.

3.3. Variable importance based on SHAP values

Fig. 5 summarises variable ranks for all shortlisted input variables (along y-axis) and the five modelling subsets considered (main columns), evaluated according to the three ranking metrics discussed in Section 2.4 (sub-columns). While some variables are consistently important across all models (especially topography derivatives), the overall impression is of significant variation between models in terms of the number of variables used (the "Bare" model is especially simple), the specific variables selected, and their relative importance across models. For example, the Bare Soil Index (BSI) was very important in the "Short Vegetation" and "Open Forest" models (where the distinction between vegetation and bare soil is especially informative), but not selected at all in the other models.

Fig. 6 uses beeswarm plots to provide more detail on the most important variables in each subset model (ranked by mean absolute SHAP value). Each dot in a given beeswarm represents a single data point (GLO-30 grid cell), showing the relative value of the selected input variable (by colour) and its impact on the final prediction made for this grid cell (by its position on the x-axis). This is a more informative reference than a standard variable importance bar chart, providing first indications of the form and strength of learned patterns (based on the direction and consistency of the horizontal colour gradient) and the distribution of SHAP values. For a specific example, consider the top-ranked variable (Bare Soil Index) in the "Short Vegetation" model (Fig. 6c), where we find evidence of a negative relationship with SHAP values. Blue points are low BSI values (i.e. higher vegetation cover) and are associated with large positive contributions to predicted error (to the right along the x-axis). On the other hand, pink points (high BSI values, more likely to be exposed soil) are clustered over a relatively narrow range of small, negative SHAP values, indicating lower-than-average predicted errors for those cells (as would be expected in relatively bare ground).

Continuing the high-level overview, Fig. 7 shows the relative importance of different variable categories (rather than individual variables) and of spatial neighbourhood sizes (used when calculating certain topographical derivatives, ranging from 3×3 to 21×21 grid cells). Comparing the relative importance of variable categories (by total mean absolute SHAP values) across all modelling subsets (Fig. 7a) reveals that some categories are consistently important (especially topography and multispectral) while others are relevant only in particular environments (e.g. urbanisation variables in the "Buildings" model). To evaluate the relative importance of the spatial extent used when processing topographical derivatives (Fig. 7b), we show total mean absolute SHAP values (by neighbourhood size) and calculate a weighted-average neighbourhood size for each modelling subset (weighted by

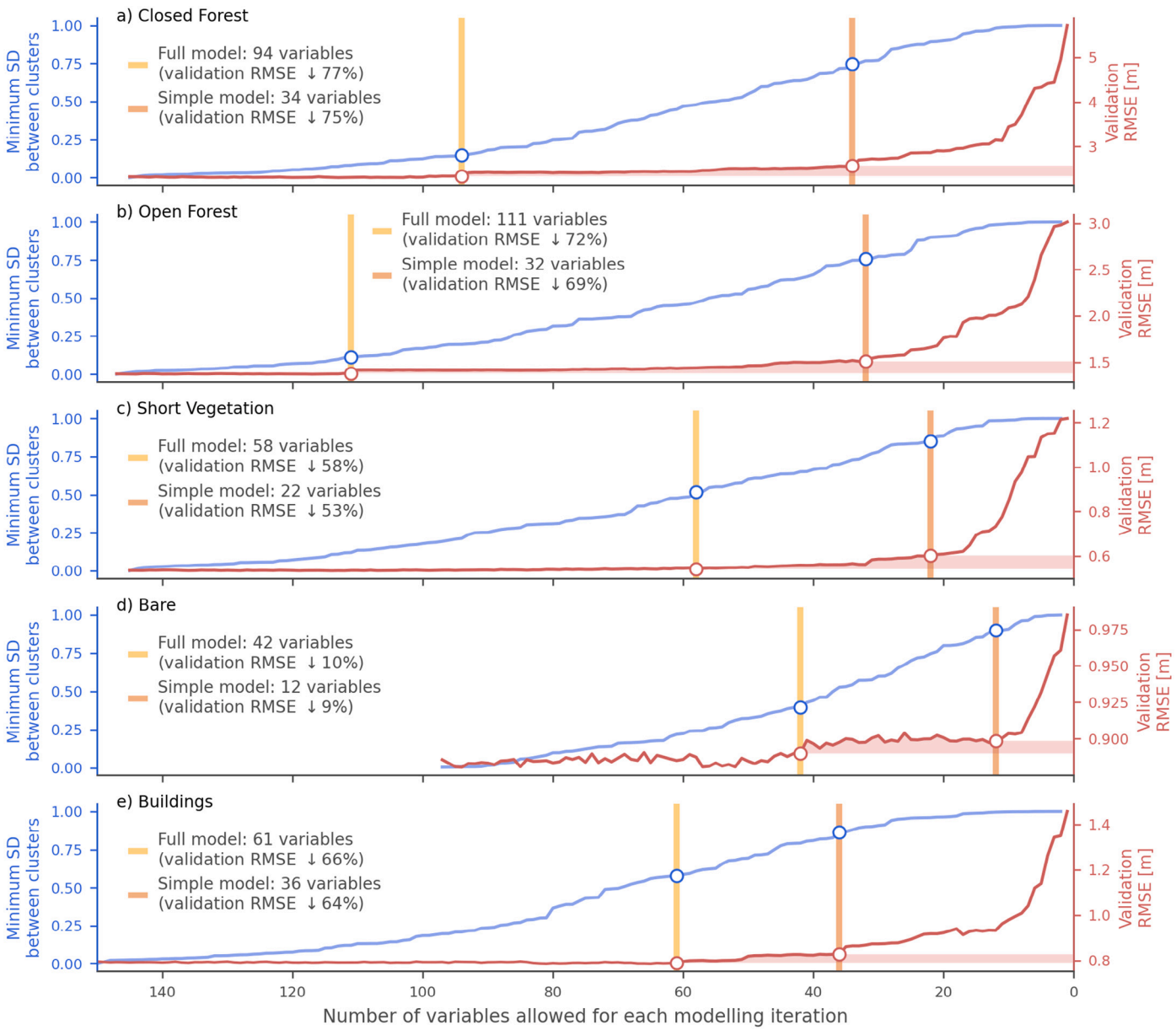


Fig. 3. Results of the recursive variable elimination for each subset model (a–e), showing the change in minimum supervised distance between variable clusters (blue line) and validation RMSE (red line). The two input variable sets extracted (full and simple) are indicated by vertical lines and the corresponding difference in validation RMSE by the horizontal pink bar.

the mean absolute SHAP values). As might be expected, most models clearly benefit from spatial context derived from a combination of small and large neighbourhoods (grid dimensions of 3–7 and 21 cells, respectively), except for the "Bare" model (for which there was apparently limited value in looking further afield than a 7×7 neighbourhood).

3.4. SHAP dependence plots for selected variables

SHAP dependence plots (Fig. 8) show the relationship between the values of a selected variable (x-axis) and the corresponding SHAP values (y-axis), revealing how that variable influences the model's predictions. This influence is expressed with reference to the model's baseline (the mean prediction over its background sample), with positive SHAP values contributing to a larger-than-average prediction and vice versa. Vertically dispersed points (i.e. where the same variable value can have different impacts) reveal interaction effects with other variables. These can be explored visually by colouring points based on the value of a second input variable, providing further insights. In interpreting these plots, note that grey histograms along the *x*-axis show variable data distributions, horizontal ticks along the *y*-axis indicate grid cells for which the selected input variable was missing, and that the interaction variable visualised (by colour) in each plot is denoted in the colourbar label.

To demonstrate the value of these dependence plots, we present a small selection in Fig. 8, focusing on the two "Forest" models (where vertical errors are largest) and the "Buildings" model (most relevant to downstream applications involving human populations). These exemplify a variety of learned patterns (expected, surprising and problematic) and include some variables identified as important in past models, such as elevation and terrain aspect. (For a more systematic overview, dependence plots for the 12 highest-ranked variables in each model are provided in Figures S2–S6.)

The learned patterns visualised in Fig. 8 take many forms and vary in their complexity. Some are mostly linear, as for the Bare Soil Index

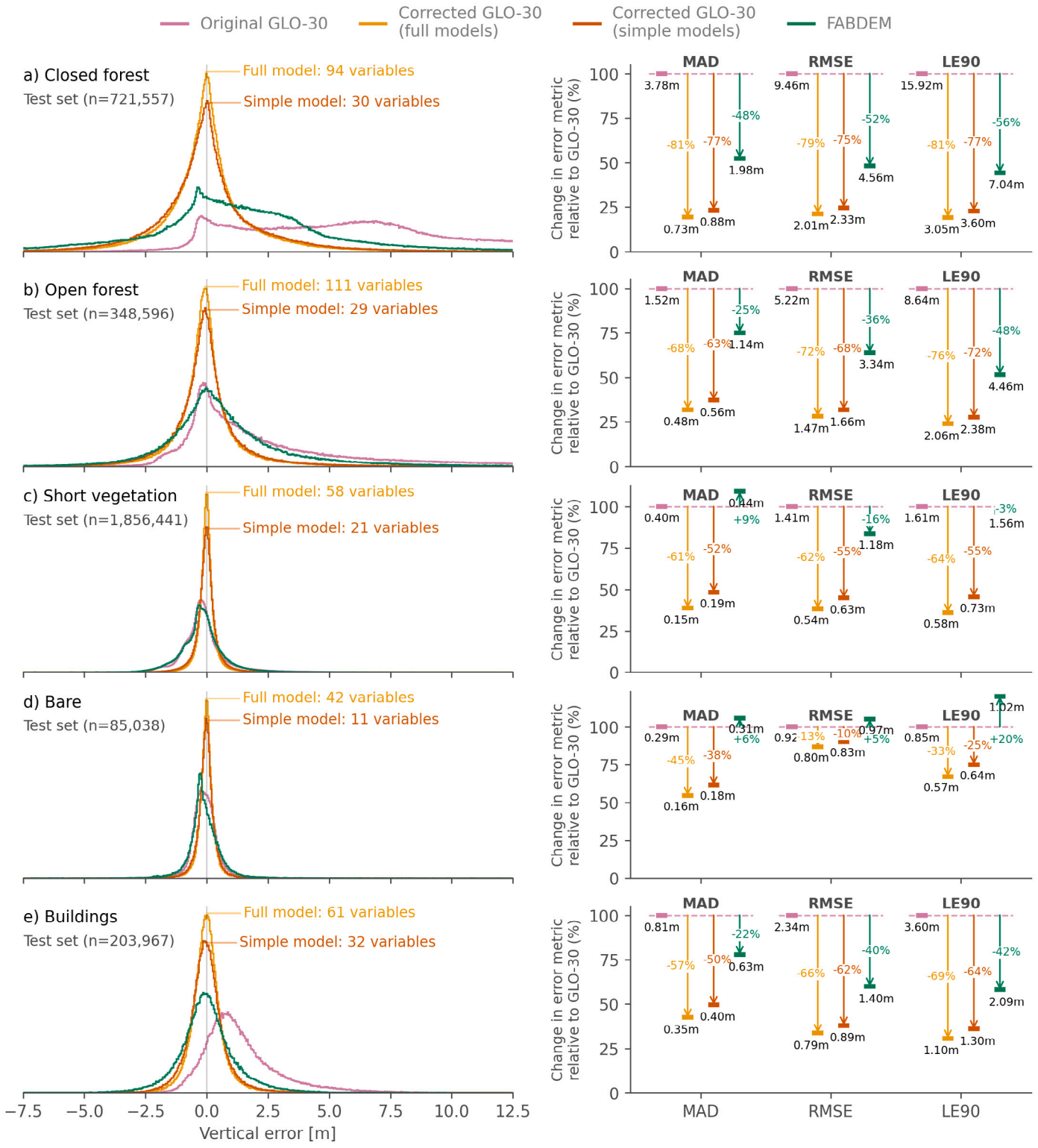


Fig. 4. Vertical error distributions (left column) for GLO-30 before and after correction (by our two models and FABDEM) and relative change in three selected error metrics (right column) to visualise the magnitude of improvements made.

(**b**) and neighbourhood elevation percentile (**i**) variables, noting strong interaction effects in the latter case (vertical colour gradients). Others are non-linear but monotonic, as seen for the Height Error Map (**f**) and the two SAR variables (**g–h**). Even more complex relationships were learned: roughly parabolic for the Normalised Difference Built-up Index (**c**), sinusoidal for terrain aspect (**e**), and increasing rapidly before returning to an asymptote of zero for elevation (**c**). Variables for which no clear pattern could be discerned (e.g. **a**) may relate to

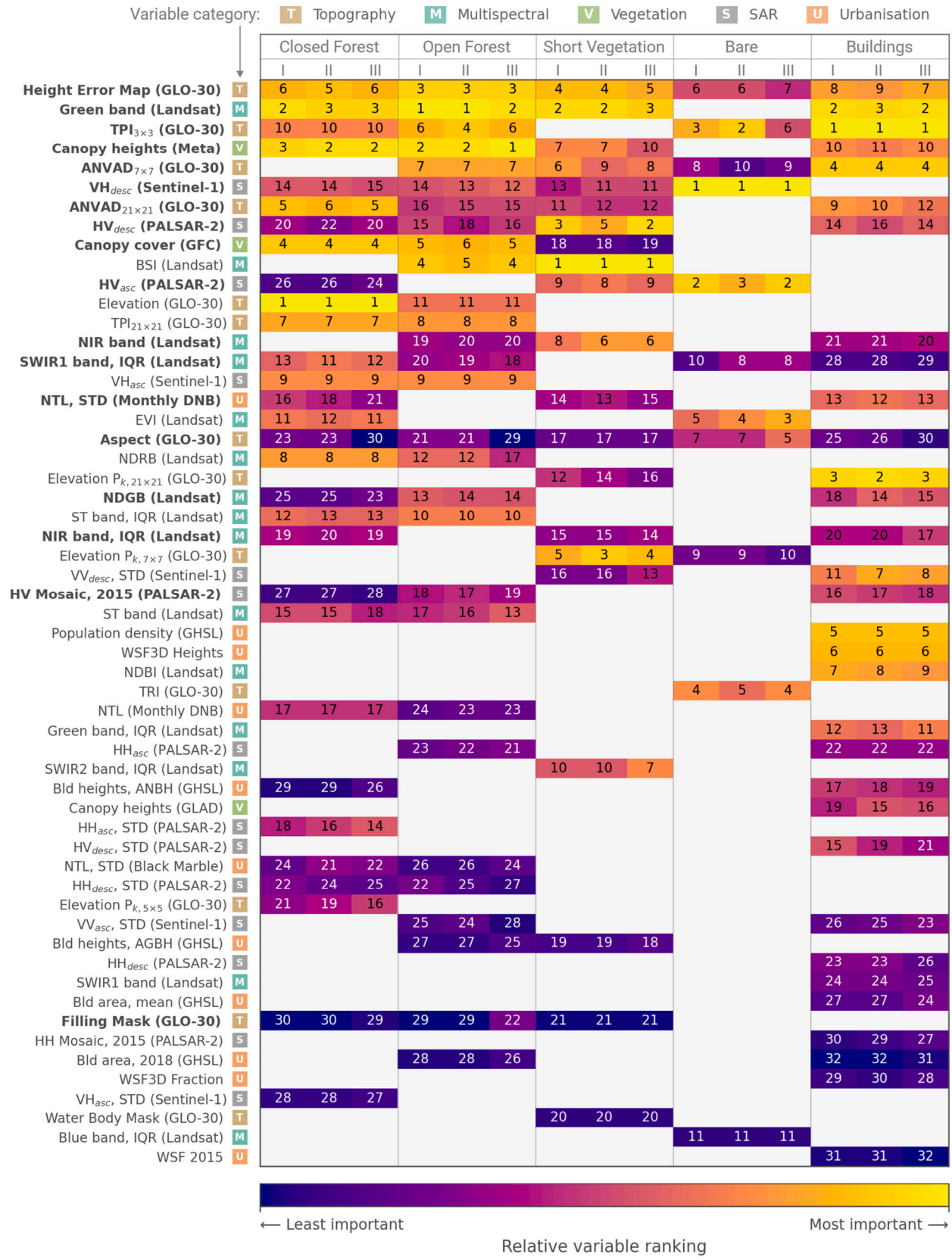


Fig. 5. Ranks for all shortlisted variables (by row) across all modelling subsets (main columns), based on three ranking metrics (by sub-column): mean absolute SHAP value (I), typical range of SHAP values (II) and frequency in top three (III). Each cell corresponds to a particular variable-model-metric combination, giving the variable's numeric rank (by that metric) and coloured according to its relative ranking (within that model), ranging from purple (low-ranked) to yellow (high-ranked), or light grey if that variable was not selected in that model. The coloured icons alongside each variable name indicate its category and variable names are in bold if they were selected for three or more subset models.

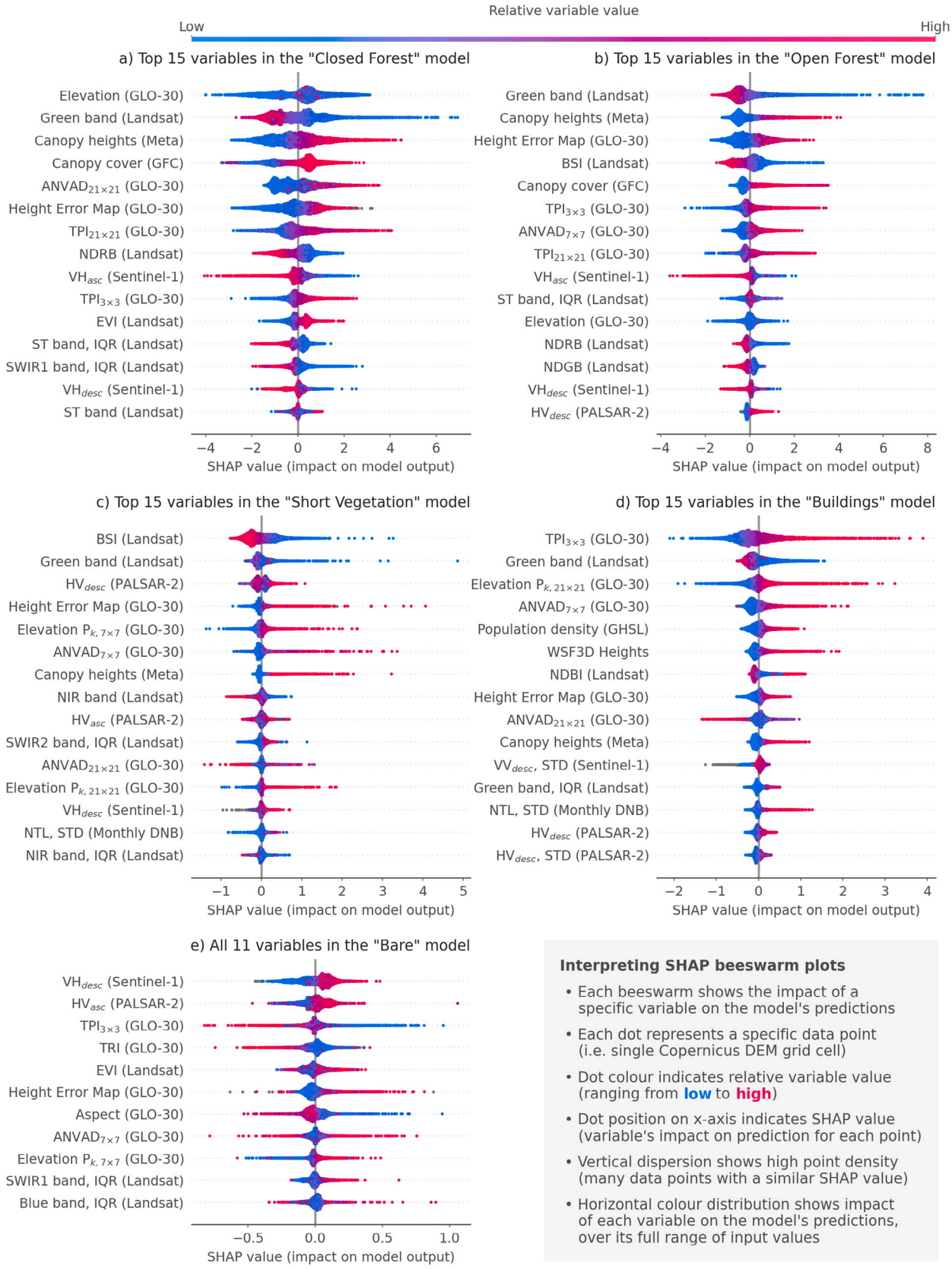


Fig. 6. Beeswarm plots for the highest-ranked variables in each model, illustrating how predictions for individual GLO-30 grid cells (represented as dots) are influenced by a given variable. The dot's colour indicates the variable's relative value, while its position on the *x*-axis reflects its impact on the model's prediction for that grid cell. Where multiple dots share similar *x*-axis positions, they stack vertically to represent density.

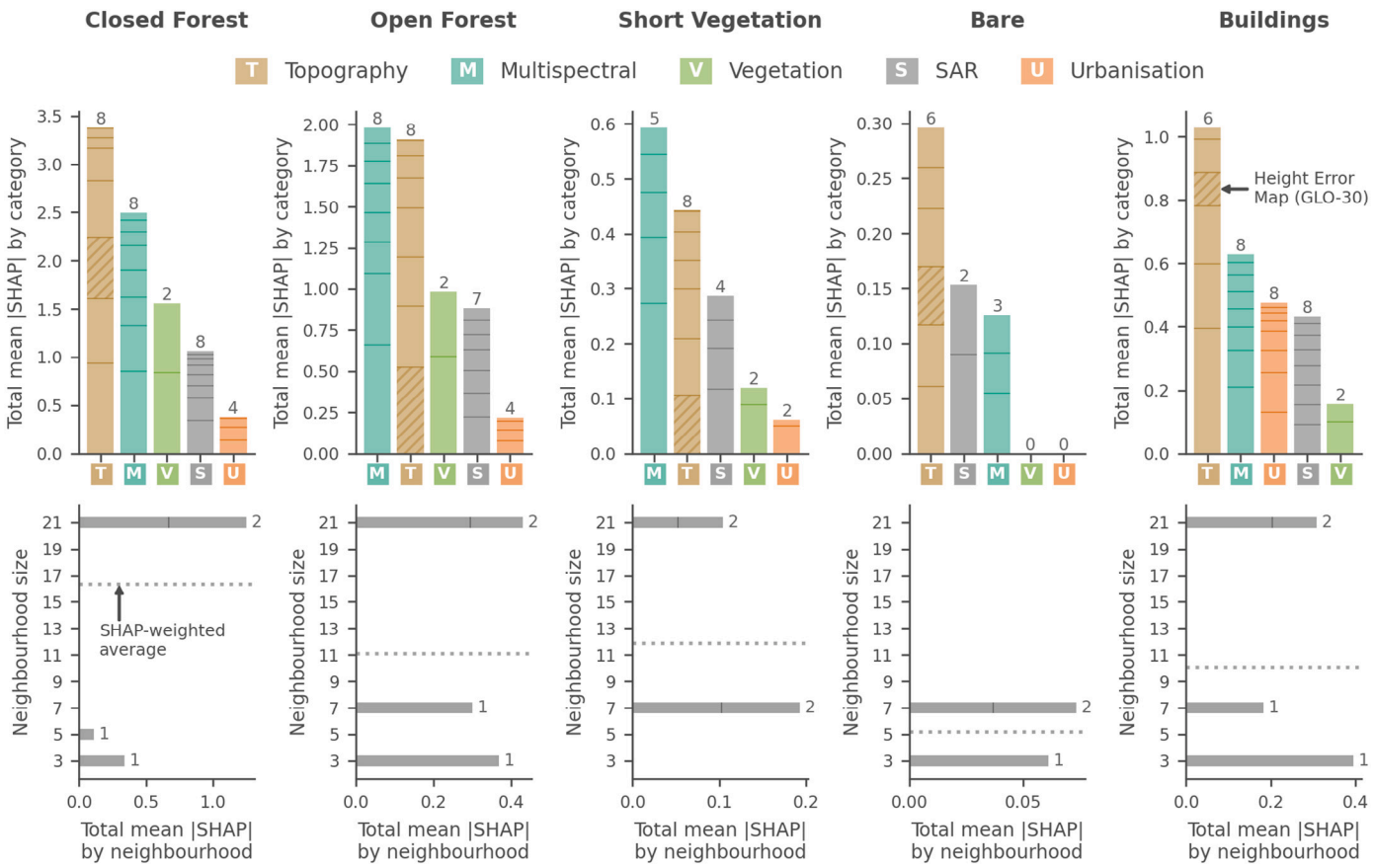


Fig. 7. Relative importance of variable categories (top row) and neighbourhood size (ranging from 3 × 3 to 21 × 21 grid cells) used when calculating topographical derivatives (bottom row), for all modelling subsets (by column). Bars represent the sum of mean absolute SHAP values for selected variables of a given category (top) or neighbourhood size (bottom, topographical derivatives only), with annotations providing the number of variables represented by each bar (also indicated by bar subdivisions). Note: hatched bars in the upper subplots indicate the Height Error Map (HEM) quality layer provided with the Copernicus DEM, rather than an actual topography derivative.

"learning shortcuts" (as discussed in Section 3.1) and were removed from the final models.

4. Discussion

4.1. Model performance

Key takeaways

- Limited improvement for "Bare" subset (errors already low)
- Significant improvements made for all other subsets, reducing RMSE by 55–75% ("simple" models) or 62–79% ("full" models)
- Relatively small performance differences between "simple" and "full" models (using 11–32 and 42–111 variables, respectively)

Explaining a predictive model is only meaningful if it has learned useful patterns relevant to new, independent test data (Roberts et al., 2017). In the work presented in this paper, we evaluated all of our models ("full" and "simple" versions for the five modelling subsets) using corresponding data from the spatial blocks set aside for testing; judging them against FABDEM as the current independently validated gold standard (Bielski et al., 2024) (Fig. 4). Model performance was generally good. With the exception of the "Bare" subset (where GLO-30 is already very accurate and not much further improvement could be made), our "full" models reduced RMSE by 62%–79% and the "simple" models by 55%–75% (surpassing the improvements made by FABDEM). This is unsurprising given that our models are trained locally on very similar data distributions (whereas FABDEM relied on global models trained over a wider range of environments), but it does suggest that our models have learned meaningful patterns worth explaining.

4.2. Differences in variable importance across subset models

Key takeaways

- Subset models vary significantly (by number of variables found relevant, specific variables selected and their relative importance)
- Topographical and multispectral variables consistently important
- Despite patchy coverage and limited use in past studies, SAR variables clearly relevant
- Neighbourhood elevation statistics always important but optimal neighbourhood size varies by subset

Comparing the models trained on the five modelling subsets ("simple" models only), we found significant variation in the number of input variables found to be relevant, the specific variables selected, and the relative importance of these variables (Fig. 5). The most complex models (29–32 variables) were for "Buildings" and the two "Forest" subsets (where vertical errors are highest) and the simplest model (11 variables) for "Bare" ground (where errors are already very low). In terms of the specific variables selected, the greatest overlap was observed between the two forest models (22 variables in common) and

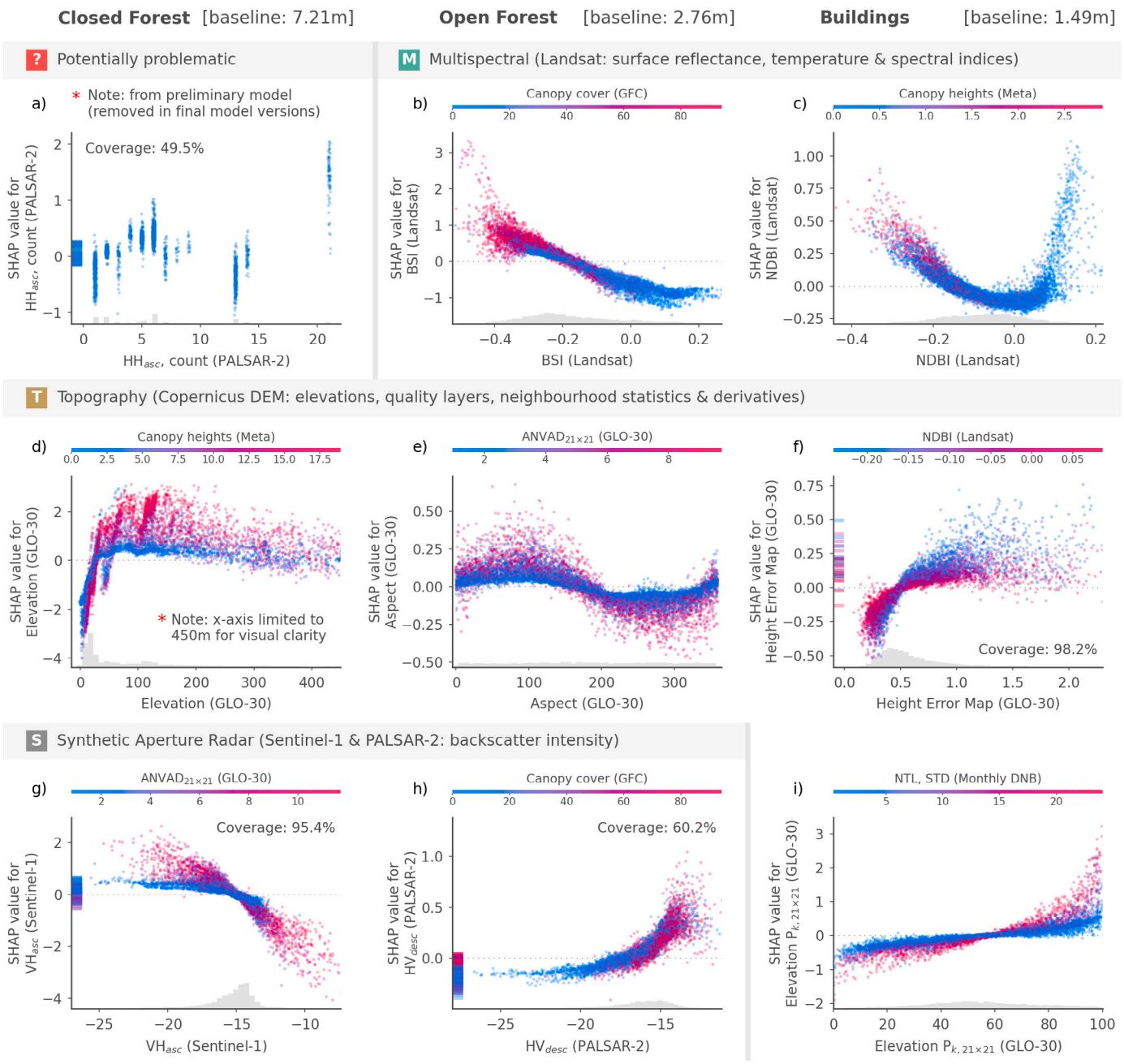


Fig. 8. Selected SHAP dependence plots for the "Closed Forest", "Open Forest" and "Buildings" models (by column, with model baseline values annotated at the top), grouped by category: problematic (a), multispectral (b–c), topography (d–f, i) and SAR (g–h). Each plot relates a specific input variable (x-axis) to its impact on the model's prediction (y-axis), with labels noting coverage (where incomplete) and points coloured to indicate interaction effects with a second input variable (see colourbar label), where relevant.

the least for the "Bare" and "Buildings" models (only five variables in common). However, variable-level analysis is complicated by cases where highly-correlated alternative datasets (e.g. three canopy height maps; Potapov et al., 2021; Lang et al., 2023; Tolan et al., 2024) were initially made available for selection.

To address this, we also evaluated importance by variable category (Fig. 7, top row), finding that topographical and multispectral variables were consistently important, while other categories varied somewhat between the model subsets. Unsurprisingly, the derived global datasets were most relevant in their target environments: the vegetation variables (forest canopy height and coverage) contributed most to the two "Forest" models and urbanisation variables (building properties, nighttime light and population density) to the "Buildings" model. The clear relevance of Synthetic Aperture Radar (SAR) data – despite patchy coverage – is noteworthy given they have not been widely used in DEM error prediction studies to date.

For certain topographical derivatives, we initially provided the models with multiple versions evaluated over ten different neighbourhood sizes (ranging from 3 × 3 to 21 × 21 grid cells), to see which extents were most informative for each model (Fig. 7, bottom row). While most models used a combination of small (3 × 3 to 7 × 7) and large (21 × 21) extents, the balance was different across models. Looking at weighted-average extents (weighted by mean absolute SHAP value, shown as dotted lines in Fig. 7), we see the largest value for "Closed Forest" and the smallest for "Bare" ground. This may relate to the obscuration or availability of bare earth elevation references — where these are rare (e.g. in dense forest), larger extents are needed, while much smaller extents are sufficient where bare earth cells are common. These findings

are relevant for future modelling efforts using spatially-structured data inputs (such as Convolutional Neural Networks), although it seems that even larger extents may be useful (four of our five models used variables evaluated over a 21 × 21 neighbourhood, the largest we considered).

4.3. Assessing fitted relationships across subset models

Key takeaways

- Patterns learned can be visualised using SHAP dependence plots and assessed against known correlations or expected causal factors
- Some patterns match those previously derived from physical principles (e.g. aspect dependency due to geolocation errors Nuth and Kääb, 2011)
- Other patterns make intuitive sense (e.g. lower predicted errors where the Bare Soil Index is high in "Forests", indicating clearings)
- Others initially surprising and may motivate future research (e.g. importance of night-time light variation in identifying tall buildings)
- Care should be taken where variables may act as proxies (for more directly-relevant factors) or relate to site-specific biases (rather than general patterns), especially if models will be applied elsewhere
- Influence of SAR backscatter intensity may differ by wavelength (L-band: vegetation canopy height/structure, C-band: imaging geometry favourability), with further research required

Going a step further than simply assigning importance scores, SHAP dependence plots reveal the patterns learned for each variable (including interaction effects with other variables). Assessing these patterns with reference to expected causal mechanisms or well-established correlations can help build trust in the model (where they seem grounded in physical reality and likely to be valid in new sites) or to identify potentially-problematic variables. The latter case includes proxy variables (correlated with variables which do have a direct relationship but are not available to the model) and spurious patterns relating to "learning shortcuts" taken by the model.

A potential example of such a spurious pattern is for the variable providing counts of PALSAR-2 HH polarisation images available (Fig. 8a). This is evaluated by grid cell but is often constant across a given study site, such that (sufficiently flexible) models might learn to associate particular image counts with site-specific elevation biases. If the modelling objective is simply to make predictions around the same set of study sites, taking advantage of an arbitrary relationship such as this may allow higher predictive accuracies. (This explains the selection and apparent importance of these variables in the first place, given that they were evaluated using withheld spatial blocks drawn from the same set of sites, where that arbitrary relationship still applied.) On the other hand, such relationships are problematic if the objective is prediction in new sites (where those arbitrary relationships are unlikely to apply) or model explanation (where our interest is in general patterns that support improved models in the future).

In other cases, the learned patterns revealed by the dependence plots make intuitive sense and provide some confidence in the model's generalisability. For example, the Bare Soil Index (BSI) has a negative, mostly linear relationship with its SHAP values in the "Open Forest" model (Fig. 8b). High BSI values indicate openings in the forest canopy, where the bare ground could be detected more easily, and are associated with lower error predictions. As BSI reduces (increasing canopy closure), SHAP values increase linearly (contributing to higher error predictions), although this relationship weakens considerably below a BSI value of around −0.3, likely relating to spectral saturation over dense forest (Gao et al., 2023).

The influence of the Normalised Difference Built-up Index (NDBI) in the "Buildings" model (Fig. 8c) is more complicated, taking a U-shaped form where larger errors are predicted when NDBI is either very low (relating to urban vegetation, see interaction effect colouring) or very high (dense urban centres). In between these two extremes, NDBI values around zero result in slightly negative SHAP values (contributing to lower-than-average error predictions). Looking at their spatial distribution, we found these grid cells were primarily roads and other spaces between buildings (low vertical errors in GLO-30).

Given the relative importance of topographical variables and their frequent use in past DEM error prediction studies, we selected four for discussion here. Elevation itself has been identified as important in many previous studies (e.g. Liu et al., 2021; Li et al., 2023a; Shen et al., 2023; Okolie et al., 2024b; Chen et al., 2024), usually explained with reference to higher elevations tending to have steeper slopes, increased forest cover and/or reduced Ground Control Point (GCP) availability (Liu et al., 2020; Nuth and Kääb, 2011). Interestingly, we found that elevation was only selected in our "Forest" models (ranking first in "Closed Forest", regardless of ranking metric), and not used at all by the other models. This suggests that the relevance of elevation to DEM errors relates more to forest canopies/structure than to slope or GCP availability, with the caveat that steep slopes (although not high elevations) are more common in forest than non-forest subsets (by a factor of more than three, for slopes above 25° in our study sites).

Looking at the dependence plot for elevation in the "Closed Forest" model (Fig. 8d), we found a particularly strong impact at very low elevations (up to around 30 m above mean sea level), above which SHAP values tend towards zero (i.e. minimal influence on the model's baseline value). Looking at other variables more directly relevant to vertical error in forests (especially canopy heights, the Height Error Map and cross-polarised SAR backscatter intensity), we found a similar pattern (i.e. those variables also increase significantly with elevation up to around 30 m before levelling off, see Figure S7). Based on this, we venture that elevation may serve in this context as a parsimonious proxy variable summarising the combined effects of numerous such variables in one predictor and that it is particularly significant at low elevations.

The next topographical variable considered here is aspect (the direction that a terrain surface faces), one of only two input variables to be selected in all models (the other being the Height Error Map, considered next). Across all models, fitted relationships for aspect are similar to that shown here for "Open Forest" (Fig. 8e) and match very closely the sinusoidal pattern that would be expected where DEM raster tiles suffer from geolocation errors (Nuth and Kääb, 2011). Consistent with that explanation, the influence of aspect is most pronounced in rough terrain (pink points), quantified here using the Average Normal Vector Angular Deviation (ANVAD) (Lindsay et al., 2019). Interestingly, previous analysis of these same sites found some evidence of small geolocation offsets but this varied by site (Meadows et al., 2024), in contrast to the consistent (albeit low-impact) pattern learned here by all subset models.

The other input variable selected in all models is the Height Error Map (HEM), one of the quality layers provided with the Copernicus DEM. Note that "error" denotes here the standard deviation of estimated elevations for each grid cell (relating to interferometric coherence and geometrical considerations) (Fahrland et al., 2022); distinct from our usage of the term in this study (difference with the bare earth elevation). We found HEM to be particularly important in the "Forest" models (where volume decorrelation of SAR signals is most problematic; Martone et al., 2012), consistent with past work in similar environments (Marešová et al., 2021). All models learned a pattern of the form shown in Fig. 8f ("Buildings" model): SHAP values initially increase rapidly, reach an inflection point at HEM values of around 0.5 m, and then gradually level off. Interestingly, the "Buildings" model

pattern reveals a strong interaction effect with NDBI: for a given HEM value, the influence on model predictions is greater for vegetated (low NDBI) than built-up (high NDBI) grid cells. This is evident in the raw data too (see Figure S8): vertical errors show no correlation at all with HEM in "Bare" ground but increasingly positive correlations as vegetation density/heights increase.

Also consistently important were variables expressing the elevation of a given grid cell relative to those in its neighbourhood, such as the Topographic Position Index (TPI) or as a percentile of neighbourhood elevations (Elevation P_k). As would be expected, higher variable values lead to higher SHAP values (i.e. cells elevated above their neighbours are more likely to contain positive vertical errors) but strong interaction effects are evident in all cases. For example, Fig. 8i shows the impact of Elevation $P_{k,21\times21}$ (cell elevation as a percentile of all elevations in its 21 × 21 grid neighbourhood) in the "Buildings" model. Especially interesting here is the strong interaction effect with the standard deviation of monthly night-time light radiance (Elvidge et al., 2021), an initially surprising finding. However, a recent study by Li et al. (2019) in urban areas found that significant temporal variation in remotely-sensed night-time light may be associated with tall buildings, which obscure light sources on/near the surface for some satellite viewing angles but not others.

The last two variables explored using dependence plots are mean backscatter intensities derived from two Synthetic Aperture Radar (SAR) missions. As noted previously, past studies on DEM error prediction have rarely made use of SAR input variables (with two recent exceptions; Dusseau et al., 2023; Uhe et al., 2025) but our results suggest they can be informative. We considered two sources, differing by radar wavelength and polarisation channels available: PALSAR-2 (L-band, 24.6 cm, HH and HV) and Sentinel-1 (C-band, 5.6 cm, VV and VH). Given the side-looking imaging geometry used for SAR, data captured on ascending and descending passes may differ significantly (i.e. for the same surface/object, viewed from different angles), particularly in steep terrain (Kellndorfer, 2019). For this reason, we considered pass directions separately when preparing input variables.

Looking first at the influence of PALSAR-2 HV backscatter intensity in the "Open Forest" model (Fig. 8h), we found that SHAP values increase exponentially with backscatter intensity, as might be expected. Due to its relatively long wavelength, the L-band PALSAR-2 radar is able to penetrate most forest canopies, resulting in significant volume scattering throughout the vertical canopy and strong cross-polarised backscatter returns from forests (Saatchi, 2019). This is consistent with the learned pattern shown in Fig. 8h — high PALSAR HV backscatter is associated with tall/dense forest canopies, where GLO-30 contains large positive errors (due to its use of a shorter-wavelength radar, less able to penetrate forest canopies).

However, for the other cross-polarised signal shown here (Sentinel-1 VH in the "Closed Forest" model, Fig. 8g), a very different relationship is learned: increasing VH backscatter intensity is associated with decreasing SHAP values (especially in rough terrain). As a reminder, Sentinel-1 uses a shorter radar wavelength (C-band, 5.6 cm), meaning shallower penetration into forest canopies, presumably providing less information about vegetation structure (Meyer, 2019). We were therefore surprised that Sentinel-1 VH backscatter turned out to be much more influential in the two "Forest" models than PALSAR-2 HV backscatter (by ranking and the SHAP value ranges evident in Fig. 8).

Closer inspection suggests that this may relate to the imaging geometry considerations faced by side-looking SAR sensors, especially over sloped terrain. Given similarities in radar wavelength and orbital inclination, backscatter intensities from Sentinel-1 (5.6 cm, 98.2° inclination) may contain information closely relevant to terrain-dependent biases affecting the TanDEM-X SAR sensors (3.1 cm, 97.4° inclination), from which GLO-30 was derived. We found that higher Sentinel-1 VH backscatter intensities are strongly associated with sensor-facing slopes (see Figure S9), where SAR-derived elevations are generally most accurate (Toutin, 2002; Shortridge and Messina, 2011), potentially explaining the negative SHAP values (lower-than-average vertical error predictions). In other words, Sentinel-1 VH backscatter intensity may function here as a proxy for favourable TanDEM-X imaging geometry: high backscatter intensities are associated with sensor-facing slopes that could be captured well by TanDEM-X too, resulting in relatively low vertical errors. This explanation is consistent with the strong interaction effects observed in Fig. 8g for terrain roughness — Sentinel-1 VH backscatter has only minor impacts on SHAP values in smooth terrain (blue points) but becomes much more significant in rough terrain (pink points).

However, this analysis is complicated by hemispheric differences in the TanDEM-X data acquisitions and potential model bias introduced by the imbalance in reference data availability (more LiDAR-derived DTMs in the northern hemisphere). Following mission specifications, the TanDEM-X satellites mapped the northern hemisphere on ascending passes and the southern when descending ("nominal" acquisitions), reversing this pattern towards the end of the mission for additional coverage of challenging terrain from the opposite viewing geometry ("crossing" acquisitions) (Rizzoli et al., 2017). Analysing TanDEM-X coverage of our study sites (using the Source Data Layers provided with the Copernicus DEM), we found that only 39% of GLO-30 grid cells benefitted from "crossing" acquisitions (with the remainder relying on "nominal" coverage only). Although this fraction is considerably higher for rough terrain (e.g. 74% where $ANVAD_{21\times21} \geq 7°$; the pink points in Fig. 8g), even here coverage is predominantly from the "nominal" acquisition geometry (70% on average, in these rough cells), meaning that hemispheric differences remain important.

The dependence plot presented in Fig. 8g is for **ascending** Sentinel-1 passes and so the explanation proposed above for that input variable is particularly relevant to the northern hemisphere (where most TanDEM-X acquisitions were also made on ascending passes). This points to a potential bias in data-driven DEM error prediction models generally: the higher availability of LiDAR-derived DTMs in the northern hemisphere (making up nearly 80% of reference grid cells in our case) may drive models to learn the patterns most relevant there, to the possible detriment of predictions made in the southern hemisphere (where they are needed most). Further research is needed to assess this in more detail and mitigate such biases.

4.4. Synthesis, limitations and recommendations for future research

Before making any recommendations for future research, it is important to acknowledge potential limitations relating to our reference data and methodology. While we tried to source reference LiDAR DTMs from diverse sites around the world, their scarcity in low-income countries (Pronk et al., 2024) means our dataset is likely biased towards the land cover and urban typologies found in higher-income countries. It may also be subject to site-specific biases (e.g. processing artefacts introduced when transforming elevations from a local vertical datum), which could affect model explanations in contexts where the true vertical error is relatively low (if models focus on these site-specific biases rather than the true vertical error). As for potential issues relating to gap-filling or editing of the Copernicus DEM, we note that this was very low across our study sites: only 0.7% of cells were filled (mostly using the 1 arc-second SRTM) and 1.1% had been edited.

When accounting for spatial autocorrelation in our modelling workflow, we adopted a spatial blocking approach (Roberts et al., 2017; Valavi et al., 2019) and based our block sizes on variogram ranges estimated for GLO-30 errors. These were too small to address the extensive spatial autocorrelation present in the climate variables though, such that local elevation biases learned from training blocks in a particular site (identifiable by climate variable value) were still relevant in test blocks from that same site. To ensure model explanations were not compromised by the "learning shortcuts" such correlations allow, we discarded all climate variables from our models. However, they may still be informative for DEM error prediction models, if evaluated using

an appropriate spatial cross-validation approach (much larger spatial blocks or else sampled at the site level).

Despite considering a very large set of candidate input variables, there are inevitably others not included here that may be even more informative. These could be entirely new variables or else alternative characterisations of a variable we did include, such as the GLAM-OUR (Li et al., 2024) and 3D-GloBFP (Che et al., 2024) building height datasets, released midway through our analysis. In some cases, we had to choose from a wide variety of alternative formulations, with a good example being surface roughness. We represented this using the Average Normal Vector Angular Deviation (ANVAD) and found it be an important predictor in all subset models. However, there is a wide variety of other roughness indices available, ranging from relatively simple options (e.g. standard deviation of slopes) to more advanced formulations such as the omnidirectional roughness and roughness anisotropy indices recently proposed by Trevisani and Guth (2024). Particularly given the clear significance of roughness to DEM error prediction, future studies might evaluate more of these alternatives.

It is also worth noting that the DEM error prediction models trained and explained here (as in almost all previous studies), take a somewhat narrow view of DEM "error" by simply considering vertical errors on a cell-by-cell basis. For many applications however, representing local morphology or enabling accurate derivatives (such as slope) may be more important (Polidori and El Hage, 2020; Trevisani et al., 2023). In their recent ranking of global DEMs, Guth et al. (2024) found that while model-corrected DEMs were generally more accurate (in the narrow sense), they often underperformed their source DEMs when it came to topographical derivatives (e.g. extracting drainage networks). To address this, Uhe et al. (2025) used a multi-criteria loss function when training the FathomDEM model, accounting for structural similarity and the spatial gradient of errors (as well as cell-by-cell vertical errors). The added nuance learned by such models would be an interesting target for future studies.

Despite those limitations, we contend that some robust recommendations for future DEM error correction studies can still be drawn from the analysis presented here. The first is that variable importance does vary significantly for models trained in different land cover environments, in terms of how many variables are relevant, which specific variables are selected, and their relative importance within each model. This suggests that it may be preferable to train an ensemble of models (each specialised in a particular land cover environment), rather than expecting a single global model to learn all of these diverse relationships. To date, such approaches have been relatively simple, using two separate models — either buildings and forests (Hawker et al., 2022) or buildings and everything else (Kim et al., 2021; Nguyen et al., 2022). Our results suggest that further specialisation may be useful, with more research needed to identify the optimal subdivision approach (e.g. by land cover, geographical region, climate zone, urban typology, etc.).

Across all subset models, we found consistently high importance metrics for topographical variables, especially neighbourhood statistics (expressions of a given grid cell's elevation relative to those of its neighbours). These were evaluated for a range of extents (from 3×3 to 21×21 grid cells) and we found that most models benefitted from a combination of small and large extents (3–7 and 21 grid cells, respectively). This dependence on local context suggests that model architectures using spatially-structured data inputs (e.g. image patches for Convolutional Neural Networks) are likely to outperform models that look at each grid cell in isolation. As for the neighbourhood size to be considered, we note that most of our models used variables for the 21×21 extent (the largest we evaluated), implying that even larger extents would likely be useful.

Multispectral data (both surface reflectance bands and derived indices) were also important in all subset models (especially "Short Vegetation" and "Open Forest"), consistent with expectations and suggested by previous studies (Nguyen et al., 2022; Dusseau et al., 2023; Chen et al., 2024). More surprising were the noteworthy contributions made by SAR variables (given their limited use in past studies), although they do present challenges (patchy spatial coverage and dependencies on terrain aspect, especially for Sentinel-1 VH polarisation). Future studies might assess the value of using the normalised Sentinel-1 Global Backscatter Model (Bauer-Marschallinger et al., 2021) instead, a mosaic product accounting for orbit geometry effects.

Recent analysis of the three global canopy height datasets (Potapov et al., 2021; Lang et al., 2023; Tolan et al., 2024) suggests they all suffer from particular biases, especially a tendency to underestimate tall canopies (Moudrý et al., 2024). Despite this, we found they were very useful predictors of GLO-30 error (in the top ten for all models except "Bare" ground), with the high-resolution (1 m grid spacing) option recently released by Tolan et al. (2024) consistently preferred. As for building heights, the World Settlement Footprint 3D (Esch et al., 2022) dataset was preferred (rank six) over the Global Human Settlement Layer (Pesaresi and Politis, 2023) option (rank 17) in our "Buildings" model, although its relatively coarse resolution (90 m) likely limits its impact.

5. Conclusions

Machine learning models are increasingly used for predicting vertical errors in global DEMs, enabling corrected versions for use in downstream applications requiring bare earth elevations (Hawker et al., 2022; Dusseau et al., 2023; Uhe et al., 2025). However, research to date has focused primarily on optimising predictive performance rather than interrogating or explaining these models. In this study, we trained machine learning models to predict vertical errors in the Copernicus DEM (GLO-30) in five different land cover environments ("Closed Forest", "Open Forest", "Short Vegetation", "Bare" and "Buildings") using diverse sites from across the globe and then explained these models using SHAP values. Our main finding is that variable importance varies significantly across these models (in terms of variables selected and their relative importance), suggesting that an ensemble of specialised models is likely to be more effective than a single global model. Further research is needed to better understand how those specialisations might be optimally defined.

In addition, the wide range of potentially-relevant input variables (160) evaluated here will support future DEM error prediction studies, in choosing between alternative datasets (e.g. the three global canopy height maps now available), appreciating the informational value of local topographical context (supporting the use of CNNs or similar model architectures), and exploring further the value of SAR backscatter intensities (rarely used so far, likely due to limited spatial coverage and imaging geometry challenges).

More generally, we present a framework for explaining models trained in parallel on different data subsets and demonstrate the potential value of using dependence plots to assess the patterns learned. In some cases, this may build trust in the models (where learned patterns are consistent with domain knowledge and seem likely to generalise well), while in others it can indicate problematic variables (proxies or spurious correlations) for closer assessment. Particularly for complex models whose outputs (and potential biases) feed into downstream applications impacting people's lives and livelihoods (e.g. flood or landslide simulations), such model explanations are increasingly important.

CRediT authorship contribution statement

Michael Meadows: Writing – original draft, Visualization, Methodology, Conceptualization, Software, Formal analysis. **Karin Reinke:** Supervision, Methodology, Writing – review & editing, Resources, Conceptualization. **Simon Jones:** Writing – review & editing, Resources, Conceptualization, Supervision, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We are grateful to all of the data providers and platforms that made this work possible, especially NASA, ESA, JAXA, NEON, Google Earth Engine, OpenTopography and the individual researchers cited in Table S1. Thank you also to Dr Nermine Hendy and Dr Roberto Del Prete for helpful discussions about the potential influence of SAR backscatter intensities in predicting DEM errors, and to two anonymous reviewers for providing detailed and insightful feedback that helped improve clarity and add nuance to many points. Michael Meadows was supported by a RTP Stipend Scholarship from the Australian Government and a Postgraduate Research Scholarship from Natural Hazards Research Australia (NHRA).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.aiig.2025.100141.

References

Aas, K., Jullum, M., Løland, A., 2021. Explaining individual predictions when features are dependent: more accurate approximations to Shapley values. Artificial Intelligence 298, 103502. http://dx.doi.org/10.1016/j.artint.2021.103502.

Abatzoglou, J.T., Dobrowski, S.Z., Parks, S.A., Hegewisch, K.C., 2018. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015. Sci. Data 5 (1), 170191. http://dx.doi.org/10.1038/sdata.2017.191.

Basu, I., Maji, S., 2022. Multicollinearity correction and combined feature effect in Shapley values. In: Long, G., Yu, X., Wang, S. (Eds.), AI 2021: Advances in Artificial Intelligence. Springer International Publishing, Cham, pp. 79–90. http://dx.doi.org/10.1007/978-3-030-97546-3_7.

Bauer-Marschallinger, B., Cao, S., Navacchi, C., Freeman, V., Reuß, F., Geudtner, D., Rommen, B., Vega, F.C., Snoeij, P., Attema, E., Reimer, C., Wagner, W., 2021. The normalised sentinel-1 global backscatter model, mapping earth's land surface with c-band microwaves. Sci. Data 8 (1), 277. http://dx.doi.org/10.1038/s41597-021-01059-7.

Baugh, C.A., Bates, P.D., Schumann, G., Trigg, M.A., 2013. SRTM vegetation removal and hydrodynamic modeling accuracy. Water Resour. Res. 49 (9), 5276–5289. http://dx.doi.org/10.1002/wrcr.20412.

Beck, H.E., Zimmermann, N.E., McVicar, T.R., Vergopolan, N., Berg, A., Wood, E.F., 2018. Present and future köppen-geiger climate classification maps at 1-km resolution. Sci. Data 5 (1), 180214. http://dx.doi.org/10.1038/sdata.2018.214.

Bentéjac, C., Csörgő, A., Martínez-Muñoz, G., 2021. A comparative analysis of gradient boosting algorithms. Artif. Intell. Rev. 54 (3), 1937–1967. http://dx.doi.org/10.1007/s10462-020-09896-5.

Bergstra, J., Yamins, D., Cox, D.D., 2013. Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. In: Proceedings of the 30th International Conference on Machine Learning (ICML 2013). pp. I–115 to I–23.

Bielski, C., López-Vázquez, C., Grohmann, C.H., Guth, P.L., Hawker, L., Gesch, D., Trevisani, S., Herrera-Cruz, V., Riazanoff, S., Corseaux, A., Reuter, H.I., Strobl, P., 2024. Novel approach for ranking DEMs: copernicus DEM improves one arc second open global topography. IEEE Trans. Geosci. Remote Sens. 62, 1–22. http://dx.doi.org/10.1109/TGRS.2024.3368015.

Brenning, A., 2012. Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. In: 2012 IEEE International Geoscience and Remote Sensing Symposium. pp. 5372–5375. http://dx.doi.org/10.1109/IGARSS.2012.6352393.

Brock, J., Schratz, P., Petschko, H., Muenchow, J., Micu, M., Brenning, A., 2020. The performance of landslide susceptibility models critically depends on the quality of digital elevation models. Geomatics, Nat. Hazards Risk 11 (1), 1075–1092. http://dx.doi.org/10.1080/19475705.2020.1776403.

Brown, C.G., Sarabandi, K., Pierce, L.E., 2010. Model-based estimation of forest canopy height in red and Austrian pine stands using shuttle radar topography mission and ancillary data: a proof-of-concept study. IEEE Trans. Geosci. Remote Sens. 48 (3), 1105–1118. http://dx.doi.org/10.1109/TGRS.2009.2031635.

Buchhorn, M., Smets, B., Bertels, L., Roo, B.D., Lesiv, M., Tsendbazar, N.-E., Li, L., Tarko, A., 2020. Copernicus global land service: Land cover 100m: version 3 globe 2015–2019: product user manual. Zenodo, Geneva, Switzerland, http://dx.doi.org/10.5281/ZENODO.3938963.

Che, Y., Li, X., Liu, X., Wang, Y., Liao, W., Zheng, X., Zhang, X., Xu, X., Shi, Q., Zhu, J., Zhang, H., Yuan, H., Dai, Y., 2024. 3D-GloBFP: The first global three-dimensional building footprint dataset. Earth Syst. Sci. Data 16 (11), 5357–5374. http://dx.doi.org/10.5194/essd-16-5357-2024.

Chen, H., Covert, I.C., Lundberg, S.M., Lee, S.-I., 2023. Algorithms to estimate Shapley value feature attributions. Nat. Mach. Intell. 5 (6), 590–601. http://dx.doi.org/10.1038/s42256-023-00657-x.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. In: KDD '16, Association for Computing Machinery, New York, NY, USA, pp. 785–794. http://dx.doi.org/10.1145/2939672.2939785.

Chen, C., Liu, Y., Li, Y., Chen, D., 2024. Explainable artificial intelligence framework for urban global digital elevation model correction based on the shapley additive explanation-random forest algorithm considering spatial heterogeneity and factor optimization. Int. J. Appl. Earth Obs. Geoinf. 129, 103843. http://dx.doi.org/10.1016/j.jag.2024.103843.

Chen, C., Yang, S., Li, Y., 2020. Accuracy assessment and correction of SRTM DEM using icesat/GLAS Data under data coregistration. Remote. Sens. 12 (20), 3435. http://dx.doi.org/10.3390/rs12203435.

Climent, F., Momparler, A., Carmona, P., 2019. Anticipating bank distress in the eurozone: an extreme gradient boosting approach. J. Bus. Res. 101, 885–896. http://dx.doi.org/10.1016/j.jbusres.2018.11.015.

Crippen, R., Buckley, S., Agram, P., Belz, E., Gurrola, E., Hensley, S., Kobrick, M., Lavalle, M., Martin, J., Neumann, M., Nguyen, Q., Rosen, P., Shimada, J., Simard, M., Tung, W., 2016. NASADEM global elevation model: Methods and progress. ISPRS - Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. XLI-B4, 125–128. http://dx.doi.org/10.5194/isprsarchives-XLI-B4-125-2016.

Dubayah, R., Blair, J.B., Goetz, S., Fatoyinbo, L., Hansen, M., Healey, S., Hofton, M., Hurtt, G., Kellner, J., Luthcke, S., Armston, J., Tang, H., Duncanson, L., Hancock, S., Jantz, P., Marselis, S., Patterson, P.L., Qi, W., Silva, C., 2020. The global ecosystem dynamics investigation: High-resolution laser ranging of the earth's forests and topography. Sci. Remote. Sens. 1, 100002. http://dx.doi.org/10.1016/j.srs.2020.100002.

Dusseau, D., Zobel, Z., Schwalm, C.R., 2023. DiluviumDEM: Enhanced accuracy in Global Coastal digital elevation models. Remote Sens. Environ. 298, 113812. http://dx.doi.org/10.1016/j.rse.2023.113812.

Elvidge, C.D., Zhizhin, M., Ghosh, T., Hsu, F.-C., Taneja, J., 2021. Annual time series of global VIIRS nighttime lights derived from monthly averages: 2012 to 2019. Remote. Sens. 13 (5), 922. http://dx.doi.org/10.3390/rs13050922.

Esch, T., Brzoska, E., Dech, S., Leutner, B., Palacios-Lopez, D., Metz-Marconcini, A., Marconcini, M., Roth, A., Zeidler, J., 2022. World settlement footprint 3D - a first three-dimensional survey of the global building stock. Remote Sens. Environ. 270, 112877. http://dx.doi.org/10.1016/j.rse.2021.112877.

Fahrland, E., Paschko, H., Jacob, P., Kahabka, H., 2022. Technical Report AO/1-9422/18/I-LG, Airbus Defence and Space GmbH, Potsdam, Germany.

Farr, T.G., Rosen, P.A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., Alsdorf, D., 2007. The Shuttle radar topography mission. Rev. Geophys. 45 (2), http://dx.doi.org/10.1029/2005RG000183.

Gao, S., Zhong, R., Yan, K., Ma, X., Chen, X., Pu, J., Gao, S., Qi, J., Yin, G., Myneni, R.B., 2023. Evaluating the saturation effect of vegetation indices in forests using 3D radiative transfer simulations and satellite observations. Remote Sens. Environ. 295, 113665. http://dx.doi.org/10.1016/j.rse.2023.113665.

Guth, P.L., Trevisani, S., Grohmann, C.H., Lindsay, J., Gesch, D., Hawker, L., Bielski, C., 2024. Ranking of 10 global one-arc-second DEMs reveals limitations in terrain morphology representation. Remote. Sens. 16 (17), 3273. http://dx.doi.org/10.3390/rs16173273.

Hancock, S., McGrath, C., Lowe, C., Davenport, I., Woodhouse, I., 2021. Requirements for a global lidar system: Spaceborne lidar with wall-to-wall coverage. R. Soc. Open Sci. 8 (12), 211166. http://dx.doi.org/10.1098/rsos.211166.

Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O., Townshend, J.R.G., 2013. High-resolution global maps of 21st-century forest cover change. Science 342 (6160), 850–853. http://dx.doi.org/10.1126/science.1244693.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: Data Mining, inference, and prediction, second ed. Springer Series in Statistics, Springer, New York, NY.

Hawker, L., Neal, J., Bates, P., 2019. Accuracy assessment of the TanDEM-X 90 digital elevation model for selected floodplain sites. Remote Sens. Environ. 232 (111319), 0–15. http://dx.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 Resour. Res. 54 (10), 7910–7928. http://dx.doi.org/10.1029/2018WR023279.

Hawker, L., Uhe, P., Paulo, L., Sosa, J., Savage, J., Sampson, C., Neal, J., 2022. A 30 m global map of elevation with forests and buildings removed. Environ. Res. Lett. 17 (2), 024016. http://dx.doi.org/10.1088/1748-9326/ac4d4f.

Höhle, J., Höhle, M., 2009. Accuracy assessment of digital elevation models by means of robust statistical methods. ISPRS J. Photogramm. Remote Sens. 64 (4), 398–406. http://dx.doi.org/10.1016/j.isprsjprs.2009.02.003.

Janzing, D., Minorics, L., Bloebaum, P., 2020. Feature relevance quantification in explainable AI: A causal problem. In: Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics. PMLR, pp. 2907–2916.

Karasiak, N., Dejoux, J.-F., Monteil, C., Sheeren, D., 2022. Spatial dependence between training and test sets: Another pitfall of classification accuracy assessment in remote sensing. Mach. Learn. 111 (7), 2715–2740. http://dx.doi.org/10.1007/s10994-021-05972-1.

Kellndorfer, J., 2019. Using SAR data for mapping deforestation and forest degradation. In: Flores, A., Herndon, K., Thapa, R., Cherrington, E. (Eds.), SAR Handbook: Comprehensive Methodologies for Forest Monitoring and Biomass Estimation. NASA, pp. 65–79. http://dx.doi.org/10.25966/68c9-gw82.

Kim, D.E., Liu, J., Liong, S.-Y., Gourbesville, P., Strunz, G., 2021. Satellite DEM improvement using multispectral imagery and an artificial neural network. Water 13 (11), 1551. http://dx.doi.org/10.3390/w13111551.

Kulp, S.A., Strauss, B.H., 2018. Coastaldem: A Global Coastal digital elevation model improved from SRTM using a neural network. Remote Sens. Environ. 206, 231–239. http://dx.doi.org/10.1016/j.rse.2017.12.026.

Lang, N., Jetz, W., Schindler, K., Wegner, J.D., 2023. A high-resolution canopy height model of the earth. Nat. Ecol. Evol. 7 (11), 1778–1789. http://dx.doi.org/10.1038/s41559-023-02206-6.

Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J., Bretagnolle, V., 2014. Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. Glob. Ecol. Biogeogr. 23 (7), 811–820. http://dx.doi.org/10.1111/geb.12161.

Lefsky, M.A., 2010. A global forest canopy height map from the moderate resolution imaging spectroradiometer and the geoscience laser altimeter system. Geophys. Res. Lett. 37 (15), http://dx.doi.org/10.1029/2010GL043622.

Legendre, P., Fortin, M.J., 1989. Spatial pattern and ecological analysis. Vegetatio 80 (2), 107–138. http://dx.doi.org/10.1007/BF00048036.

Li, Y., Li, L., Chen, C., Liu, Y., 2023a. Correction of global digital elevation models in forested areas using an artificial neural network-based method with the consideration of spatial autocorrelation. Int. J. Digit. Earth 16 (1), 1568–1588. http://dx.doi.org/10.1080/17538947.2023.2203953.

Li, X., Ma, R., Zhang, Q., Li, D., Liu, S., He, T., Zhao, L., 2019. Anisotropic characteristic of artificial light at night – Systematic investigation with VIIRS DNB multi-temporal observations. Remote Sens. Environ. 233, 111357. http://dx.doi.org/10.1016/j.rse.2019.111357.

Li, R., Sun, T., Ghaffarian, S., Tsamados, M., Ni, G., 2024. GLAMOUR: GLobAl building morphology dataset for URban hydroclimate modelling. Sci. Data 11 (1), 618. http://dx.doi.org/10.1038/s41597-024-03446-2.

Li, B., Xie, H., Tong, X., Tang, H., Liu, S., 2023b. A global-scale DEM elevation correction model using icesat-2 laser altimetry data. IEEE Trans. Geosci. Remote Sens. 61, 1–15. http://dx.doi.org/10.1109/TGRS.2023.3321956.

Li, H., Zhao, J., Yan, B., Yue, L., Wang, L., 2022. Global DEMs vary from one to another: An evaluation of newly released copernicus, NASA and AW3D30 DEM on selected terrains of China using icesat-2 altimetry data. Int. J. Digit. Earth 15 (1), 1149–1168. http://dx.doi.org/10.1080/17538947.2022.2094002.

Lindsay, J., 2016. Whitebox GAT: A case study in geomorphometric analysis. Comput. Geosci. 95, 75–84. http://dx.doi.org/10.1016/j.cageo.2016.07.003.

Lindsay, J.B., Newman, D.R., Francioni, A., 2019. Scale-optimized surface roughness for topographic analysis. Geosciences 9 (7), 322. http://dx.doi.org/10.3390/geosciences9070322.

Liu, Y., Bates, P.D., Neal, J.C., Yamazaki, D., 2021. Bare-earth DEM generation in urban areas for flood inundation simulation using global digital elevation models. Water Resour. Res. 57 (e2020WR028516), 1–25. http://dx.doi.org/10.1029/2020WR028516.

Liu, K., Song, C., Ke, L., Jiang, L., Pan, Y., Ma, R., 2019. Global open-access DEM performances in earth's Most Rugged Region high mountain Asia: A multi-level assessment. Geomorphology 338, 16–26. http://dx.doi.org/10.1016/j.geomorph.2019.04.012.

Liu, Z., Zhu, J., Fu, H., Zhou, C., Zuo, T., 2020. Evaluation of the vertical accuracy of open global DEMs over Steep Terrain Regions using icesat data: A case study over hunan province, China. Sensors 20 (17), 4865. http://dx.doi.org/10.3390/s20174865.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I., 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2 (1), 56–67. http://dx.doi.org/10.1038/s42256-019-0138-9.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems. 30, Curran Associates, Inc., pp. 1–10.

Mälicke, M., 2022. SciKit-GStat 1.0: A scipy-flavored geostatistical variogram estimation toolbox written in Python. Geosci. Model. Dev. 15 (6), 2505–2532. http://dx.doi.org/10.5194/gmd-15-2505-2022.

Marconcini, M., Metz-Marconcini, A., Üreyen, S., Palacios-Lopez, D., Hanke, W., Bachofer, F., Zeidler, J., Esch, T., Gorelick, N., Kakarla, A., Paganini, M., Strano, E., 2020. Outlining where humans live, the world settlement footprint 2015. Sci. Data 7 (1), 242. http://dx.doi.org/10.1038/s41597-020-00580-5.

Marešová, J., Gdulová, K., Pracná, P., Moravec, D., Gábor, L., Prošek, J., Barták, V., Moudrý, V., 2021. Applicability of data acquisition characteristics to the identification of local artefacts in global digital elevation models: comparison of the copernicus and TanDEM-X DEMs. Remote. Sens. 13 (19), 3931. http://dx.doi.org/10.3390/rs13193931.

Martone, M., Bräutigam, B., Rizzoli, P., Gonzalez, C., Bachmann, M., Krieger, G., 2012. Coherence evaluation of tandem-x interferometric data. Innovative Applications of SAR Interferometry from Modern Satellite Sensors, ISPRS J. Photogramm. Remote Sens. Innovative Applications of SAR Interferometry from Modern Satellite Sensors, 73, 21–29.http://dx.doi.org/10.1016/j.isprsjprs.2012.06.006,

Meadows, M., Jones, S., Reinke, K., 2024. Vertical accuracy assessment of freely available global DEMs (FABDEM, copernicus DEM, NASADEM, AW3D30 and SRTM) in flood-prone environments. Int. J. Digit. Earth 17 (1), 2308734. http://dx.doi.org/10.1080/17538947.2024.2308734.

Meadows, M., Wilson, M., 2021. A comparison of machine learning approaches to improve free topography data for flood modelling. Remote. Sens. 13 (2), 275. http://dx.doi.org/10.3390/rs13020275.

Meyer, F., 2019. Spaceborne synthetic aperture radar – Principles, data access, and basic processing techniques. In: Flores, A., Herndon, K., Thapa, R., Cherrington, E. (Eds.), SAR Handbook: Comprehensive Methodologies for Forest Monitoring and Biomass Estimation. NASA, pp. 21–43. http://dx.doi.org/10.25966/ez4f-mg98.

Moudrý, V., Gábor, L., Marselis, S., Pracná, P., Barták, V., Prošek, J., Navrátilová, B., Novotný, J., Potůčková, M., Gdulová, K., Crespo-Peremarch, P., Komárek, J., Malavasi, M., Rocchini, D., Ruiz, L.A., Torralba, J., Torresani, M., Cazzolla Gatti, R., Wild, J., 2024. Comparison of three global canopy height maps and their applicability to biodiversity modeling: accuracy issues revealed. Ecosphere 15 (10), e70026. http://dx.doi.org/10.1002/ecs2.70026.

Moudrý, V., Lecours, V., Gdulová, K., Gábor, L., Moudrá, L., Kropáček, J., Wild, J., 2018. On the use of global DEMs in ecological modelling and the accuracy of new bare-earth DEMs. Ecol. Model. 383, 3–9. http://dx.doi.org/10.1016/j.ecolmodel.2018.05.006.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D.G., Piles, M., Rodríguez-Fernández, N.J., Zsoter, E., Buontempo, C., Thépaut, J.-N., 2021. ERA5-land: A state-of-the-art global reanalysis dataset for land applications. Earth Syst. Sci. Data 13 (9), 4349–4383. http://dx.doi.org/10.5194/essd-13-4349-2021.

Natekin, A., Knoll, A., 2013. Gradient boosting machines, a tutorial. Front. Neurorobotics 7, http://dx.doi.org/10.3389/fnbot.2013.00021.

Nguyen, N.S., Kim, D.E., Jia, Y., Raghavan, S.V., Liong, S.Y., 2022. Application of multi-channel convolutional neural network to improve DEM data in urban cities. Technologies 10 (3), 61. http://dx.doi.org/10.3390/technologies10030061.

Nuth, C., Kääb, A., 2011. Co-registration and bias corrections of satellite elevation data sets for quantifying glacier thickness change. Cryosphere 5 (1), 271–290. http://dx.doi.org/10.5194/tc-5-271-2011.

Okeson, A., Caruana, R., Craswell, N., Inkpen, K., Lundberg, S.M., Nori, H., Vaughan, J.W., 2021. Summarize with caution: Comparing global feature attributions. Bull. the IEEE Comput. Soc. Tech. Comm. Data Eng. 44 (4), 14–27.

Okolie, C., Adeleke, A., Mills, J., Smit, J., Maduako, I., Bagheri, H., Komar, T., Wang, S., 2024a. Assessment of explainable tree-based ensemble algorithms for the enhancement of copernicus digital elevation model in agricultural lands. Int. J. Image Data Fusion 1–31. http://dx.doi.org/10.1080/19479832.2024.2329563.

Okolie, C., Mills, J., Adeleke, A., Smit, J., 2024b. Digital elevation model correction in urban areas using extreme gradient boosting, land cover and terrain parameters. Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. XLVIII-4-W9-2024, 275–282. http://dx.doi.org/10.5194/isprs-archives-XLVIII-4-W9-2024-275-2024.

Okolie, C.J., Mills, J.P., Adeleke, A.K., Smit, J.L., Peppa, M.V., Altunel, A.O., Arungwa, I.D., 2024c. Assessment of the global copernicus, NASADEM, ASTER and AW3D digital elevation models in central and Southern Africa. Geo- Spat. Inf. Sci. 1–29. http://dx.doi.org/10.1080/10095020.2023.2296010.

Olajubu, V., Trigg, M.A., Berretta, C., Sleigh, A., Chini, M., Matgen, P., Mojere, S., Mulligan, J., 2021. Urban correction of global DEMs using building density for nairobi, Kenya. Earth Sci. Informatics 14 (3), 1383–1398. http://dx.doi.org/10.1007/s12145-021-00647-w.

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 Sens. Environ. 182, 49–59. http://dx.doi.org/10.1016/j.rse.2016.04.018.

Pesaresi, M., 2023. GHS-BUILT-s R2023a - GHS built-up surface grid, derived from Sentinel2 composite and landsat, multitemporal (1975–2030). http://dx.doi.org/10.2905/9F06F36F-4B11-47EC-ABB0-4F8B7B1D72EA.

Pesaresi, M., Politis, P., 2023. GHS-BUILT-h R2023a - GHS building height, derived from AW3D30, SRTM30, and Sentinel2 composite (2018). http://dx.doi.org/10.2905/85005901-3A49-48DD-9D19-6261354F56FE.

Pimenova, O., Roberts, C., Rizos, C., 2022. Regional "Bare-Earth" digital terrain model for costa rica based on NASADEM corrected for vegetation bias. Remote. Sens. 14 (10), 2421. http://dx.doi.org/10.3390/rs14102421.

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., Pélissier, R., 2020. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nat. Commun. 11 (1), 4540. http://dx.doi.org/10.1038/s41467-020-18321-y.

Polidori, L., El Hage, M., 2020. Digital elevation model quality assessment methods: a critical review. Remote. Sens. 12 (21), 3522. http://dx.doi.org/10.3390/rs12213522.

Potapov, P., Li, X., Hernandez-Serna, A., Tyukavina, A., Hansen, M.C., Kommareddy, A., Pickens, A., Turubanova, S., Tang, H., Silva, C.E., Armston, J., Dubayah, R., Blair, J.B., Hofton, M., 2021. Mapping global forest canopy height through integration of GEDI and landsat data. Remote Sens. Environ. 253, 112165. http://dx.doi.org/10.1016/j.rse.2020.112165.

Pronk, M., Hooijer, A., Eilander, D., Haag, A., de Jong, T., Vousdoukas, M., Vernimmen, R., Ledoux, H., Eleveld, M., 2024. DeltaDTM: A Global Coastal digital terrain model. Sci. Data 11 (1), 273. http://dx.doi.org/10.1038/s41597-024-03091-9.

Qiu, W., Chen, H., Dincer, A.B., Lundberg, S., Kaeberlein, M., Lee, S.-I., 2022. Interpretable machine learning prediction of all-cause mortality. Commun. Med. 2 (1), 1–15. http://dx.doi.org/10.1038/s43856-022-00180-x.

Rizzoli, P., Martone, M., Gonzalez, C., Wecklich, C., Borla Tridon, D., Bräutigam, B., Bachmann, M., Schulze, D., Fritz, T., Huber, M., Wessel, B., Krieger, G., Zink, M., Moreira, A., 2017. Generation and performance assessment of the global tandem-x digital elevation model. ISPRS J. Photogramm. Remote Sens. 132, 119–139. http://dx.doi.org/10.1016/j.isprsjprs.2017.08.008.

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40 (8), 913–929. http://dx.doi.org/10.1111/ecog.02881.

Román, M.O., Wang, Z., Sun, Q., Kalb, V., Miller, S.D., Molthan, A., Schultz, L., Bell, J., Stokes, E.C., Pandey, B., Seto, K.C., Hall, D., Oda, T., Wolfe, R.E., Lin, G., Golpayegani, N., Devadiga, S., Davidson, C., Sarkar, S., Praderas, C., Schmaltz, J., Boller, R., Stevens, J., Ramos González, O.M., Padilla, E., Alonso, J., Detrés, Y., Armstrong, R., Miranda, I., Conte, Y., Marrero, N., MacManus, K., Esch, T., Masuoka, E.J., 2018. NASA's black marble nighttime lights product suite. Remote Sens. Environ. 210, 113–143. http://dx.doi.org/10.1016/j.rse.2018.03.017.

Saatchi, S., 2019. SAR methods for mapping and monitoring forest biomass. In: Flores, A., Herndon, K., Thapa, R., Cherrington, E. (Eds.), SAR Handbook: Comprehensive Methodologies for Forest Monitoring and Biomass Estimation. NASA, pp. 207–246. http://dx.doi.org/10.25966/hbm1-ej07.

Sampson, C.C., Smith, A.M., Bates, P.D., Neal, J.C., Alfieri, L., Freer, J.E., 2015. A high-resolution global flood hazard model. Water Resour. Res. 51 (9), 7358–7381. http://dx.doi.org/10.1002/2015WR016954.

Schiavina, M., Freire, S., MacManus, K., 2023a. GHS-POP R2023a - GHS population grid multitemporal (1975–2030). http://dx.doi.org/10.2905/2FF68A52-5B5B-4A22-8F40-C41DA8332CFE.

Schiavina, M., Melchiorri, M., Pesaresi, M., 2023b. GHS-SMOD R2023a - GHS settlement layers, application of the degree of urbanisation methodology (stage i) to GHS-POP R2023a and GHS-BUILT-s R2023a, multitemporal (1975–2030). http://dx.doi.org/10.2905/A0DF7A6F-49DE-46EA-9BDE-563437A6E2BA.

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., Brenning, A., 2019. Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. Ecol. Model. 406, 109–120. http://dx.doi.org/10.1016/j.ecolmodel.2019.06.002.

Shapley, L.S., 1953. A value for N-person games. In: Kuhn, H.W., Tucker, A.W. (Eds.), Contributions To the Theory of Games II. Princeton University Press, Princeton, pp. 307–317. http://dx.doi.org/10.1515/9781400881970-018.

Shen, X., Zhou, C., Zhu, J., 2023. Improving the accuracy of TanDEM-X digital elevation model using least squares collocation method. Remote. Sens. 15 (14), 3695. http://dx.doi.org/10.3390/rs15143695.

Shimada, M., Itoh, T., Motooka, T., Watanabe, M., Shiraishi, T., Thapa, R., Lucas, R., 2014. New global forest/non-forest maps from ALOS PALSAR data (2007–2010). Remote Sens. Environ. 155, 13–31. http://dx.doi.org/10.1016/j.rse.2014.04.014.

Shortridge, A., Messina, J., 2011. Spatial structure and landscape associations of SRTM error. Remote Sens. Environ. 115 (6), 1576–1587. http://dx.doi.org/10.1016/j.rse.2011.02.017.

Simard, M., Pinto, N., Fisher, J.B., Baccini, A., 2011. Mapping forest canopy height globally with spaceborne lidar. J. Geophys. Res.: Biogeosciences 116 (G4), http://dx.doi.org/10.1029/2011JG001708.

Tadono, T., Nagai, H., Ishida, H., Oda, F., Naito, S., Minakawa, K., Iwamoto, H., 2016. Generation of the 30 M-Mesh global digital surface model by ALOS PRISM. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. XLI-B4, Copernicus GmbH, pp. 157–162. http://dx.doi.org/10.5194/isprs-archives-XLI-B4-157-2016.

Telford, R.J., Birks, H.J.B., 2009. Evaluation of transfer functions in spatially structured environments. Quat. Sci. Rev. 28 (13), 1309–1316. http://dx.doi.org/10.1016/j.quascirev.2008.12.020.

Tobler, W.R., 1970. A computer movie simulating urban growth in the detroit region. Econ. Geogr. 46, 234–240. http://dx.doi.org/10.2307/143141, arXiv:143141.

Tolan, J., Yang, H.-I., Nosarzewski, B., Couairon, G., Vo, H.V., Brandt, J., Spore, J., Majumdar, S., Haziza, D., Vamaraju, J., Moutakanni, T., Bojanowski, P., Johns, T., White, B., Tiecke, T., Couprie, C., 2024. Very high resolution canopy height maps from RGB imagery using self-supervised vision transformer and convolutional decoder trained on aerial lidar. Remote Sens. Environ. 300, 113888. http://dx.doi.org/10.1016/j.rse.2023.113888.

Toutin, T., 2002. Impact of terrain slope and aspect on radargrammetric DEM accuracy. Image Spectroscopy and Hyperspectral Imaging (Special Section), ISPRS J. Photogramm. Remote Sens. Image Spectroscopy and Hyperspectral Imaging (Special Section), 57 (3), 228–240.http://dx.doi.org/10.1016/S0924-2716(02)00123-5,

Trevisani, S., Guth, P.L., 2024. Terrain analysis according to multiscale surface roughness in the taklimakan desert. Land 13 (11), 1843. http://dx.doi.org/10.3390/land13111843.

Trevisani, S., Skrypitsyna, T.N., Florinsky, I.V., 2023. Global digital elevation models for terrain morphology analysis in mountain environments: Insights on copernicus GLO-30 and ALOS AW3D30 for a large alpine area. Environ. Earth Sci. 82 (9), 198. http://dx.doi.org/10.1007/s12665-023-10882-7.

Tsendbazar, N., Herold, M., Li, L., Tarko, A., de Bruin, S., Masiliunas, D., Lesiv, M., Fritz, S., Buchhorn, M., Smets, B., Van De Kerchove, R., Duerauer, M., 2021. Towards operational validation of annual global land cover maps. Remote Sens. Environ. 266, 112686. http://dx.doi.org/10.1016/j.rse.2021.112686.

Uhe, P., Lucas, C., Hawker, L., Brine, M., Wilkinson, H., Cooper, A., Saoulis, A.A., Savage, J., Sampson, C., 2025. Fathomdem: An improved global terrain map using a hybrid vision transformer model. Environ. Res. Lett. 20 (3), 034002. http://dx.doi.org/10.1088/1748-9326/ada972.

Uuemaa, E., Ahi, S., Montibeller, B., Muru, M., Kmoch, A., 2020. Vertical accuracy of freely available global digital elevation models (ASTER, AW3D30, MERIT, TanDEM-X, SRTM, and NASADEM). Remote. Sens. 12 (21), 3482. http://dx.doi.org/10.3390/rs12213482.

Valavi, R., Elith, J., Lahoz-Monfort, J.J., Guillera-Arroita, G., 2019. Blockcv: an r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. Methods Ecol. Evol. 10 (2), 225–232. http://dx.doi.org/10.1111/2041-210X.13107.

Verhaeghe, J., Van Der Donckt, J., Ongenae, F., Van Hoecke, S., 2023. Powershap: a power-full Shapley feature selection method. In: Amini, M.-R., Canu, S., Fischer, A., Guns, T., Kralj Novak, P., Tsoumakas, G. (Eds.), Machine Learning and Knowledge Discovery in Databases. In: Lecture Notes in Computer Science, Springer International Publishing, Cham, pp. 71–87. http://dx.doi.org/10.1007/978-3-031-26387-3_5.

Wendi, D., Liong, S.-Y., Sun, Y., Doan, C.D., 2016. An innovative approach to improve SRTM DEM using multispectral imagery and artificial neural network. J. Adv. Model. Earth Syst. 8 (2), 691–702. http://dx.doi.org/10.1002/2015MS000536.

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 J. Photogramm. Remote Sens. 139, 171–182. http://dx.doi.org/10.1016/j.isprsjprs.2018.02.017.

Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J.C., Sampson, C.C., Kanae, S., Bates, P.D., 2017. A high-accuracy map of global terrain elevations. Geophys. Res. Lett. 44 (11), 5844–5853. http://dx.doi.org/10.1002/2017GL072874.