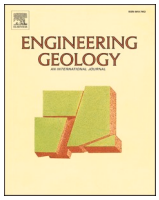
Engineering Geology 362 (2026) 108541

Contents lists available at ScienceDirect

Engineering Geology

journal homepage: www.elsevier.com/locate/enggeo



Seismic site characterization using satellite-derived terrain morphometry and geological data: A machine learning approach for predominant frequency prediction

Harish Thakur, P. Anbazhagan *

Department of Civil Engineering, Indian Institute of Science, Bengaluru 560012, India

ARTICLE INFO

Keywords:
Predominant frequency
GLO-30 DEM
Terrain morphometry
Seismic site characterization
Stacked Modelling, Residual Kriging

ABSTRACT

Predominant frequency (*fo*) characterization across large seismically active regions remains challenging due to limited field measurements and cost constraints. Existing *fo* mapping approaches rely exclusively on spatial interpolation methods (kriging, inverse distance weighting, natural neighbor) that redistribute measured values without incorporating terrain morphometry, geological context, or subsurface parameters as predictors. This study develops a DEM-based machine learning methodology for regional-scale *fo* prediction in the Himalayan region and Indo-Gangetic Plains, addressing critical data scarcity in earthquake-prone developing countries. We compiled 4400 *fo* measurements from 26 published HVSR studies using systematic georeferencing procedures to ensure spatial consistency. The methodology employs a two-stage regression kriging framework: (1) stacked ensemble machine learning models trained on 20 predictor variables using GLO-30 DEM morphometric parameters (elevation, slope, curvature indices), geological classifications, and bedrock depth information to capture nonlinear terrain-frequency relationships; and (2) ordinary kriging of model residuals to account for spatial correlation patterns. Cross-validation partitioning ensures unbiased residuals, while Bayesian optimization determines optimal hyperparameters for base model selection. Feature importance analysis reveals that valley bottom identification (MRVBF), geological formation characteristics, and bedrock depth provide primary predictive capability (Shapley values ~0.15–0.18), demonstrating that terrain morphometry and subsurface parameters effectively control *fo* variation at regional scales. The stacked ensemble achieves $R^2 = 0.516$ and RMSE = 0.634 log units, with variogram analysis revealing spatial correlation extending 7.3 km and structured variance accounting for 52 % of model residuals. High-resolution *fo* maps (50 m grid) generated for Delhi, Kathmandu, and Dhaka differentiate site response zones: low frequencies (<1.0 Hz) in deep sedimentary basins versus high frequencies (>3.0 Hz) in bedrock-controlled areas.

This work represents the first regional-scale application of DEM-derived terrain morphometry for direct *fo* prediction, utilizing a much larger compiled dataset for this purpose than previous basin-scale studies. Unlike previous studies that employed purely interpolation techniques without predictive parameters, this hybrid framework integrates physical predictors (terrain morphometry, geology, bedrock depth) with spatial modelling to produce more robust *fo* maps. Results demonstrate that incorporating satellite-derived morphometric and geological parameters—readily available globally—significantly enhances prediction reliability beyond interpolation-only approaches. This cost-effective methodology enables preliminary seismic hazard assessment in data-sparse mountainous regions where traditional field surveys are resource-constrained, with applications for earthquake risk assessment, regional zonation, and building code implementation in developing countries.

1. Introduction

Predominant frequency (*fo*) characterization across large seismically active regions remains fundamentally important for earthquake hazard assessment, yet achieving comprehensive spatial coverage is constrained by limited field measurements and cost considerations in data-sparse regions. In literature, site classification is usually done based on parameters quantifying near-surface stiffness – such as shear wave

* Corresponding author.
E-mail addresses: harishthakur@iisc.ac.in (H. Thakur), anbazhagan@iisc.ac.in (P. Anbazhagan).

https://doi.org/10.1016/j.enggeo.2025.108541
Received 13 September 2025; Received in revised form 29 November 2025; Accepted 31 December 2025
Available online 2 January 2026

velocity, Vs, and the fundamental dynamic response – such as the site's predominant or natural frequency, *fo* (Akin et al., 2011). For a one-dimensional response analysis, these two parameters are related by a fundamental expression $fo = Vs/(4H)$, where Vs represents the average shear wave velocity of overlying sediments and H denotes the depth to seismic bedrock (Kramer and Stewart, 2024). Furthermore, these site characteristics influence the built environment's response during the seismic events (Panzera et al., 2018; Brando et al., 2020; Senkaya et al., 2024). In the microzonation studies, these parameters have been used either independently or in combination for site classification (Yilar et al., 2017; Maklad et al., 2020; Abbasnejadfard et al., 2023; Martínez-Segura et al., 2024; Di Martino et al., 2024). It has been well established in the earthquake engineering literature that beyond predictor such as Vs30 (shear wave velocity of top 30 m of a site) for characterizing a site's seismic response, the *fo* value of the site also plays a critical role as it directly influences the amplification characteristics of the site (Delgado et al., 2000; Haghshenas et al., 2008; Hassani and Atkinson, 2018; Senkaya et al., 2024).

For regional-scale microzonation purposes, significant literature exists for shear wave velocity, Vs30-based site characterization using remotely sensed data such as Digital Elevation Models (DEMs) and their derivatives (Allen and Wald, 2009; Mori et al., 2020; Kim et al., 2021; Geyin and Maurer, 2023; Thakur and Anbazhagan, 2025). Earlier studies for site characterization using Vs30 relied on relatively simple DEM parameters such as topographic slope (Allen and Wald, 2009). In recent studies, more advanced models utilizing different geomorphometric, geomorphological predictors and spatial analysis have been developed for this purpose (Liu et al., 2017; Geyin and Maurer, 2023; Abbasnejadfard et al., 2023; Thakur and Anbazhagan, 2025). The *fo* estimation represents a fundamental parameter in seismic site characterization. Past studies have demonstrated that sites typically overlain with softer sedimentary materials over stronger bedrock are particularly susceptible to amplification and liquefaction effects during earthquake events (Lin et al., 2021; Kramer and Stewart, 2024). The primary cause of this phenomenon is the entrapment and resonance of seismic energy within the softer surface layer, often leading to devastating effects on infrastructure above.

The *fo* is calculated using ground vibration measurements to quantify this site-specific amplification effect. These vibration measurements, obtained during earthquake events or from ambient noise recordings, are processed to calculate Horizontal to Vertical Spectral Ratio (HVSR) curves, which are then utilized for natural frequency estimation (Nakamura, 1989; SESAME Project, 2004; Haghshenas et al., 2008). The HVSR technique has become a standard method in earthquake engineering due to its simplicity and effectiveness in characterizing local site effects. Past studies have also utilized *fo* for site characterization relating it to site class similar to Vs30 as given in Table 1 (Zhao et al., 2006; Fukushima et al., 2007; JRA, 2019; Laouami, 2020). However, despite the critical importance of *fo* in seismic hazard assessment and the proven success of remote sensing approaches for Vs30 estimation, no comprehensive methodologies currently exist for predicting *fo* at regional scales for the Himalayan region. This represents a significant gap in the ability to conduct reliable, cost-effective, regional scale seismic site characterization using *fo* prediction.

Table 1
Predominant/Natural frequency (*fo*) based site-classification system proposed in past studies. For *fo* – JRA (2019), Zhao et al. (2006), Fukushima et al. (2007), for Site Class, Vs30 and SPT-N – BSSC (2015).

Site Class	Description	*fo* (Hz)	Vs30 (m/s)	SPT-N
B	Rock	> 5	760–1500	> 50
C	Hard soil	2.5–5	360–760	15–50
D	Medium soil	1.66–2.5	180–360	< 15
E	Soft soil	< 1.66	< 180	–

1.1. Research objectives

The Himalayan region presents particular challenges for predominant frequency (*fo*) -based seismic site characterization due to its geological complexity, high seismic activity, and rapidly expanding urban centers. Current approaches rely predominantly on ground-based measurements that are spatially limited and insufficient for regional-scale hazard assessment. The availability of high-resolution global DEM datasets such as SRTM (NASA JPL, 2013) and GLO-30 (European Space Agency, 2024) provides an opportunity to develop DEM-based methodologies for comprehensive seismic site characterization. To predict *fo* using application of geological information, bedrock depth data and DEM-based parameters for seismic site characterization, our study aims to:

1. Establish a comprehensive regional database by systematically compiling predominant frequency measurements from multiple published sources across the Himalayan region.
2. Extract and evaluate DEM-derived predictors by developing geomorphometric parameters from GLO-30 DEM data (elevation, slope, curvature, terrain indices) and assessing their predictive capability for *fo* estimation along with geological and subsurface variables.
3. Implement a regression kriging framework integrating stacked ensemble machine learning with spatial geostatistical modelling to optimize prediction accuracy across heterogeneous terrain types.
4. Generate high-resolution predominant frequency (*fo*) maps for three Himalayan capital regions using the developed framework, providing regional-scale seismic site characterization.

This study demonstrates the application of satellite-derived topographic variables combined with geological classifications and bedrock depth information as predictors for predominant frequency (*fo*) estimation, contributing to the methodological development of quantitative remote sensing for seismic site characterization. We implement a two-stage hybrid framework that integrates stacked ensemble machine learning with residual kriging to capture both nonlinear terrain-frequency relationships and spatial autocorrelation patterns. This approach differs from previous methodologies through its regional-scale application utilizing a larger dataset across diverse geological settings rather than basin-specific analyses, its focus on direct *fo* prediction rather than indirect Vs30-based inference, and its comprehensive morphometric analysis from readily available GLO-30 DEM data.

Previous study on *fo* prediction for the Kathmandu basin found that geomorphometric secondary variables did not enhance interpolation accuracy due to limited data availability, with only distance to bedrock outcrops showing improvement (Trevisani et al., 2021). Our study demonstrates that morphometric parameters can effectively contribute to *fo* prediction when combined with extensive spatial coverage and advanced ensemble techniques. The methodology is designed for regional seismic zonation applications where cost-effective satellite-based assessment is advantageous, though site-specific engineering applications require field validation. This integrated approach addresses critical data scarcity constraints in earthquake-prone developing regions, with potential applications extending to other seismically active areas worldwide.

2. Data

2.1. Existing studies and measurements

We compiled approximately 4400 *fo* data points from 26 published studies across the Himalayan region (Table 2). Fig. 1 shows these measurement sites, which are concentrated in and around major urban centers, with most located in India, followed by Bangladesh and Nepal. All these studies used the HVSR technique to determine *fo* values. The

Table 2
List of all studies considered for the dataset preparation classified according to sites' geolocation information (or GI): S1 – sites with reported/known lat and long values, and S2 – Sites with locations marked on a map.

ID	Study	Tests*	GI	Datapoints	Region*
1	Singh et al. (2019)	HVSR	S1	34	AR
2	DST (2007)	HVSR, SPT	S1	141	AS
3	Kuldeep et al. (2022)	HVSR, ANI	S2	54	AS
4	Ansary and Arefin (2020)	HVSR	S1	92	DAC
5	Ansary et al. (2022)	HVSR	S2	580	DAC
6	NCS (2016)	HVSR, MASW, CS	S1	511	DL
7	Mundepi et al. (2010)	HVSR	S2	115	DL
8	Mahajan et al. (2021)	HVSR	S2	191	HP
9	Kumar et al. (2023)	HVSR, MASW	S2	44	HP
10	Sandhu et al. (2022)	HVSR	S1	19	HR
11	Zahoor et al. (2023)	HVSR, MASW	S2	119	J&K
12	Gupta et al. (2023)	HVSR	S2	125	J&K
13	Paudyal et al. (2013)	HVSR	S1	172	NP
14	Chamlagain et al. (2025)	HVSR	S2	48	NP
15	Trevisani et al. (2021)	HVSR	S1	39	NP
16	Gupta and Kumari (2023)	HVSR	S1	194	NI
17	Walling et al. (2009)	HVSR	S1	35	OD
18	Mishra et al. (2020)	HVSR	S1	52	SK
19	Rahman et al. (2018)	HVSR, MASW, SPT	S1	167	SYL
20	Chowdhuri et al. (2011)	HVSR, SPT	S2	74	TR
21	Shankar et al. (2021a)	HVSR, MAM	S2	180	UP
22	Shankar et al. (2021b)	HVSR	S2	75	UP
23	Kundu et al. (2024)	HVSR	S2	53	UP
24	Kumar et al. (2025)	HVSR	S1	81	UK
25	Mundepi et al. (2015)	HVSR	S2	182	UK
26	Nath et al. (2015)	HVSR	S2	1063	WB

* Abbreviations: HVSR – Horizontal to Vertical Spectral Ratio, MASW – Multichannel Analysis of Surface Waves, ANI – Ambient Noise Interferometry, CS – Crosshole Seismic, SPT – Standard Penetration Test. Regions' abbreviations: Arunachal Pradesh – AR, Assam – AS, Bihar – BR, Dhaka – DAC, Delhi – DL, Himachal Pradesh – HP, Punjab – PB, Haryana – HR, Jammu and Kashmir – J&K, Nepal – NP, Northern-India – NI, Odisha – OD, Sikkim – SK, Sylhet – SYL, Tripura – TR, Uttarakhand – UK, Uttar Pradesh – UP, West Bengal – WB.

data sources varied in their measurement approaches: some studies recorded ambient vibrations under normal conditions, while others captured earthquake-induced vibrations during seismic events.

Several studies combined both measurement types. Previous studies have applied various spatial interpolation techniques to create regional site characterization maps from their *fo* measurements. These include Kriging (Walling et al., 2009; Singh et al., 2019; Trevisani et al., 2021), Natural Neighbor Interpolation (Kuldeep et al., 2022), Inverse Distance Weighting (Zahoor et al., 2023), and Nearest Neighborhood Interpolation (Gupta et al., 2023). Additionally, several studies provided Vs30 contour maps for site characterization (DST, 2007; Mundepi et al., 2010; Kuldeep et al., 2022; Kumar et al., 2023). For studies presenting only contour maps, we extracted *fo* values at the original measurement locations identified in the studies. Since the source studies employed exact interpolation methods (Kriging, IDW, Natural Neighbor), the interpolated contour values at these measurement points equal the original observed values. We interpolated representative values from contour intervals specifically at these georeferenced measurement locations, ensuring extracted data represents actual field observations rather than interpolated intermediate points.

Most existing studies have not incorporated geomorphological variables in their analysis. Trevisani et al. (2021) represents a notable exception, having incorporated terrain characteristics in their analysis of the Kathmandu basin. However, they concluded that geomorphometric information offered limited benefits for *fo* prediction due to insufficient data, except for distance to outcropping bedrock. This limited exploration of topographic variables suggests potential for further investigation of DEM-derived predictors in natural frequency estimation.

2.2. Georeferencing and data extraction

The compiled studies varied in their spatial data reporting. Some provided precise geographic coordinates, while others presented measurement locations only through maps with marked points (S1 and S2, respectively, in Table 2). For the latter category, we applied systematic georeferencing procedures following established protocols (Yao Xiaobai, 2020). Our georeferencing process involved three key steps. First, we identified Ground Control Points (GCPs) on published maps using recognizable features. Second, we registered these points to standard coordinate systems using geographic information systems. Third, we verified accuracy by cross-checking multiple reference features, including rivers, lakes, major roads, railways, administrative boundaries, and urban development patterns. The georeferencing procedure was performed in QGIS (QGIS Development Team, 2023) environment. The detailed methodology of this procedure is explained in Thakur and Anbazhagan (2025).

For *fo* value extraction, we employed two methods based on the data format. Studies with tabulated coordinates allowed direct extraction of *fo* values. For studies presenting only contour maps, we interpolated representative values from contour intervals at each measurement location. This systematic approach ensures consistent data quality across all sources, providing a robust dataset for developing and validating our DEM-based prediction methodology.

2.3. fo Dataset features

Fig. 2 shows the distribution of *fo* data points across 4440 sites in the final dataset from the Himalayan region and Indo-Gangetic Plains. Based on the JRA (2019) criteria (Table 1), most sites fall within the Soft soil category (Site Class E – 3395 sites), followed by Medium soil (Site Class D – 356 sites) and Hard soil (Site Class C – 335 sites). Compared to sites in soil categories, relatively few sites have f_0 values that fall under the Rock category (Site Class B – 354 sites).

The spatial distribution of *fo* values (Fig. 1) reflects distinct regional geological controls. Low-frequency sites are concentrated in deep sediment-filled areas, including Dhaka, Delhi's Yamuna floodplain, Kolkata's deltaic region, and areas near the Brahmaputra River around Guwahati, where thick alluvial deposits can amplify longer-period seismic waves. High-frequency sites occur at locations with shallow bedrock or stiff soils, found both at elevated areas within these same cities and in Himalayan settings such as the Kangra Valley foothills and steep terrain locations. This pattern demonstrates the fundamental control of subsurface geology and topography on seismic site response across the Indo-Gangetic Plains and Himalayan foreland.

2.4. GLO-30 DEM, bedrock depth and geological dataset

In the present study, we have used the GLO-30 Digital Elevation Model (DEM) for *fo* predictions. The Copernicus GLO-30 DEM, derived from TanDEM-X bistatic SAR interferometry (2010–2016), provides global 30-m resolution topographic data with ~2 m relative vertical accuracy and ~ 4 m absolute vertical accuracy, representing a significant improvement over legacy DEMs for seismic applications (Wessel et al., 2018). We employed GLO-30 DEM and its morphometric derivatives to enhance our *fo* prediction models for seismic site characterization. The dataset's 30-m resolution supports multi-scale terrain analysis, capturing local site effects and regional geological controls on natural frequency that complement traditional Vs30-based site classification schemes (Geyin and Maurer, 2023). GLO-30's enhanced detection of sedimentary basins, where *fo* is expected to correlate strongly with basin geometry and sediment thickness, combined with its limited vegetation penetration capability for identifying surface structure

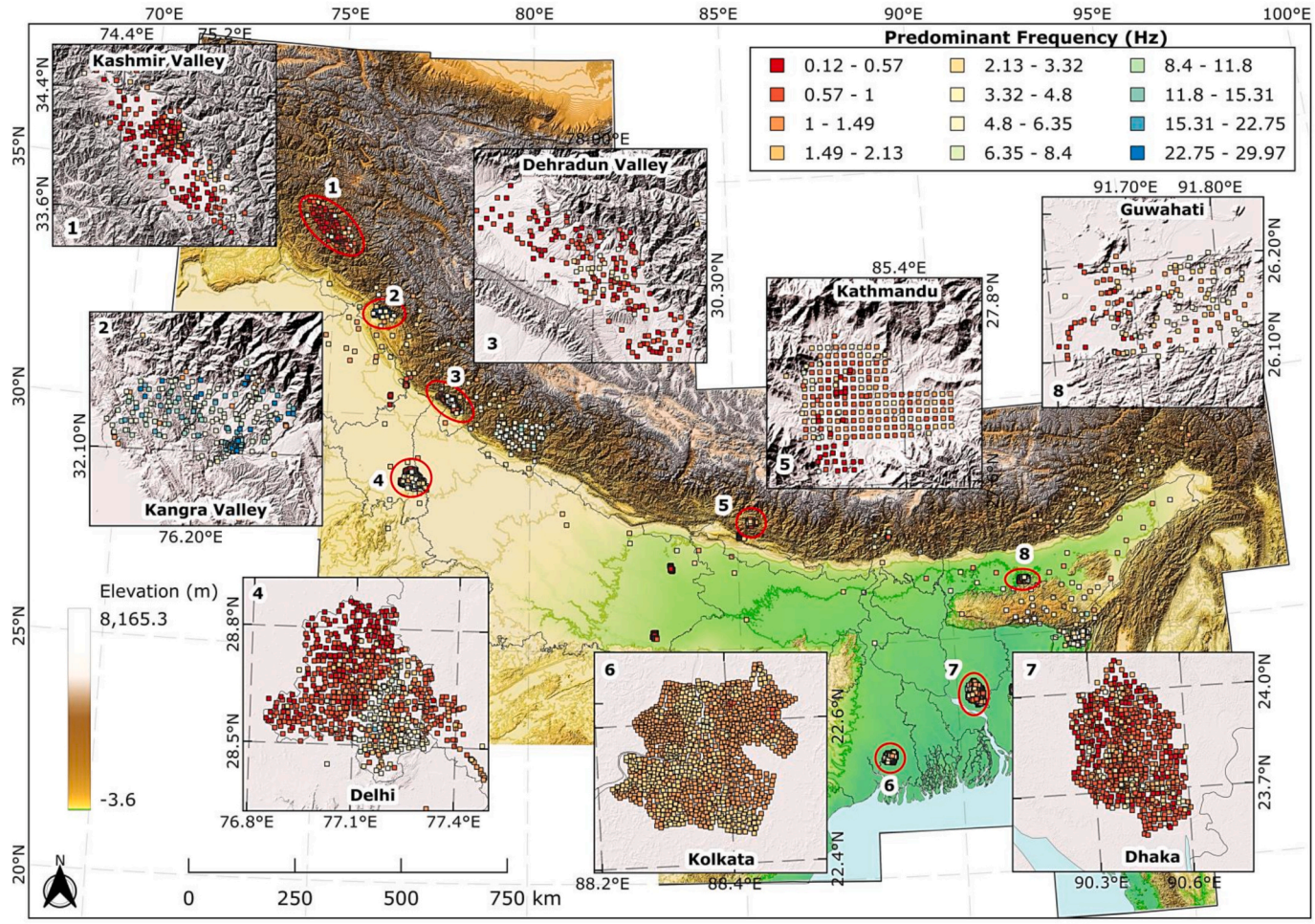


Fig. 1. *Main Map*: Final Predominant/Natural Frequency (*fo*) datapoints (see legend at top right) obtained for different locations in the Himalayan region using the geolocation data from reports, published articles and the metric georeferencing procedure. The DEM used as a base layer in the main map for Hillshade and Elevation profile (see legend at bottom left) has a resolution of 50 m. Jagged lines on the main map are contour lines at a 50 m interval. *Insets*: Numbers in inset maps are the locations marked on the main map.

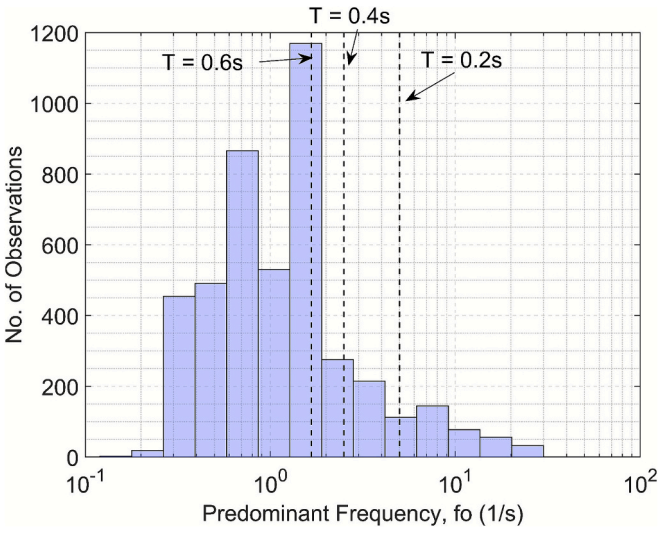


Fig. 2. Histogram of all the datapoints collected from the past studies as listed in Table 2. Vertical lines show the class boundaries as described in Table 1.

information, provides an opportunity for *fo* prediction across diverse geological settings. Fig. 3 (a) and (b) show the variation of *fo* values for the site locations as shown in Fig. 1, with respect to elevation (in meters) and slope (in percentage). The site's elevation and slope values in the figure correspond to a resampled raster with a resolution of 50 m × 50 m using the GLO-30 DEM.

Bedrock depth data were obtained from the global depth-to-bedrock dataset developed by Shangguan et al. (2017), which provides comprehensive spatial coverage at 250 m resolution. This dataset was constructed using machine learning algorithms (Random Forest and Gradient Boosting Tree) applied to approximately 1.3 million soil profile locations and 1.6 million borehole locations worldwide, along with 155 environmental covariates including DEM-based morphological derivatives, lithologic units, and MODIS surface reflectance data. The dataset represents the most comprehensive global compilation of bedrock depth information available for seismic site characterization studies, providing the fundamental geological constraint necessary for natural frequency prediction in this study. Fig. 3 (c) shows the variation of *fo* values with respect to the bedrock depth values obtained from the Shangguan et al. (2017) dataset. Bedrock values obtained from the dataset for the sites shown in Fig. 1 range between 0 and 4500 m.

The geological classification data for the present study were derived from Wandrey (1998), which provides a comprehensive digital geologic map of South Asia developed under the U.S. Geological Survey World Energy Project (also see Thakur and Anbazhagan, 2025). This dataset encompasses Afghanistan, Bangladesh, Bhutan, India, Myanmar, Nepal, Pakistan, and Sri Lanka, compiled from numerous UNESCO geological

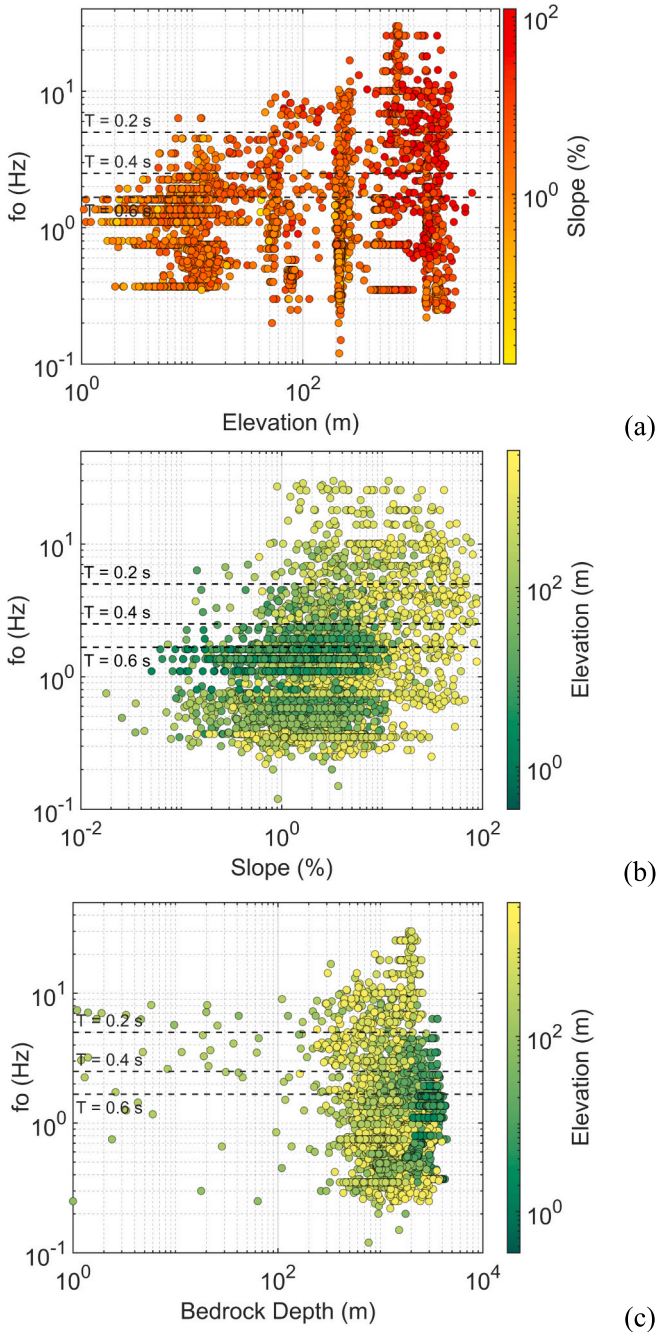


Fig. 3. Scatter plots for the catalogued dataset's sites' Predominant Frequency (*fo*) values with respect to (a) Elevation (m), (b) Slope (%), (c) Bedrock Depth (m). The colour ramp shows parameter values for the same locations. The elevation and slope values for the sites correspond to a resampled raster with a resolution of 50 m × 50 m using the GLO-30 DEM. Bedrock depth values correspond to a 250 m resolution raster using the global bedrock depth map by Shangguan et al., 2017.

maps and national geological surveys at scales ranging from 1:500,000 to 1:10,000,000. The geological units are systematically classified according to the World Energy Project standard, incorporating geologic age and general lithologic characteristics, with attributes including stratigraphic units and regional geological contacts. Based on this dataset, most of the sites in the present study fall in the 'Quaternary sediments (Q)' category per the dataset. Fig. 4 shows the proportion of different geological classes (top 6) for the *fo* sites as a percentage of all sites. In this, categories such as 'undivided Precambrian Rocks (pC)' and

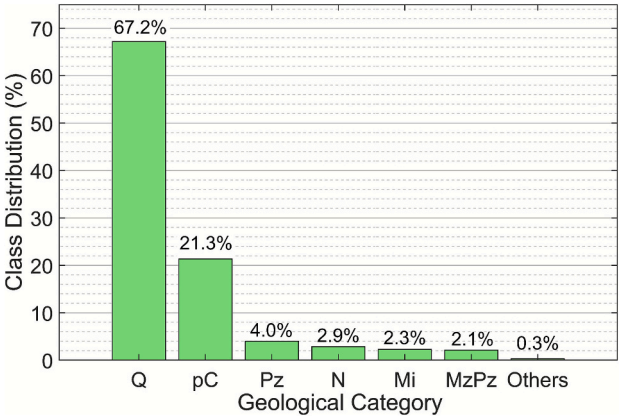


Fig. 4. Proportion of different geological classes for the *fo* site locations as a percentage of the total data points. Here, Q – Quaternary sediments, pC – undivided Precambrian rocks, Pz – Undifferentiated Paleozoic rocks, N – Neogene sedimentary rocks, Mi – Mesozoic intrusive rocks, MzPz – Paleozoic and Mesozoic metamorphic rocks.

'undifferentiated Paleozoic rocks (Pz)' were most prevalent after the Quaternary sediment category.

3. Methodology

We have adopted regression kriging to prepare *fo* maps (Hengl et al., 2007; Hengl and MacMillan, 2019). This approach combines the deterministic and stochastic parts of spatial variation for the final prediction for a location s_0 given as:

$$z_{pred}(s_0) = m_{pred}(s_0) + e_{pred}(s_0) \quad (1)$$

here $z_{pred}(s_o)$ is the final predicted value at a new location s_o, $m_{pred}(s_o)$ is the prediction from the Machine Learning (ML) model, and $e_{pred}(s_o)$ is the interpolated residual for the location using residual kriging. At the first stage, an ML model is developed to predict the *fo* using different predictors. This stage implements stacked ensemble learning, where multiple base models are trained on extracted features. The best-performing models' predictions were combined to develop a stacked ML model. At the second stage, the stacked model's residuals are spatially interpolated using ordinary kriging. The results from kriging are used to update the model's prediction values around a sampled point.

In the next sections, we explain the following steps: model prediction parameters and their extraction, ML model selection for prediction, and the residual kriging procedure adopted for the final predictions. The regression kriging procedure for natural frequency prediction in seismic site characterization employs a two-stage methodology that integrates machine learning ensemble techniques with spatial geostatistics. This approach aims to effectively capture nonlinear terrain-frequency relationships and residual spatial correlation patterns that persist after initial modelling. The stacked ML model development was done in MATLAB (2023), kriging was performed in R (R Core Team, 2024) using *gstat* package (Pebesma, 2004), and the predictors extraction, *fo* site characterization maps preparation was done in a GIS environment.

3.1. Model prediction parameters and their extraction

Twenty predictor variables were selected for *fo* prediction ML model development, comprising seventeen continuous terrain morphometric parameters and three categorical geological classifications (Table 3 and Fig. 5). Of the twenty predictors considered for the *fo* prediction model, 17 are continuous and three are categorical. The continuous variables include six topographic parameters – (Abbasnejadfard et al., 2023) Elevation, (Akin et al., 2011) Slope, (Allen and Wald, 2009) Profile curvature, (Amatulli et al., 2018) Plan curvature, (Ansary and Arefin,

Table 3
A list of all the parameters/predictors used for the predominant Frequency (*fo*) prediction. Ranges of parameters correspond to the resampled GLO-30 DEM at a 50 m resolution. Here C_V – Categorical Variable.

ID	Variable	Abbrev.	Units	Range	Reference Study
		Geometric Predictors			
1	Elevation	Elev	m	0 to 3380	–
2	Bedrock Depth	BD	m	0 to 4453	–
		Topographic Predictors			
3	Slope	Slop	%	0 to 122.7	–
4	Topographic Position Index	TPI	m	−24.3 to 17.93	Jenness (2006)
5	Terrain Ruggedness Index	TRI	m	0 to 44	Riley et al. (1999)
6	Vector Ruggedness Measure	VRM	–	0 to 0.0846	Sappington et al. (2007)
7	Roughness	Rough	m	0 to 154.33	
		Geological Predictors			
8	Geology	Geol	C_V	–	–
		Curvature Predictors			
9	Profile Curvature	Prof_Curv	m^{-1}	−0.0084 to 0.0077	
10	Plan Curvature	Plan_Curv	m^{-1}	−3.275 to 0.935	Minár et al. (2020)
11	Total Curvature	Total_Curv	m^{-1}	0 to 0.00018	
12	Tangential Curvature	Tang_Curv	m^{-1}	−0.012 to 0.011	
		Hydrological Predictors			
13	Nearest distance to the river	Dist_R	m	0 to 4670.1	–
14	Topographic Wetness Index	TWI	–	8.42 to 18.42	Sörensen et al. (2006)
		Basin Geometry Predictors			
15	Nearest distance to basin boundary	Dist_B	m	0 to 6232.7	–
		Geomorphometric Predictors			
16	Morphometric Feature	MF	C_V	1 to 6	–
17	Geomorphons	GM	C_V	1 to 10	Jasiewicz and Stepinski (2013)
18	Multiresolution Index of Valley Bottom Flatness	MRVBF	–	2.6E-14 to 8.99	Gallant and Dowling (2003)
19	Multiresolution Index of Ridge Top Flatness	MRRTF	–	4E-15 to 6.98	Gallant and Dowling (2003)
20	Morphometric Protection Index	MPI	–	0 to 0.472	–

2020) Total curvature, and (Ansary et al., 2022) Tangential curvature; six terrain analysis indices – (Brando et al., 2020) Topographic Position Index (TPI), (Breiman, 1996) Terrain Ruggedness Index (TRI), (BSSC, 2015) Topographic Wetness Index (TWI), (Chamlagain et al., 2025) Vector Ruggedness Measure (VRM), (Chen et al., 2024) surface roughness, and (Chowdhuri et al., 2011) Morphometric Protection Index (MPI); two landform characteristics (Gallant and Dowling, 2003) – (Delgado et al., 2000) Multiresolution Valley Bottom Flatness (MRVBF) and (Di Martino et al., 2024) Multiresolution Ridge Top Flatness (MRRTF); two distance-based parameters – (Dong and Shan, 2013) distance to nearest basin boundary and (DST, 2007) distance to nearest river; and one critical subsurface parameter – (European Space Agency, 2024) bedrock depth. The categorical variables include (Freund and Schapire, 1997) Geological formation type, (Fukushima et al., 2007) Geomorphons landform classification, GM, and (Gallant and Dowling, 2003) Morphometric features classification, MF.

These predictors were calculated using the DEM layer, bedrock depth layer, and geological layers. The terrain morphometric parameters were derived from the GLO-30 DEM using established geomorphometric analysis techniques in a GIS environment (Amatulli et al., 2018). The distance-based predictors were calculated using a river network layer and basin boundary layer for the Himalayan region from HydroSHEDS database (Lehner and Grill, 2013). Distance variables were calculated in the GIS to determine proximity to basin boundaries and drainage networks that can influence sediment deposition and site response characteristics.

These variables were selected based on their theoretical and potential relationships to the fundamental 1D site response equation $fo = Vs/(4H)$ (Kramer and Stewart, 2024), where bedrock depth directly represents the impedance contrast parameter H, while topographic and geological variables serve as proxies for velocity structure (Vs) and local site effects (Zhao et al., 2006). The combination of the subsurface parameters (such as bedrock depth and geology) and DEM-based terrain characteristics provides comprehensive coverage of factors controlling *fo* variations across diverse geological environments.

Next, a correlation analysis was done on the extracted predictors. The analysis was conducted on a filtered dataset of 3970 observations (89.4 % retention rate) comprising 31 predictor variables: 17 continuous variables and 14 categorical dummy variables representing the top 6 geological formations (98.4 % coverage), top 4 morphological features (99.2 % coverage), and top 7 Geomorphons (91.3 % coverage). This was done to enhance visualization clarity. Categorical variables were filtered to retain only the most frequent categories, reducing matrix overcrowding while maintaining high data representation and excluding rare categories with significantly smaller sample sizes. The 465 correlations were distributed as strong ($|r| > 0.5$): 30 correlations, moderate ($0.3 < |r| \leq 0.5$): 45 correlations, and weak ($|r| \leq 0.3$): 390 correlations. In geological categories, Quaternary deposits (Geol_Q in Fig. 6) showed the strongest topographic associations with negative correlations to elevation ($r = -0.676$) and positive correlations to valley bottom flatness (MRVBF, $r = 0.781$), indicating preferential occurrence in low-lying terrain. Precambrian rocks (Geol_pC in Fig. 6) demonstrated contrasting patterns with positive correlations to elevation ($r = 0.413$) and negative correlations to valley bottom flatness (MRVBF, $r = -0.586$), suggesting occurrence in elevated terrain. Morphological features and Geomorphons showed moderate correlations with topographic variables, notably MF_3 with topographic position index (TPI, $r = -0.422$) and Geomorphons GM_3 with TPI ($r = 0.414$). The systematic correlation patterns validate the inclusion of these categorical predictors in the present study.

3.2. Dataset partitioning scheme and workflow

For the implementation of regression kriging and to obtain unbiased residuals for the kriging step while incorporating stacked models (SMs) for prediction, we used the following data partitioning and model-building scheme (also see Fig. 7):

1. *Initial Data Partitioning*: The complete dataset was divided using 5-fold cross-validation, creating five folds – FN1, …, FN5 (Fig. 7). For each fold FNi, the remaining four folds formed the training set (D_i) and fold FNi became the test set (Te_i).
2. *Training Set Subdivision*: Each training set D_i was further split into:
   - Training subset (Tr_i): 85 % of D_i for base model training.
   - Validation subset (Va_i): 15 % of D_i for meta-learner training
3. *Base Model Training*: For each Tr_i, multiple base models were trained using 5-fold cross-validation within Tr_i to prevent overfitting.
4. *Base Model Selection*: Base models were evaluated on Va_i using RMSE, MAE and R^2 (Table 4). The top two performing models were selected for each stacked model SM_i.
5. *Meta-learner/ Stacked Model Training*: The predictions from the two selected base models on Va_i were used as features to train a stacked model for SM_i. For this stage a number of model types were trained and best performing was selected for SM_i.
6. *Final Predictions*: Each stacked model SM_i generated predictions on its corresponding test set Te_i.

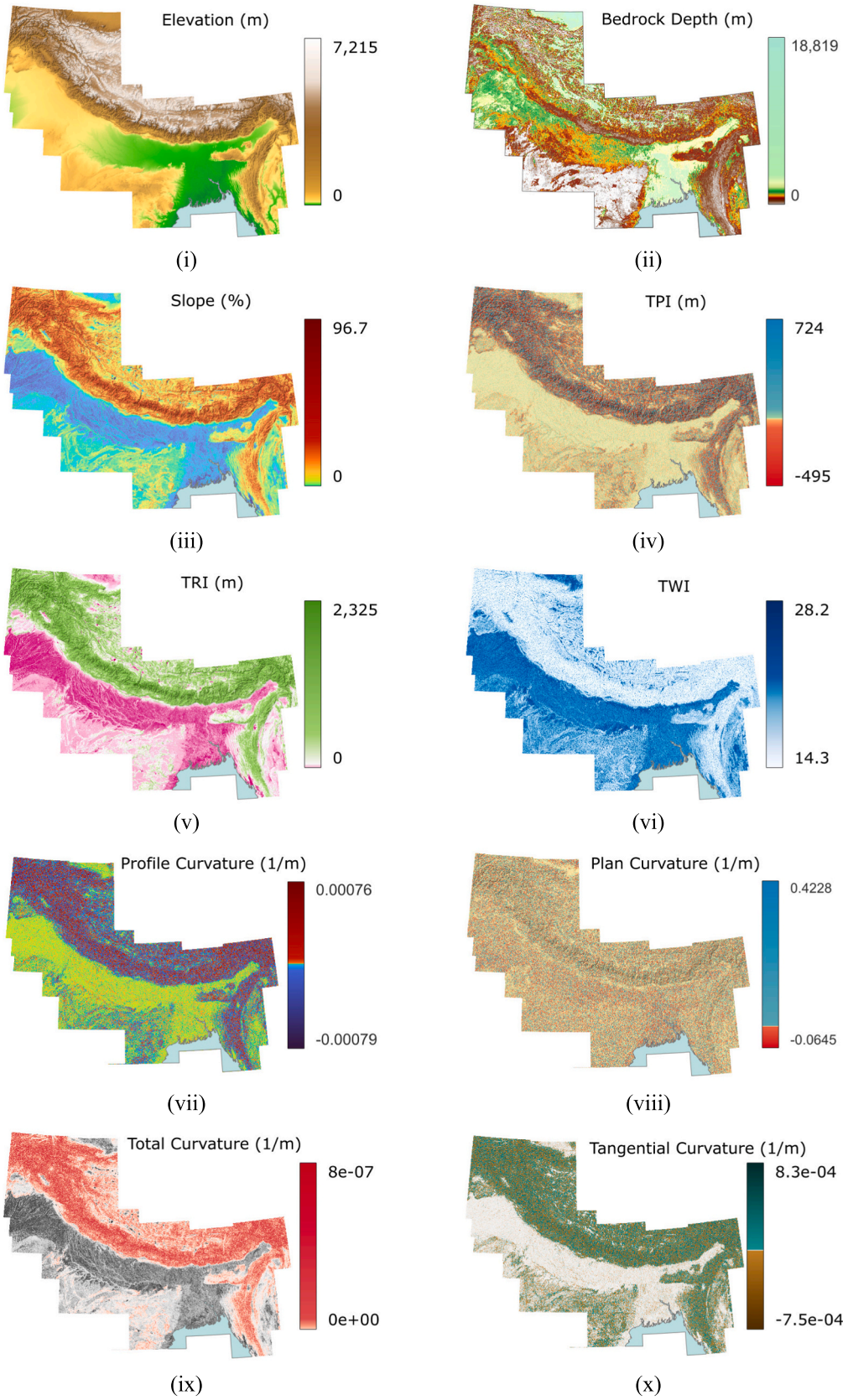


Fig. 5. Maps of predictors considered in the present study at a 1 km resolution for the region of interest spanning India, Nepal, Bangladesh and Bhutan.

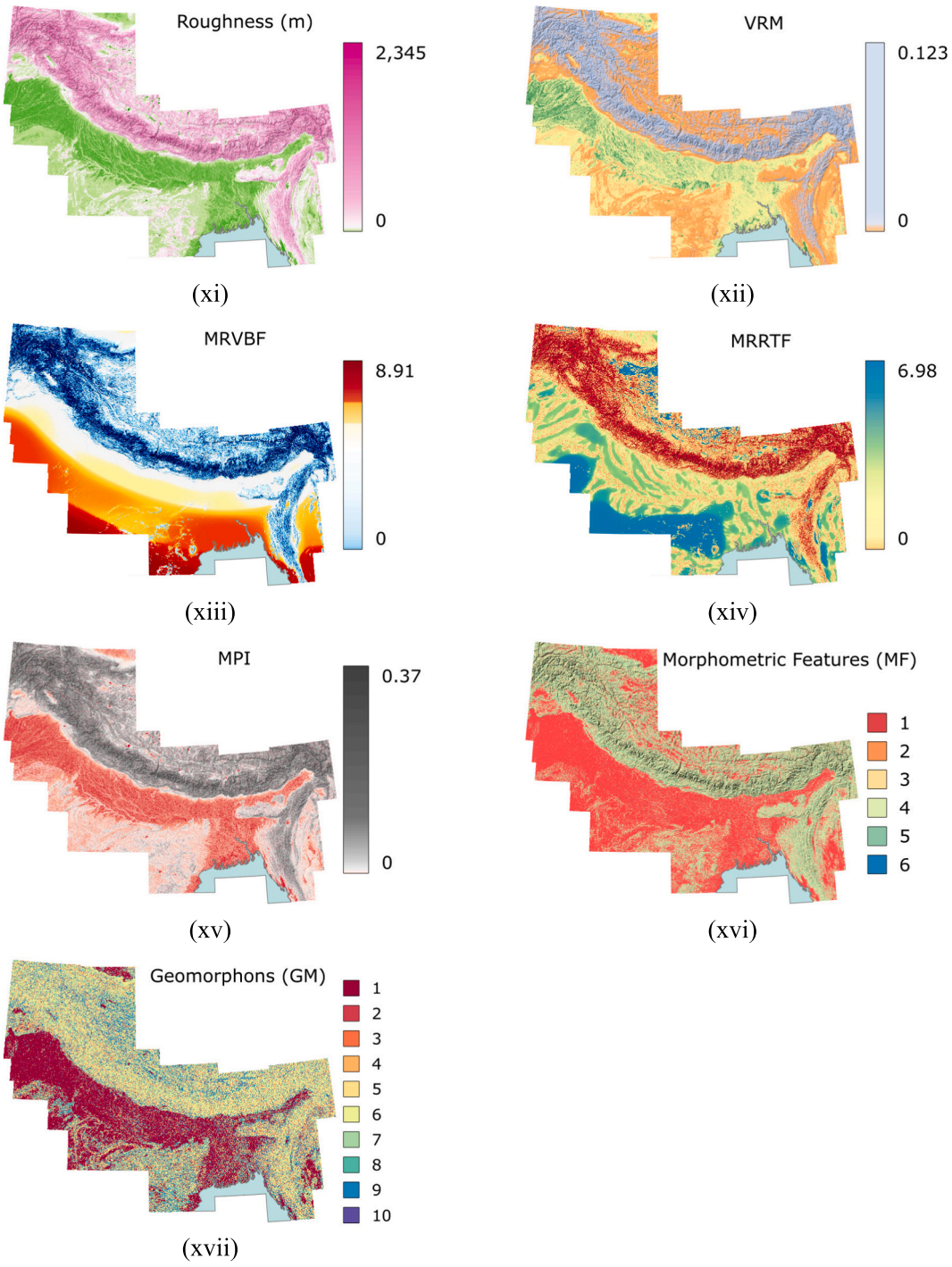


Fig. 5. (*continued*).

7. *Residual Calculation*: Residuals were calculated as (observed - predicted) for each test point. Since each data point appears in exactly one test set, this ensured unbiased residuals.
8. *Kriging*: The combined residuals from all test sets were used for spatial kriging to capture the remaining spatial correlation.

For the site characterization maps preparation, the average of all stacked models' predictions was used for prediction on a new location with known/extracted values of predictors.

3.3. First stage – Building stacked ensemble model

For the stacked ensemble construction, we began with a comprehensive evaluation of multiple machine learning algorithms to identify optimal base models for each cross-validation fold. Following the data partitioning scheme outlined in Section 3.2, we systematically tested different algorithm families, including ensemble methods, Neural networks, Gaussian Process Regression (GPR), Linear regression models, and Support Vector Machines (SVM). Each algorithm was evaluated using 5-fold cross-validation within the training subset to ensure robust performance assessment while preventing overfitting.

For base model evaluation we employed Root Mean Square Error

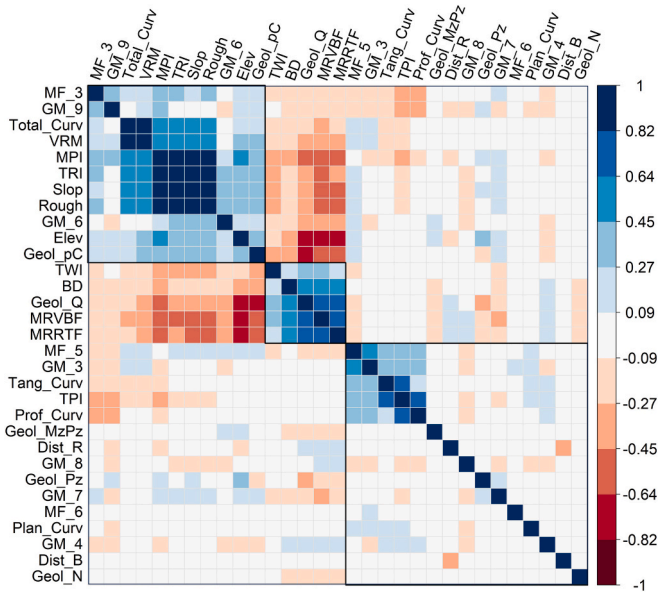


Fig. 6. Correlation matrix of all predictors for Predominant Frequency (*fo*), showing relationships between geomorphometric parameters, geological formations (Geol_), morphological features (MF_), and Geomorphons' types (GM_). Variables are clustered by similarity using hierarchical clustering.

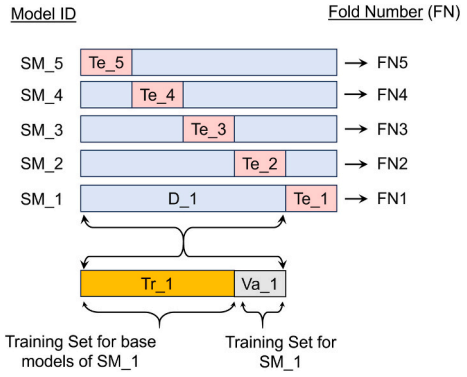


Fig. 7. Dataset partitioning scheme adopted in the present study. SM_i is the stacked model for Fold Number (FN) i.

Table 4
List of performance measures calculated and used for the two stages of model building.

Model Performance Measure	Formula*
Stage – 1. Stacked Ensemble Model	
Root Mean Square Error (RMSE)	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
Mean Absolute Error (MAE)	$MAE = \sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$
Coefficient of Determination (R^2)	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$
Stage – 2. Residual Kriging	
Residual Sum of Squares (RSS)	$RSS = \sum_{i=1}^{m}\left[\gamma_{empirical}(h_i) - \gamma_{model}(h_i)\right]^2$
Akaike Information Criterion (AIC)	$AIC = n \times \ln\left(\frac{RSS}{n}\right) + 2k$

* Notations: n – number of data points, y_i – observed value for the ith datapoint, $\hat{y}$ – predicted value for the ith datapoint, $\bar{y}$ - mean value of the observations, m – number of lag classes, k – number of model parameters.

(RMSE), Mean Absolute Error (MAE), and coefficient of determination (R^2) as primary selection criteria (Table 4). The evaluation process prioritized prediction accuracy and algorithmic diversity to ensure complementary error patterns in the final ensemble. Multiple ML algorithms were evaluated for base model selection, with hyperparameters optimized using Bayesian optimization. During hyperparameter optimization, fifty trials were evaluated for each model, with the RMSE being minimized. Algorithm-specific parameters were systematically tuned: ensemble parameters (learning cycles, learning rate, minimum leaf size), Gaussian process parameters (kernel functions and hyperparameters), and support vector machine parameters (regularization and kernel scale). Performance evaluation using cross-validation metrics (Table 4) was done to identify the top two algorithms per fold for the stacked ensemble framework.

The selected base models' interpretability was assessed using SHAP values (Lundberg and Lee, 2017), which decompose each natural frequency prediction into individual predictor contributions relative to the model baseline. Shapley values quantify how terrain morphometry (curvature, elevation) and subsurface properties (bedrock depth, geology) influence site-specific predictions, with magnitude indicating impact strength and sign showing directional effect. Mean Shapley values reveal which terrain and subsurface features most strongly control *fo* predictions in our seismic site characterization models. Feature importance plots were generated for the base models to ensure robust interpretability and validate consistency with seismic site characterization principles.

Our algorithm selection was guided by established principles in seismic site characterization modelling. Tree-based ensemble methods (bagged and boosted trees) were prioritized due to their demonstrated effectiveness in capturing complex, nonlinear relationships between terrain morphometry and geophysical site parameters (Geyin and Maurer, 2023; Thakur and Anbazhagan, 2025). These methods inherently handle mixed-type predictors (continuous morphometric parameters and categorical geological classifications) without requiring explicit interaction term specification and are robust to the multi-scale input data integration characteristic of regional geophysical modelling. Gaussian Process Regression was evaluated for its ability to model spatial correlations and provide uncertainty estimates, while Support Vector Machines were tested for their capacity to identify complex decision boundaries in high-dimensional feature spaces.

To improve the robustness of the prediction, as discussed, we have used model stacking in the first stage. Model stacking is an ensemble learning technique that combines multiple diverse base models through a meta-learner to achieve better predictions than any individual model (Vilalta and Drissi, 2002). Prevalent methods for model stacking are bagging (Breiman, 1996), random forest (Ho, 1995), and boosting (Freund and Schapire, 1997). It works in the following stages: first, base models are trained on the dataset (Tr_i) using cross-validation to avoid overfitting. Next prediction on a test set (here Va_i) was then done using these base models to generate out-of-fold predictions. Then, a meta-learner/ stacked model is trained to combine these base model predictions optimally. The stacked model learns the best way to weight and combine the base models' outputs, leveraging their different strengths and compensating for individual weaknesses. This approach prevents overfitting through cross-validation and reduces bias and variance, resulting in more robust predictions than simple averaging or any single model alone.

3.4. Second stage – Residuals' Kriging

Following the initial stacked machine learning prediction, residual kriging was employed as the second stage to capture and model spatial autocorrelation in prediction residuals (Hengl and MacMillan, 2019). In the second stage, we applied ordinary kriging to the unbiased residuals (ϵ) given as:

$$\epsilon(S_i) = y(S_i) - y_{ML}(S_i) \quad (2)$$

where $y(S_i)$ and $y_{ML}(S_i)$ represent observed and ML predicted *fo*

values for the site S_i, respectively. Residuals from the stacked ML model ($n = 4440$) were examined for spatial structure using variogram analysis. The spatial correlation structure of these residuals is characterized through empirical variogram analysis. The empirical variogram was computed using the method of moments estimator:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \left[(Z(x_i) - Z(x_i + h))^2 \right] \quad (3)$$

where $\gamma(h)$ is the semivariance at lag distance h, $N(h)$ is the number of point pairs separated by distance h, and $Z(x_i)$ represents the residual values (Hengl and MacMillan, 2019). For the variogram, a maximum lag distance (cutoff) of 15,000 m was applied to focus on the relevant spatial scale, with lag intervals of 500 m to ensure adequate point pairs per bin.

Theoretical variogram models were fitted to capture *fo* variability. Eight theoretical variogram models were evaluated, including simple models (Nugget, Spherical, Exponential, Gaussian, Matérn, and Linear) and nested model combinations (Spherical+Nugget and Exponential+Nugget). Anisotropy was assessed by computing directional variograms at four principal directions: 0° (East-West), 45° (Northeast-Southwest), 90° (North-South), and 135° (Northwest-Southeast), each with a tolerance angle of ±22.5°. Model selection was based on Residual Sum of Squares (RSS) and Akaike Information Criterion (AIC) as given in Table 4; and visual inspection of model fit to empirical variogram. The geostatistical analysis was performed using the gstat package in R (Pebesma, 2004).

Ordinary kriging of residuals was performed using the best-performing variogram model. Spatial interpolation of natural frequency residuals was conducted using a two-stage approach combining kriging with K-nearest neighbours (KNN) gap filling on a 50-m resolution grid covering different regions. The primary kriging stage involved converting global residuals to spatial objects using UTM coordinates,

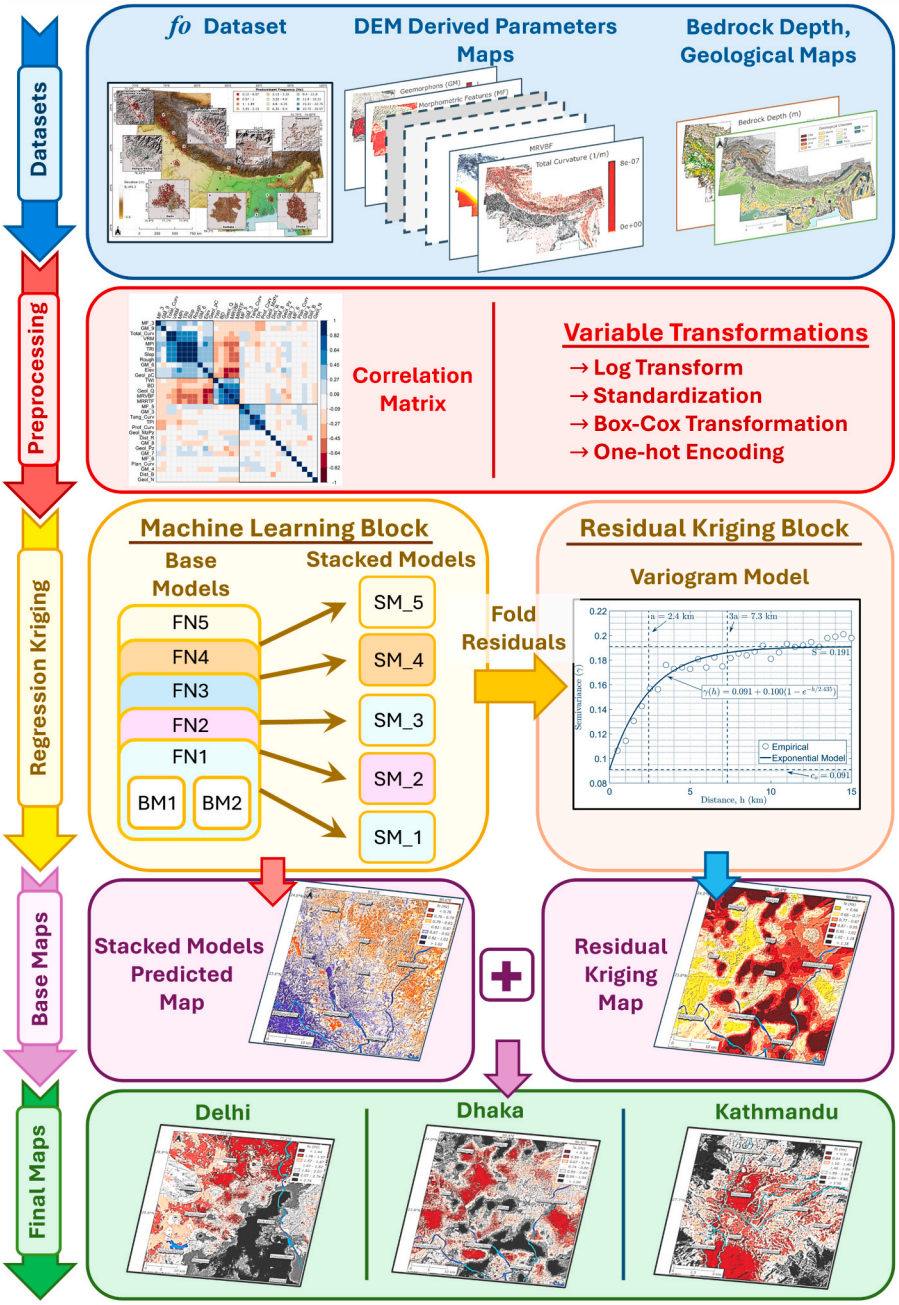


Fig. 8. Workflow showing the steps followed in the present study for the Predominant Frequency (*fo*) prediction for seismic site characterization.

with training data selected within a 30-km buffer around the region grid to ensure adequate local representation. The search parameters were selected to ensure adequate local data representation while maintaining computational tractability for the high-resolution grid. The secondary gap-filling stage addressed locations where ordinary kriging could not produce predictions due to insufficient local data density or extreme spatial configurations. This stage employed K-nearest neighbours interpolation using four nearest valid predictions through simple averaging.

3.5. Map preparation

Fig. 8 shows the workflow followed in the present study for the final site characterization maps for *fo*. The present study's two-stage approach combines the strength of machine learning algorithms in capturing complex nonlinear relationships with geostatistical methods' ability to model spatial dependencies (Hengl and MacMillan, 2019; Geyin and Maurer, 2023). We used average values obtained from the five stacked ML models for the final map preparation. The ML models' predictions and the residual kriging were combined to get the final *fo*-based seismic site characterization maps.

For the final map preparation for major capital regions in the region, we have chosen a 50 m resolution. The 50 m output grid spacing enables visualization of local topographic variations captured by the GLO-30 DEM and facilitates integration with standard urban planning datasets. However, the effective prediction resolution is constrained by the coarsest input variable – the 250 m bedrock depth dataset. The 50 m visualization grid should not be interpreted as implying corresponding precision in *fo* predictions at that scale. The 50 m grid facilitates visualization of terrain-controlled patterns but predictions remain most reliable at regional scales. Site-specific *fo* values will require field validation, particularly where complex local geology or anthropogenic modifications are not captured by regional predictors.

4. Results and discussion

4.1. Stacked ensemble model's performance evaluation

4.1.1. Base models

Table 5 presents performance metrics for the base models selected for each cross-validation fold (FN1 through FN5). The base models were developed for log-transformed *fo*. The model screening phase evaluated multiple algorithm types, including ensemble methods, Gaussian process regression (GPR), various tree-based models, and support vector machines (SVMs). From this comprehensive evaluation, the selection process consistently identified bagged ensembles as the first base model (BM1) and boosted ensembles as the second base model (BM2) across all five folds. Despite testing diverse algorithm families, this uniform selection outcome indicated that tree-based ensemble methods most effectively captured the nonlinear relationships in the *fo* prediction.

The results show that bagged ensembles achieved validation RMSE values of 0.609–0.637 and test RMSE of 0.593–0.636, while boosted ensembles showed validation RMSE of 0.636–0.655 and test RMSE of 0.633–0.662 (Table 5). The corresponding R^2 values averaged 0.507 for bagged models and 0.471 for boosted models on validation sets, with test R^2 values of 0.537 and 0.494, respectively. These results demonstrated that bagged ensembles consistently outperformed their boosted counterparts with lower prediction errors and higher explained variance. The small differences observed between validation and test performance suggested adequate generalization capability. Notably, FN3-BM1 improved from validation (RMSE = 0.630) to test (RMSE = 0.593), achieving the best test performance among all base models with $R^2 = 0.585$. The analysis of mean absolute error values, which ranged from 0.414 to 0.432 for bagged models and 0.459 to 0.479 for boosted models, yielded MAE/RMSE ratios of approximately 0.68–0.73, indicating consistent error distributions without excessive outlier influence.

The exclusive selection of bagged and boosted ensembles from the broader pool of tested algorithms highlighted their relatively better performance for our application. The systematic pairing of these two ensembles reflected their complementary characteristics: bagged ensembles reduce variance through bootstrap aggregation, while boosted ensembles reduce bias through sequential error correction. This complementarity supported their combination through stacked model, as each method addressed different sources of prediction error. The observed performance consistency across folds demonstrated stable predictions despite training data variations. The moderate R^2 values obtained indicated that while the models captured meaningful variance, substantial unexplained variation remained. These results justified the subsequent application of spatial interpolation techniques to capture localized effects not represented in the predictor variables. The dominance of ensemble methods for our present dataset over other tested algorithms suggests that the complexity and non-linearity of *fo* prediction require the flexibility of tree-based approaches.

Figs. 9 and 10 provide the plots of the quantitative performance assessment obtained from the base model selection process. Fig. 9 displays the predicted versus observed value plots for the best-performing base models (FNi-BM1 and FNi-BM2), where the scatter of points relative to the diagonal line reflects the models' predictive accuracy on their respective test sets. The distribution pattern shows that the bagged and boosted ensemble methods capture the underlying relationships between DEM-derived morphometric parameters and *fo* measurements. Fig. 10 presents the residual plots for these same models, revealing the

Table 5
Results of all stacked models (SMs) and their corresponding base models (BMs) using performance measures listed in Table 4 for validation (V) and test (T) sets. These results are in natural log units. Here, FNi – Fold Number i, SM_i – Stacked Model i, SVM – Support Vector Machines.

Stacked Model ID	Base Model ID	Model Type	RMSE (V)	R^2 (V)	MAE (V)	RMSE (T)	R^2 (T)	MAE (T)
SM_1		Linear Regression	0.621	0.538	0.418	0.633	0.508	0.434
	FN1-BM1	Bagged Ensemble	0.609	0.529	0.414	0.636	0.503	0.437
	FN1-BM2	Boosted Ensemble	0.639	0.483	0.459	0.656	0.471	0.482
SM_2		Linear SVM	0.627	0.539	0.421	0.622	0.517	0.416
	FN2-BM1	Bagged Ensemble	0.637	0.487	0.431	0.629	0.534	0.420
	FN2-BM2	Boosted Ensemble	0.654	0.46	0.479	0.662	0.484	0.479
SM_3		Robust Linear	0.588	0.595	0.402	0.587	0.553	0.401
	FN3-BM1	Bagged Ensemble	0.630	0.505	0.432	0.593	0.585	0.401
	FN3-BM2	Boosted Ensemble	0.655	0.464	0.471	0.633	0.527	0.451
SM_4		Linear Regression	0.617	0.542	0.420	0.608	0.509	0.418
	FN4-BM1	Bagged Ensemble	0.624	0.519	0.428	0.620	0.537	0.426
	FN4-BM2	Boosted Ensemble	0.655	0.469	0.470	0.650	0.491	0.465
SM_5		Linear Regression	0.625	0.541	0.414	0.626	0.540	0.417
	FN5-BM1	Bagged Ensemble	0.614	0.515	0.419	0.632	0.528	0.414
	FN5-BM2	Boosted Ensemble	0.636	0.478	0.460	0.653	0.497	0.469

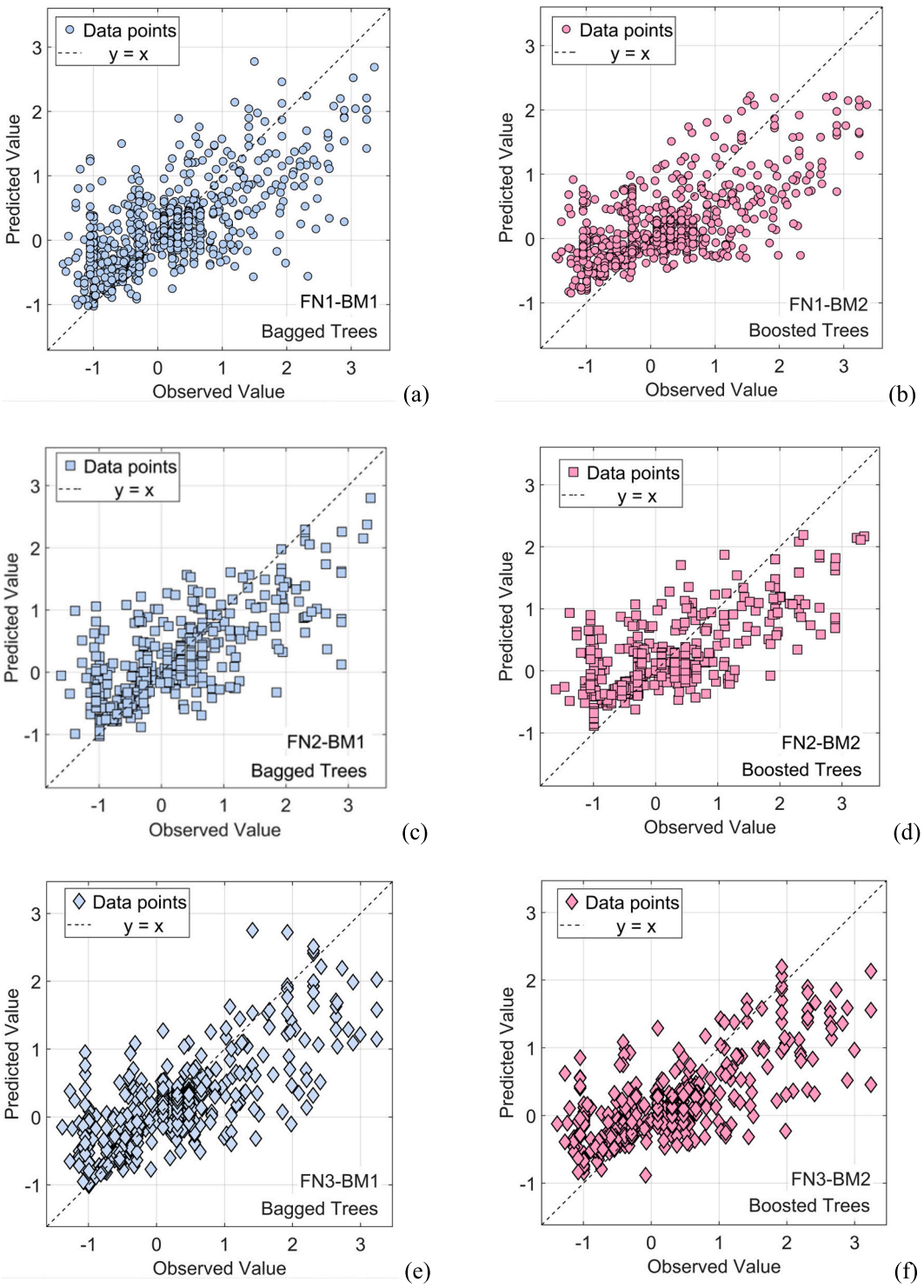


Fig. 9. Test set's predicted vs observed values plots for a particular fold number FNi (i ∈[1,5]) for their two best-performing base models – BM1 in (a), (c), (e), (g), and (i); and BM2 in (b), (d), (f), (h), and (j). Here, the test set is Va_i for a particular fold i and the values are in natural log units.

distribution of prediction errors across the range of fitted values. The residual patterns indicate no systematic bias exists in the model predictions. These figures substantiate the performance metrics showing RMSE values between 0.593 and 0.662 for the selected base models and R^2 values ranging from 0.471 to 0.585.

4.1.2. Stacked models

The performance metrics for the five stacked ensemble models (SM_1 through SM_5), each developed using their respective cross-validation folds, are presented in Table 5. Each stacked model incorporates two optimally selected base models identified through model screening, with the meta-learner combining their predictions to enhance robustness and generalization capability. Linear regression appeared in 3 of the five base learner positions, while linear support vector machines and robust linear regression appeared only in SM_2 and SM_3, respectively. The meta-learners (bagged and boosted ensembles) combine the log-

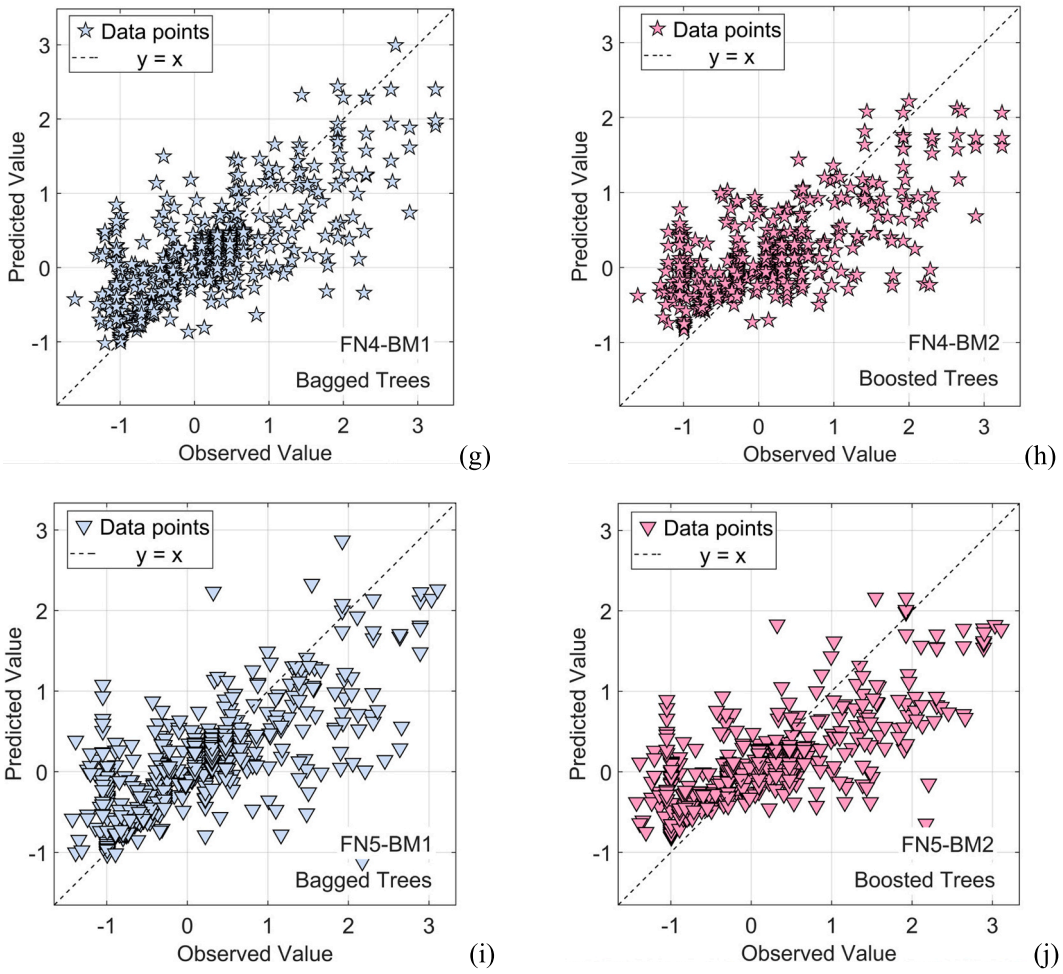


Fig. 9. (*continued*).

transformed base model predictions, which ensure positive predictions after back-transformation.

Validation sets' performance showed RMSE values ranging from 0.588 to 0.627 in log units and R^2 values between 0.538 and 0.595. Test sets' performance yielded RMSE values between 0.587 and 0.633 and R^2 values from 0.509 to 0.553. The small difference between validation and test performance of the stacked models indicates limited overfitting. The final prediction methodology averages all five stacked models to improve robustness. Averaging models trained on different data subsets reduces prediction variance by approximately $1/\sqrt{5}$. Each fold covers slightly different spatial and geological characteristics, so averaging mitigates fold-specific biases. The stacked ensemble approach provides stable natural frequency predictions for spatial interpolation and operational use. While individual base models sometimes achieved lower error rates, the ensemble framework here offers prediction stability.

Fig. 11 shows the prediction and observed value plots for the stacked models' on their respective test folds (Te_i in Fig. 7). These shows consistently scatter around the diagonal 1:1 line. Fig. 12 shows the residuals variation against predicted values corresponding to the data points in Fig. 11. These residual values show the random scatter around the x – axis indicating no systematic bias in the stacked models' predictions.

4.2. Feature importance analysis

To understand the impact of different predictors we conducted a variable importance analysis. The Shapley Importance analysis (see Fig. 13) shows that terrain morphometric parameters and geological classifications dominated natural frequency prediction. Multiresolution Valley Bottom Flatness (MRVBF) emerged as the most important predictor (Shapley value: ~ 0.18), reflecting the critical role of sediment accumulation zones in controlling site response characteristics. Valley bottom areas systematically correspond to deeper sedimentary deposits with strong impedance contrasts which correlate well with low *fo* values (Kramer and Stewart, 2024).

Undivided Precambrian rocks (pC) ranked second in importance (Shapley value: ~ 0.17), demonstrating the fundamental influence of crystalline basement geology on *fo* characteristics. Precambrian formations typically comprise high-velocity crystalline rocks (granites, gneisses, metamorphic complexes) that create strong impedance contrasts with overlying sediments and exhibit distinctly different seismic response characteristics compared to younger geological formations (Zhao et al., 2006).

Elevation lies in the top three positions in all the model results (Shapley value: ~ 0.15), confirming its role as a fundamental proxy for structural position and geological context. Higher elevations often correspond to areas where Precambrian basement approaches the surface or where structural uplift has exposed older, more competent geological formations with elevated predominant frequencies. Multiresolution Ridge Top Flatness (MRRTF) and bedrock depth (BD) exhibited similar moderate importance levels (Shapley values: ~ 0.04 and ~ 0.03, respectively). The relatively modest importance of bedrock depth data suggests that the combination of depositional environment identification and basement rock presence effectively captures the primary controls on impedance contrast variations.

Secondary importance variables, including Morphometric Protection

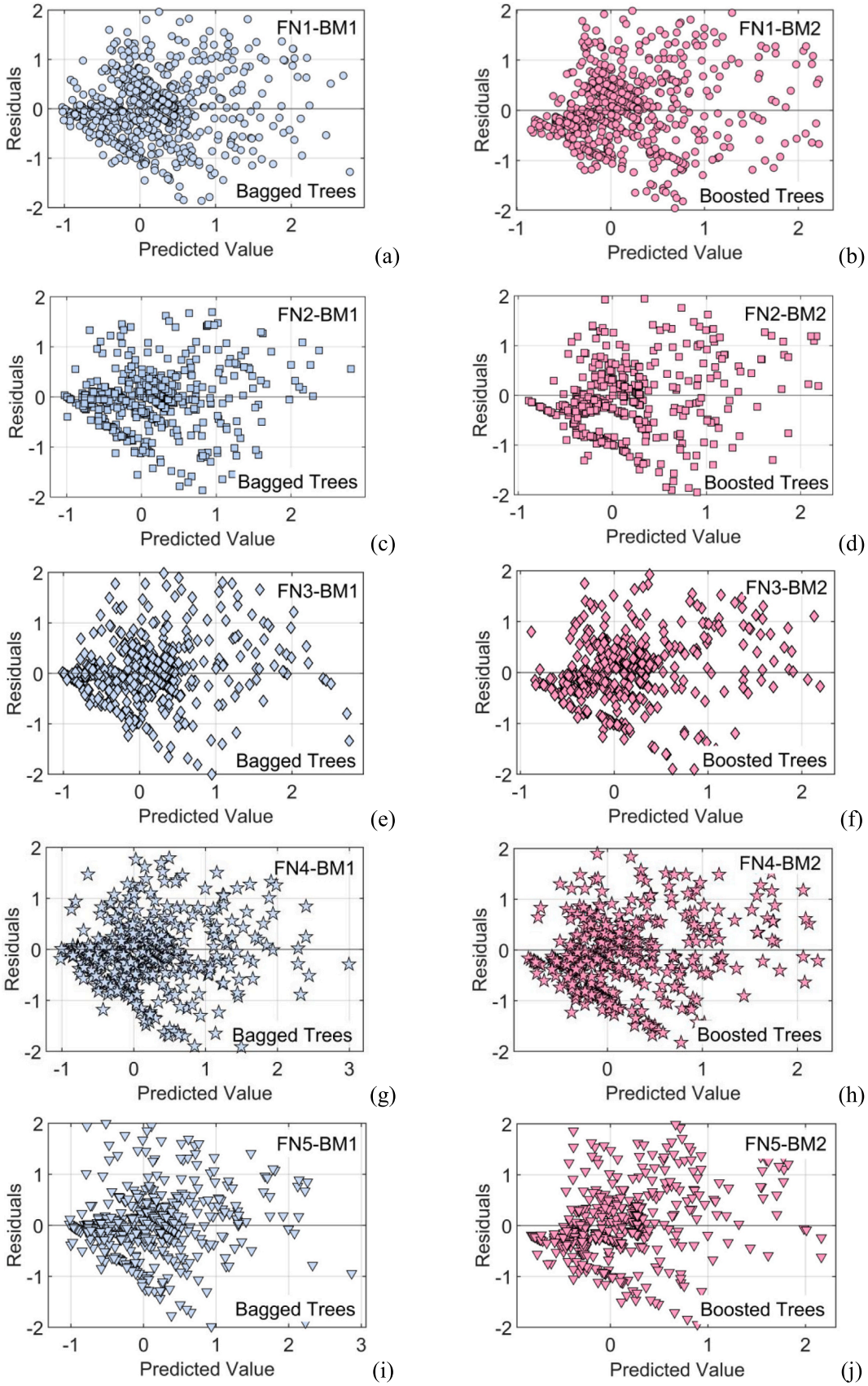


Fig. 10. Test set's residual plots for a particular fold number FNi (i ∈[1,5]) for their two best-performing base models – BM1 in (a), (c), (e), (g), and (i); and BM2 in (b), (d), (f), (h), and (j). Here, the test set is Va_i for a particular fold i and the values are in natural log units.

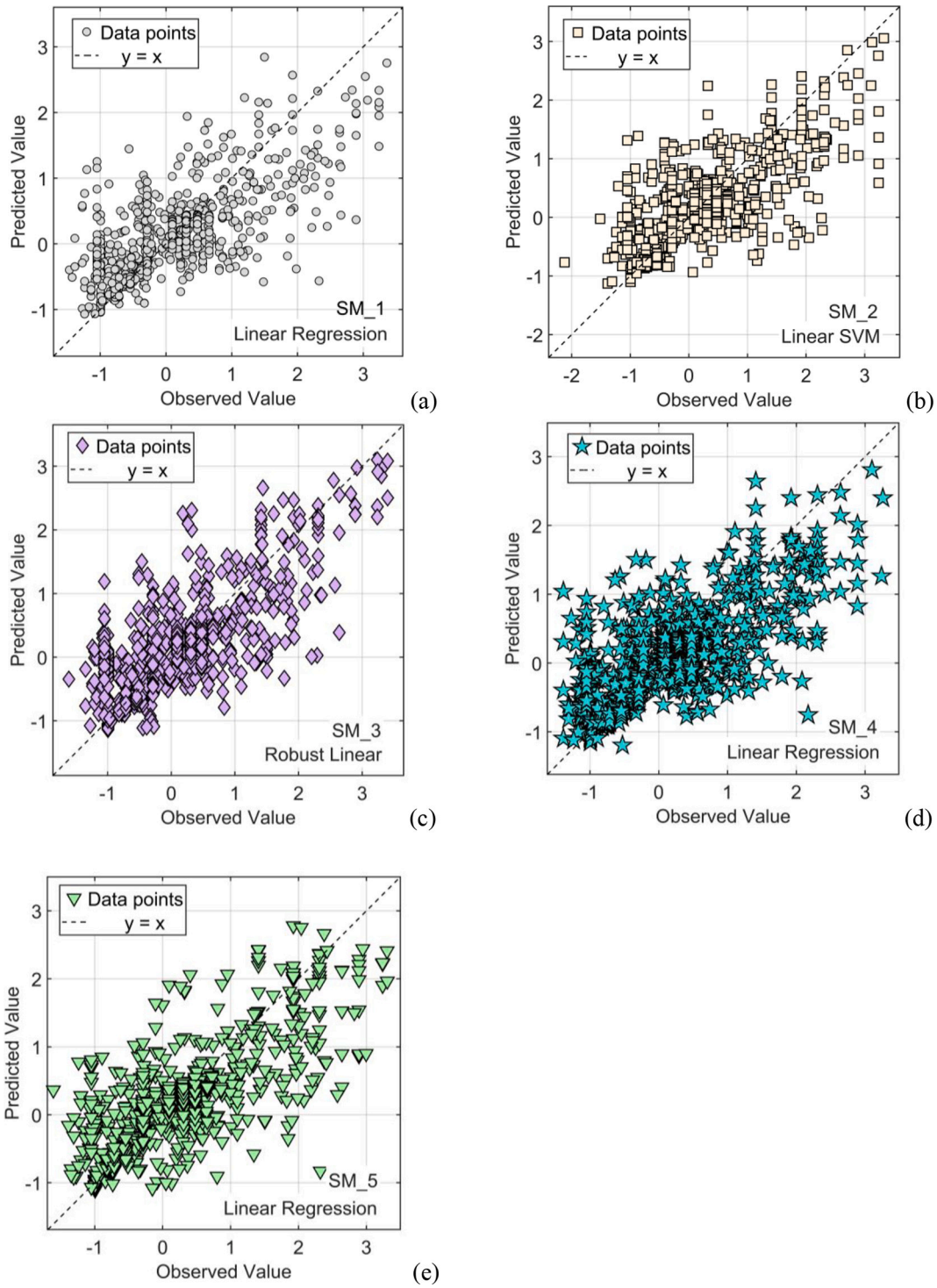


Fig. 11. Prediction vs observed value plots for stacked models (SM_i) on their respective test sets (Te_i), where i ∈[1,5]. These values are in natural log units.

Index (~ 0.03), surface roughness (~ 0.03), and slope (~ 0.025), contributed meaningful but limited predictive power through their representation of local topographic complexity and erosional-depositional processes. Notably, traditional curvature parameters, including profile curvature, plan curvature, and tangential curvature exhibited relatively low importance (< 0.01), indicating significantly less contribution to *fo* prediction in this geological context. The importance hierarchy demonstrates that sediment accumulation zone identification and Precambrian basement rock presence provide the primary predictive capability, reflecting the fundamental control of depositional processes and basement geology on seismic site response characteristics.

4.3. Residuals, Variogram and Kriging results

4.3.1. SEMs residuals

Utilizing the five stacked models' predictions on their respective test sets (Te_i) of size $n = 888$, we calculated unbiased residuals for all the locations. A summary of the statistics of these residuals is given in Table 6.

The ML model residuals ranged from −2.48 to 3.15 units with a mean

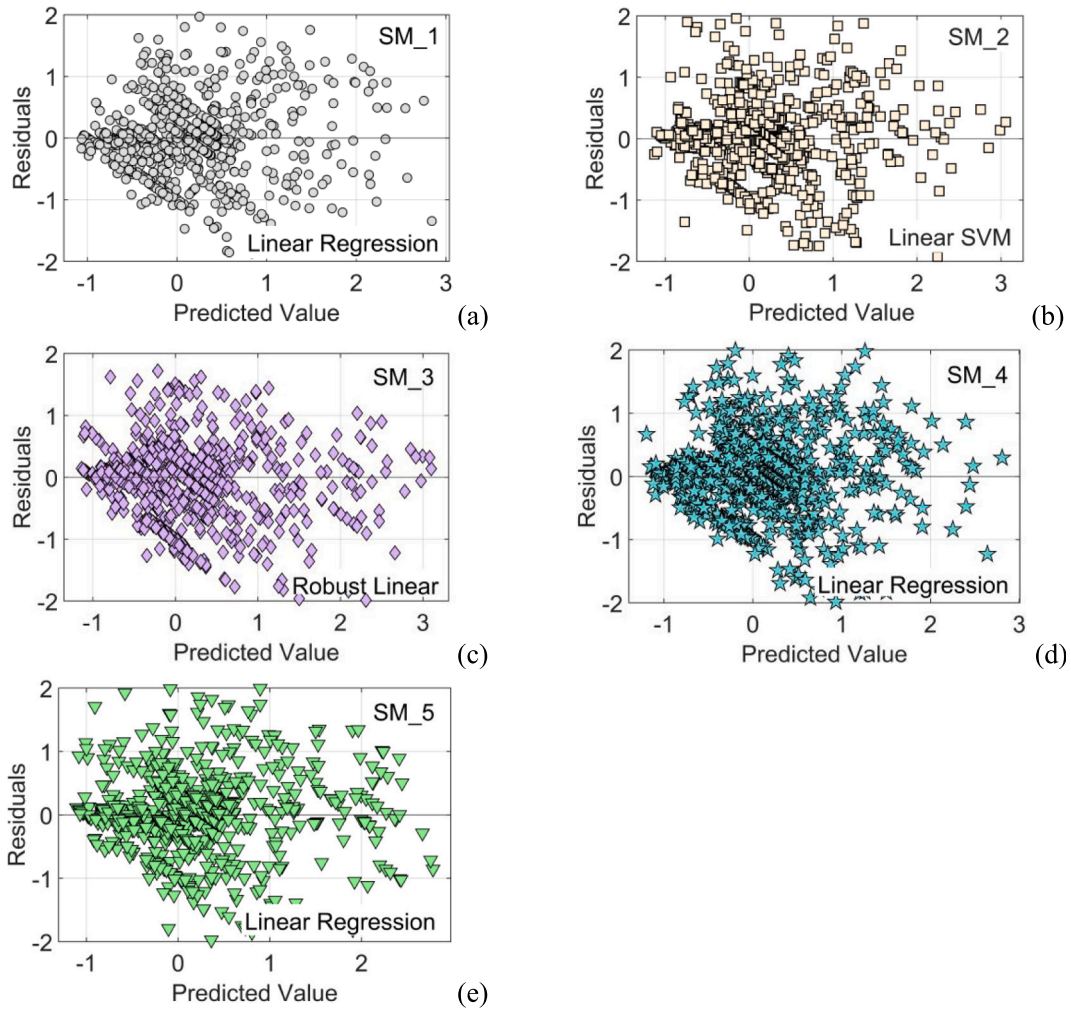


Fig. 12. Residual plots for stacked models (SM_i) on their respective test sets (Te_i), where i ∈[1,5]. These residuals were used for the kriging procedure, and their values are in natural log units.

of 0.010 ± 0.616 units (mean ± standard deviation), indicating that the stacked ML model achieved near-zero bias with moderate variability suitable for spatial analysis (Table 6). The empirical variogram was calculated using 30 lag classes with a maximum lag distance of 14,748 m, revealing clear spatial structure in the residuals with semivariance increasing systematically from short to intermediate distances before stabilizing at longer lags, showing the presence of spatial autocorrelation and has been incorporated in the prediction model using residual kriging as the second modelling stage. Directional variogram analysis was also conducted which showed isotropic spatial structure with no significant directional effects. Based on this, we use the omnidirectional variogram without incorporating anisotropic corrections.

The exponential, Matérn, and exponential+nugget models achieved nearly identical optimal fits (RSS differences <0.000001, Table 7), with the exponential model selected for its simplicity and parsimony while maintaining identical performance metrics. Following the principle of model parsimony, the simpler exponential model was preferred, as it requires fewer parameter specifications while achieving identical goodness-of-fit.

The exponential variogram model follows the form:

$$\gamma(h) = c_0 + c_1\left(1 - e^{-\frac{h}{a}}\right) \tag{4}$$

where c_0 represents the nugget effect, c_1 is the partial sill, a is the range parameter, and the effective correlation range extends to $3a$ units, representing the distance at which 95 % of the total sill is reached and spatial correlation becomes negligible.

The selected exponential model exhibited a nugget effect (c_o) of 0.091, representing 48 % of the total sill, indicating substantial short-range variation and measurement uncertainty (Fig. 14). The partial sill (c_1) of 0.1 accounts for 52 % of the total variance, representing the structured spatial correlation component. The range parameter (a) of 2435 m defines the characteristic correlation distance, while the total sill ($c_o + c_1$) of 0.191 represents the complete spatial variance. The effective range extends to 7305 m ($3a$), representing the distance at which 95 % of the total sill is reached and spatial correlation becomes negligible.

The fitted variogram parameters provide crucial insights into the spatial structure of ML residuals and optimization opportunities for the two-stage prediction approach. The nugget-to-sill ratio of 48 % indicates substantial short-range variation encompassing measurement uncertainty, micro-scale environmental heterogeneity not captured by the ML feature space. The structured spatial component (52 % of total variance) demonstrates that a considerable spatially correlated signal remains in the ML residuals, validating the residual kriging approach and suggesting meaningful improvements in prediction accuracy are achievable. The effective correlation range of 7305 m defines the spatial neighborhood within which residual values exhibit significant correlation, which can provide guidance for optimal sampling density in future data collection.

The good model fit (RSS = 0.000805, R^2 = 0.952, RMSE = 0.005179)

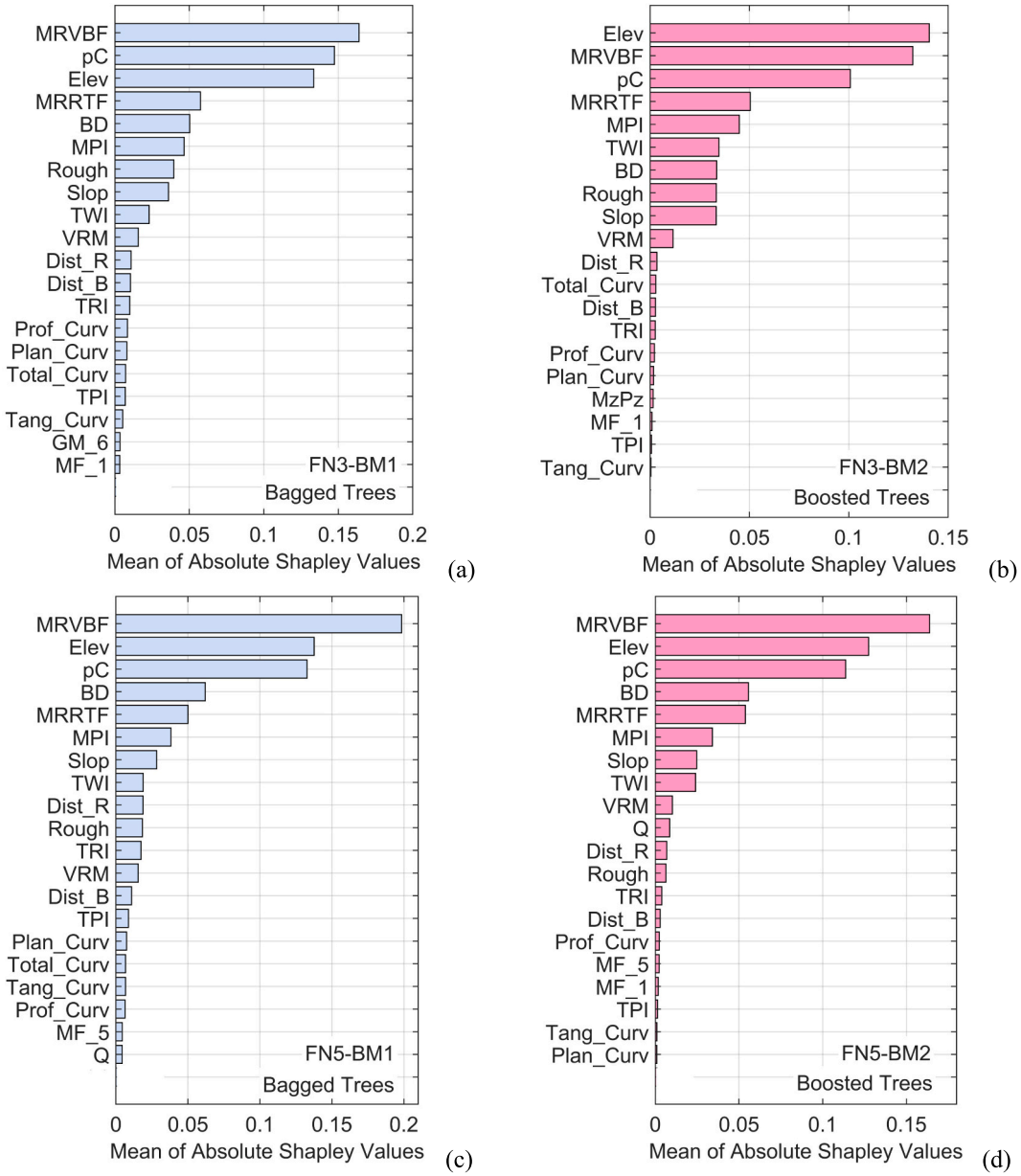


Fig. 13. Shapley importance plots for the two best-performing base models (BMs) for the stacked model SM_3 in (a) and (b), and SM_5 in (c) and (d). Query points for the Shapley importance plots were taken from the test sets (Va_i) of the base model's respective fold i. Parameters other than those mentioned in Table 3 are geological categories, morphometric features (MF) and Geomorphons (GM) classes.

Table 6
Residuals statistics for all five sets obtained using Stacked ensemble models on different folds.

ID	Count	Mean	Std	Min	Max
1	888	−0.003	0.634	−2.335	2.891
2	888	0.016	0.622	−2.363	2.650
3	888	−0.037	0.586	−2.238	2.266
4	888	0.054	0.608	−2.482	2.937
5	888	0.021	0.626	−2.011	3.152

and strong agreement between empirical and theoretical semivariance values across all 30 lag distances confirm that the exponential model successfully captures the underlying spatial covariance structure, ensuring reliable spatial predictions and uncertainty estimates for the residual kriging component of the two-stage modelling framework. Ordinary kriging was implemented using the exponential variogram model with partial sill (c_1) = 0.100, range = 2435 m, and nugget (c_0) = 0.091. The kriging procedure employed 4–16 nearest training points within a 50-km search radius under a constant mean.

4.4. Predominant frequency maps of the capital regions

4.4.1. Delhi

The residual kriging implementation for the Delhi region achieved 99.48 % spatial coverage, with ordinary kriging predicting most locations on the 50-m resolution grid. K-nearest neighbours (KNN) gap filling was applied to complete the remaining 0.52 % of locations where kriging could not produce reliable predictions due to sparse local data density, achieving complete 100 % grid coverage across the study area. The final stacked ensemble model (FSEM) predictions (Fig. 15a) show clear spatial patterns of natural frequency variation across the Delhi region. High frequency zones (>1.38 Hz) appear in the southeastern portion of the study area, corresponding to areas with shallow bedrock

Table 7
Variogram model comparison results. RSS – Residuals Sum of Squares, AIC – Akaike Information Criterion.

ID	Model	RSS	AIC	Range (m)	Sill	Nugget
1	Exponential	0.000805	−311.79	2435	0.191	0.091
2	Matérn	0.000805	−311.79	2435	0.191	0.091
3	Exponential+Nugget	0.000805	−311.79	2435	0.191	0.091
4	Spherical	0.002092	−283.13	4800	0.183	0.096
5	Spherical+Nugget	0.002093	−283.11	4799	0.183	0.096
6	Linear	0.002552	−277.16	3434	0.181	0.098
7	Gaussian	0.002642	−276.12	2091	0.180	0.105
8	Nugget	0.069230	−180.15	0	0.134	0.134

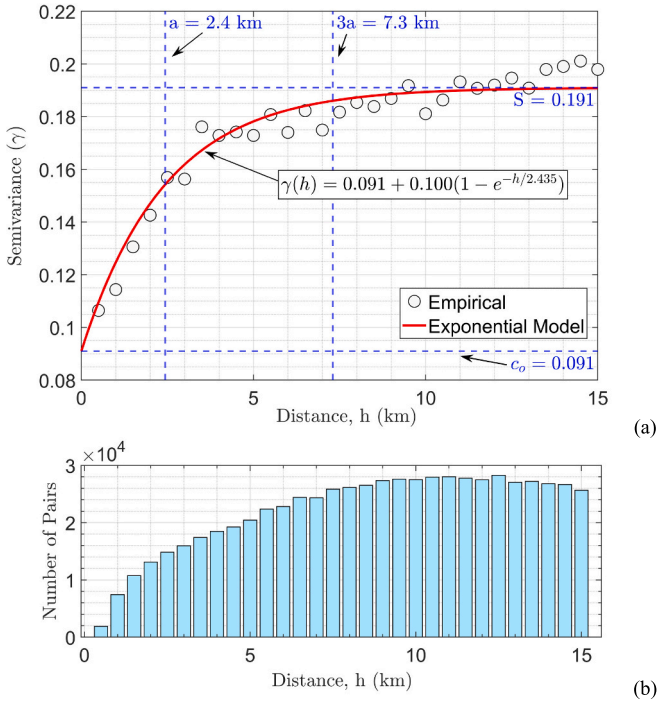


Fig. 14. Variogram modelling results for the residuals obtained from stacked ensemble models (SEMs) – (a) Exponential model parameters, (b) Variation of number of point pairs with distance corresponding to each empirical data point. Here, a – Range, S – sill ($c_o + c_1$), c_o – nugget.

and elevated topographic positions. Medium frequency regions (0.55–0.70 Hz) are located in the western sectors, while low frequency zones (<0.52 Hz) characterize the northeastern areas. The lowest predominant frequency values occur along the Yamuna River floodplains in the northeast, reflecting deep alluvial deposits and soft sedimentary conditions. A north-to-south gradient is observed, with frequencies gradually increasing from the river valley toward higher ground.

The application of residual kriging corrections (Fig. 15b) shows local improvements in prediction accuracy, particularly at and around locations where actual *fo* measurements were conducted. The final regression kriging map (Fig. 15c) integrates the regional trends captured by machine learning and the local spatial correlations modeled through geostatistical interpolation, improving prediction reliability for seismic site characterization applications.

4.4.2. Kathmandu

Residual's ordinary kriging achieved complete spatial coverage for the Kathmandu region using the global variogram model parameters derived from the residual analysis. Unlike the Delhi implementation, no gap-filling procedures were required due to the adequate spatial distribution of training data relative to the study area extent.

The FSEM predictions (Fig. 16a) show the characteristic bowl-shaped frequency distribution of the Kathmandu basin, with the lowest natural frequencies (<1.0 Hz) concentrated in the central basin floor where thick lacustrine sediments are present. Higher frequency values (> 3.2 Hz) are observed along the basin margins and surrounding hills, corresponding to bedrock outcrops and shallow soil cover. The north-south trending frequency patterns reflect the basin's structural geology, with notable low-frequency corridors following ancient lake bed deposits.

Residual kriging adjustments (Fig. 16b) show local corrections, particularly in the transition zones between the basin floor and surrounding hills, where geological contacts create sharp impedance contrasts. The final regression kriging results (Fig. 16c) capture both the large-scale basin geometry effects and localized geological influences, providing detailed natural frequency characterization for earthquake hazard assessment in this densely populated urban center.

4.4.3. Dhaka

The Dhaka region analysis employed the ordinary kriging with global variogram parameters, achieving complete spatial coverage without additional interpolation techniques. The uniform data distribution across the study area enabled consistent regional prediction quality. The FSEM map (Fig. 17a) displays the characteristic low-frequency signature of the Bengal Delta region, with most areas showing natural frequencies below 1.02 Hz, reflecting the presence of deep alluvial deposits and recent deltaic sediments. Variations are observed across the region, with higher frequencies in the western areas corresponding to older alluvial terraces, and the lowest frequencies in the eastern sectors where active floodplain processes maintain softer sedimentary conditions.

The residual kriging component (Fig. 17b) shows fine-scale spatial structure in the prediction residuals, capturing local geological variations not fully represented by the morphometric predictors. The integrated regression kriging results (Fig. 17c) provide a comprehensive natural frequency map that accounts for regional depositional patterns and local site effects, offering information for seismic hazard assessment in this expanding megacity built on unconsolidated sediments.

4.5. Methodological performance and comparative analysis

4.5.1. DEM-based natural frequency prediction performance

The stacked ensemble approach achieved test R^2 values of 0.509–0.553 and RMSE of 0.587–0.633 across cross-validation folds, demonstrating moderate predictive capability for predominant frequency (*fo*) estimation using satellite-derived morphometric parameters. The regression kriging approach provides reliable regional-scale patterns appropriate for preliminary seismic zonation and planning applications. However, the 49 % unexplained variance and multi-resolution input constraints necessitate field validation for site-specific engineering design. The performance consistency across folds indicates stable predictions despite training data variations. A comparison with Vs30 (shear wave velocity for top 30 m of soil profile) prediction methodology provide relevant context for evaluating these results. Established approaches using topographic proxies typically achieve similar performance metrics values in morphometric studies report modest R^2 ranges in diverse geological settings (Thakur and

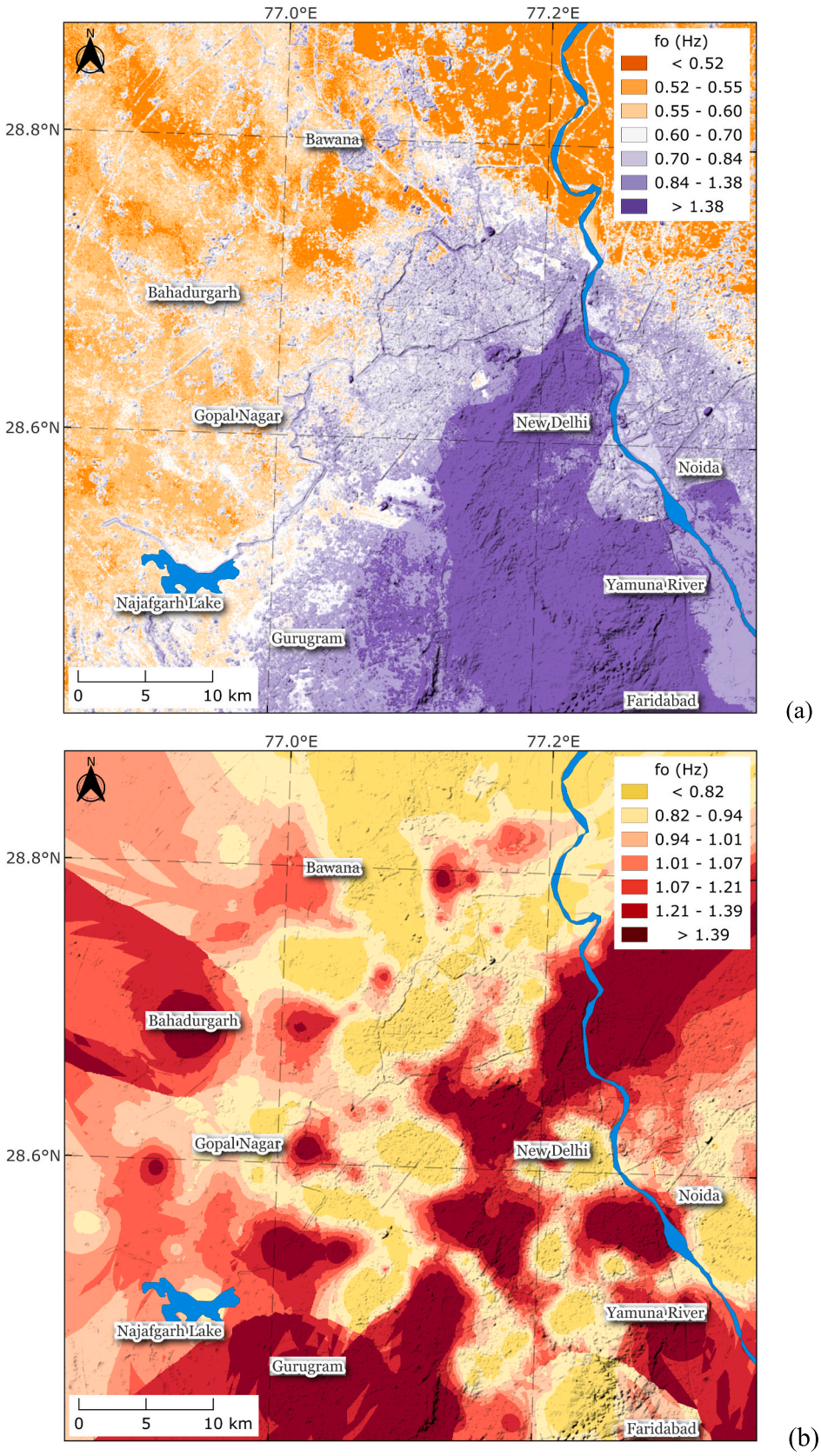


Fig. 15. Predominant Frequency (*fo* in Hz) prediction obtained for the Delhi region using (a) final Stacked Ensemble Model (FSEM), (b) residual kriging (RK), and (c) Regression Kriging (combining FSEM and RK). The map is prepared using a 50 m × 50 m grid and *fo* classes are coloured based on quantiles.

Anbazhagan, 2025). Our predominant frequency (*fo*) prediction performance falls within this range, suggesting comparable accuracy

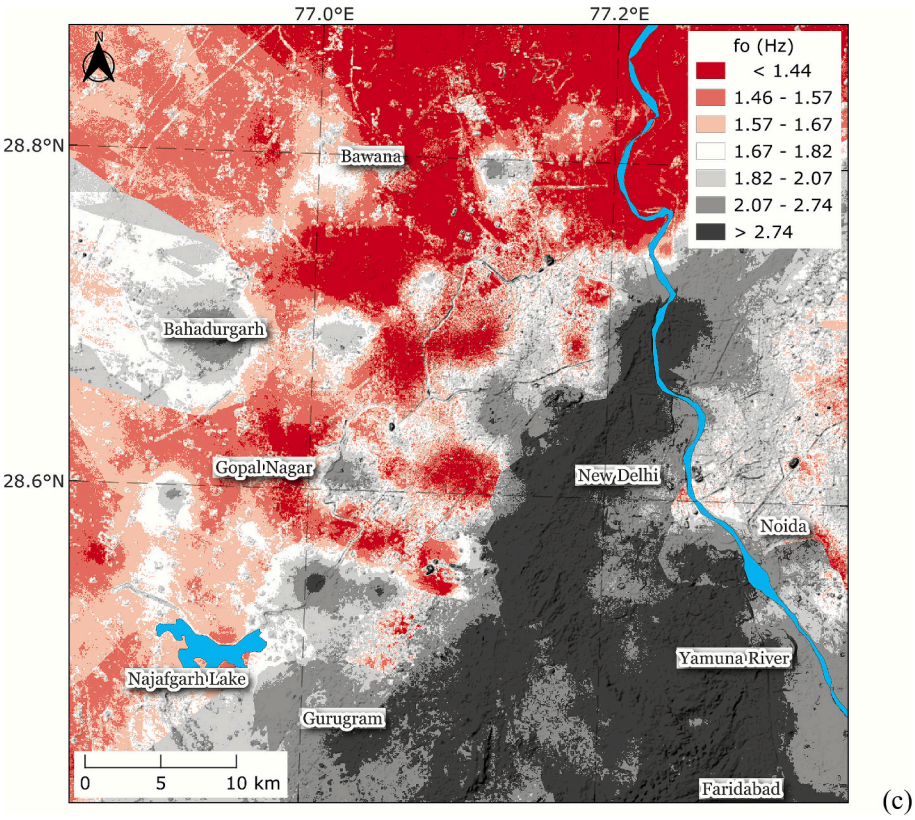


Fig. 15. (*continued*).

between DEM-based approaches for these distinct but related seismic parameters.

Tree-based ensemble methods consistently outperformed other algorithms in our study, aligning with recent trends in Vs30 prediction research where ensemble approaches increasingly show superior performance for complex terrain-geophysical relationships (Geyin and Maurer, 2023; Thakur and Anbazhagan, 2025). The exclusive selection of bagged and boosted ensembles across all cross-validation folds indicates that these methods effectively capture the nonlinear relationships between terrain morphometry and *fo*. While operational Vs30 systems rely primarily on topographic slope, our approach integrates comprehensive morphometric variables including elevation, curvature parameters, and terrain complexity indices.

This methodological difference reflects the distinct nature of *fo* against Vs30 as seismic site characterization parameters – Vs30 characterizes time-averaged subsurface velocity structure while *fo* quantifies resonant response characteristics. The prominence of MRVBF (Shapley value ~0.18) as the primary predictor reflects the physical relationship between sediment accumulation processes and seismic site response. Geological formation characteristics, particularly Precambrian rocks (Shapley value ~0.17), contribute substantially to predictions, indicating that both topographic and geological factors influence natural frequency patterns at regional scales. The comparable performance between our *fo* prediction approach and established Vs30 methods suggests that satellite-derived terrain analysis can effectively capture subsurface geological controls relevant to seismic site response. However, both parameters provide complementary rather than equivalent information for site characterization applications.

4.5.2. Resolution of scale-dependent effects in HVSR studies

The relationship between terrain morphometry and *fo* prediction shows sensitivity to spatial scale and data characteristics. Basin-scale studies have reported limited effectiveness of geomorphometric variables (Trevisani et al., 2017; Trevisani et al., 2021), while regional-scale analysis reveals substantial predictive capability. This difference appears related to spatial extent and data density rather than fundamental methodological limitations.

The variogram analysis reveals important spatial characteristics: a nugget-to-sill ratio of 48 % with an effective correlation range of 7305 m indicates substantial spatial structure in natural frequency patterns. The structured spatial component (52 % of total variance) remaining in ML residuals validates the regression kriging approach and suggests that spatial correlation modelling provides meaningful prediction improvements beyond machine learning alone. These findings suggest that DEM-derived predictors become effective when sufficient spatial coverage and geological diversity are present in the dataset. The scale-dependency has practical implications for remote sensing applications, indicating that regional-scale approaches may be more suitable for satellite-based site characterization than local-scale applications.

4.6. Limitations and future research scope

The present methodology addresses practical needs in regions where traditional geotechnical site characterization is logistically challenging. High-resolution predominant frequency (*fo*) mapping using globally available satellite data provides site response information for earthquake engineering applications, particularly in rapidly developing urban areas. However, several limitations affect broader applications. Model performance explains approximately 51 % of predominant frequency (*fo*) variance, indicating substantial unexplained variation remains. Model performance ($R^2 \approx 0.51$) is appropriate for regional-scale seismic hazard assessment and preliminary zonation but insufficient for site-specific engineering design or detailed microzonation where classification accuracy is critical. The 49 % unexplained variance reflects local subsurface heterogeneity not captured by DEM analysis which includes buried paleo-channels, weathered zones, and anthropogenic

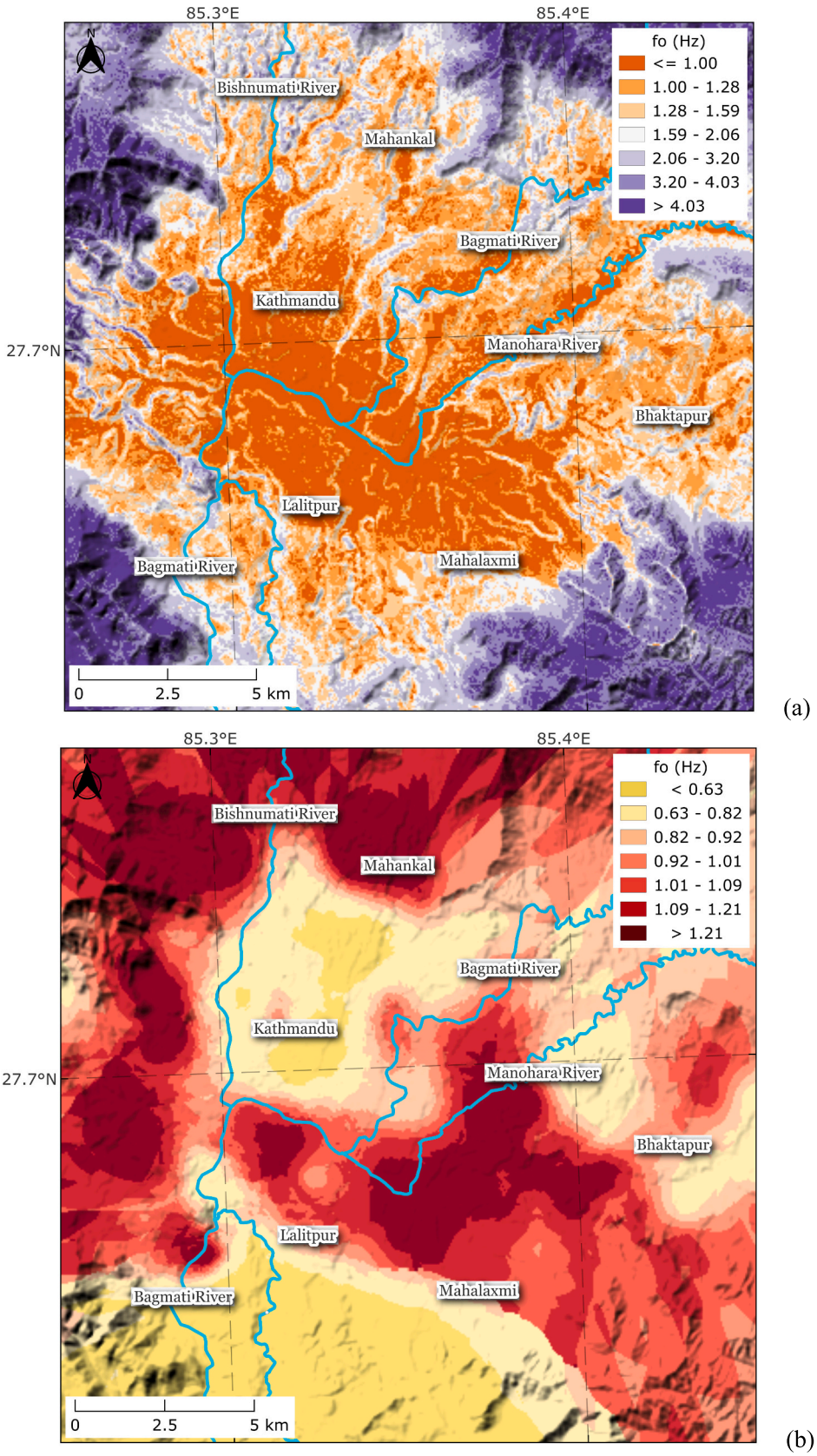


Fig. 16. Predominant Frequency (*fo* in Hz) prediction obtained for the Kathmandu region using (a) final Stacked Ensemble Model (FSEM), (b) residual kriging (RK), and (c) Regression Kriging (combining FSEM and RK). The map is prepared using a 50 m × 50 m grid and *fo* classes are coloured based on quantiles.

fills. Field validation through direct *fo* measurements remains essential for engineering design applications. The effective correlation range of

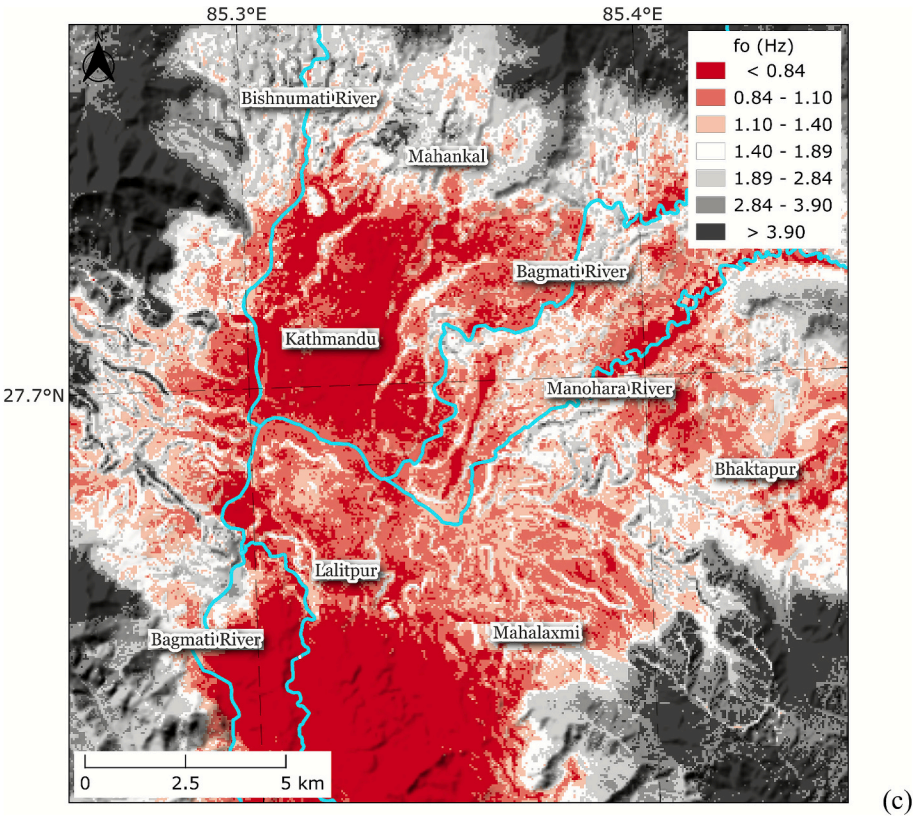


Fig. 16. (*continued*).

~7 km suggests that interpolation accuracy decreases significantly beyond this distance from training data, affecting prediction reliability in poorly sampled areas. Also, the predicted *fo* values represent characteristic site frequencies based on HVSR measurements, not event-specific resonance frequencies and actual site response during earthquakes may vary with source directionality, incident wavefield composition, and nonlinear soil behaviour. Our maps provide average site response proxies suitable for regional hazard assessment and preliminary planning, not event-specific predictions for critical infrastructure design.

Our proposed methodology integrates multi-resolution input data spanning three orders of magnitude in spatial scale. While the GLO-30 DEM provides 30 m resolution topographic information, bedrock depth data is available only at 250 m resolution, and geological formations are mapped at scales coarser than 1:500,000. Although the model successfully integrates these disparate scales through machine learning, the effective prediction resolution is fundamentally constrained by the coarsest input variable, which in this case is the bedrock depth dataset at approximately 250 m. The high-resolution 50 m output grids serve two practical purposes: enabling detailed visualization of topographic influences on *fo* predictions and facilitating integration with standard urban planning GIS datasets. However, these fine visualization grids should not be interpreted as implying corresponding precision in *fo* predictions at sub grid scales. The spatial uncertainty is further compounded by georeferencing procedures applied to source studies' published maps, which introduce positional uncertainties ranging from 50 to 200 m depending on map quality and ground control point availability. These combined factors limit the effective spatial precision of predictions, though regional-scale patterns remain robust and appropriate for preliminary seismic zonation and planning applications.

The present study's approach also requires validation in different geological and tectonic contexts before widespread application. Areas with complex geological structures, significant anthropogenic modifications, or disconnected relationships between surface topography and bedrock conditions may exhibit reduced prediction accuracy. Additionally, the reliance on existing measurement databases limits spatial coverage in some regions, and the effectiveness of terrain morphometric parameters depends on the specific relationships between surface morphology and subsurface conditions in different environments. Integration of additional satellite-derived variables could potentially improve prediction accuracy, though the moderate performance of comprehensive morphometric analysis suggests that terrain-based approaches may have inherent limitations for *fo* prediction. Multi-sensor data fusion incorporating SAR backscatter, and optical vegetation indices represents potential enhancement pathways. The framework's modular design enables adaptation to different remote sensing platforms and integration with existing monitoring systems.

The compiled dataset exhibits spatial heterogeneity, with measurement concentration in urban centers (Delhi, Kathmandu, Dhaka) reflecting the distribution of government microzonation projects and research studies in populated areas. This urban bias may limit model extrapolation accuracy to under sampled mountainous terrain, high-elevation plateaus, intermontane basins, and steep slopes where infrastructure development occurs but measurements remain sparse. While the dataset captures diverse geological contexts (Quaternary sediments, Precambrian basement, multiple depositional environments) essential for regional pattern recognition, the uneven spatial distribution represents an inherent limitation of retrospectively compiled measurements. Also, regional-scale patterns captured by terrain morphometry and geological classifications provide first-order approximations suitable for preliminary hazard assessment, but site-specific validation remains essential for infrastructure applications in under sampled terrain types. Future studies should prioritize systematic measurements across elevation gradients, diverse geological contexts, and currently sparse regions to improve model representativeness across the full range of Himalayan

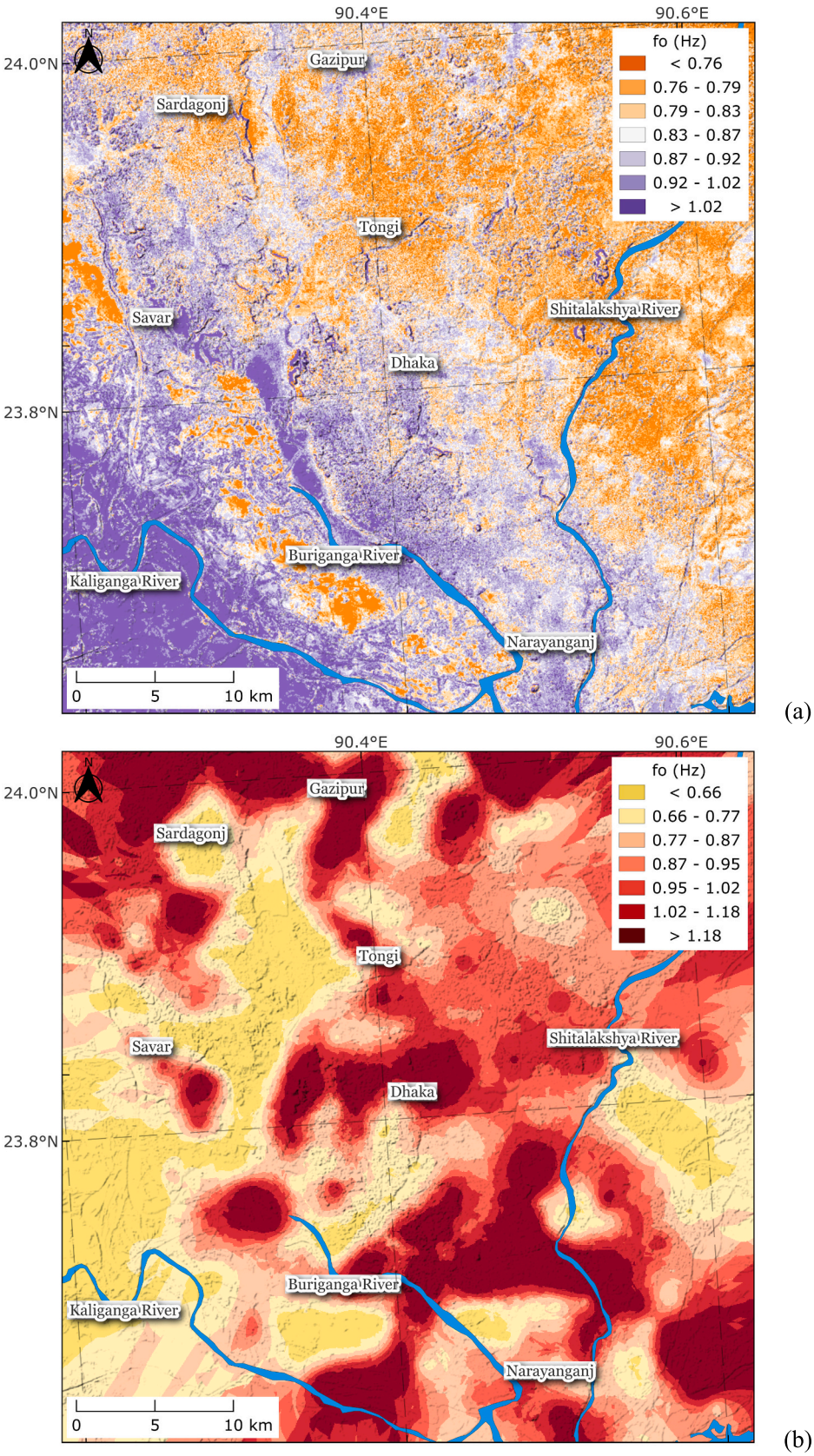


Fig. 17. Predominant Frequency (*fo* in Hz) prediction obtained for the Dhaka region using (a) final Stacked Ensemble Model (FSEM), (b) residual kriging (RK), and (c) Regression Kriging (combining FSEM and RK). The map is prepared using a 50 m × 50 m grid and *fo* classes are coloured based on quantiles.

terrain types and assess model transferability beyond accessible valley sites.

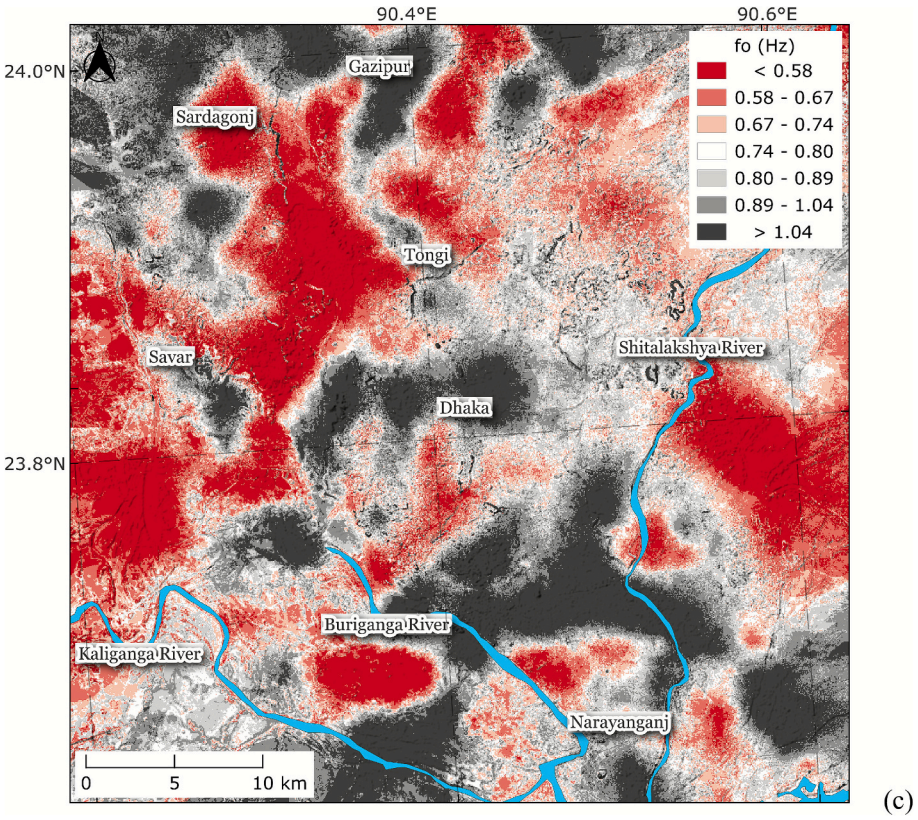


Fig. 17. (*continued*).

Future research should also focus on developing region-specific calibration approaches that account for local geological conditions and their influence on terrain – frequency (*fo*) relationships. Physics-based feature engineering approaches combining multiple terrain parameters into composite indices specifically designed for *fo* prediction could enhance accuracy by capturing complex interactions between terrain characteristics and their combined influence on site response mechanisms. Integration of additional remote sensing data sources, including synthetic aperture radar and hyperspectral imagery, could provide enhanced characterization of subsurface conditions and improve prediction accuracy in complex geological environments.

5. Summary and Conclusion

This study establishes that satellite-derived terrain morphometry, combined with geological classifications and bedrock depth data, enables robust regional-scale predominant frequency (*fo*) prediction when integrated through machine learning frameworks. This represents a methodological advancement over previous approaches that relied exclusively on spatial interpolation techniques such as kriging, inverse distance weighting, and natural neighbor methods which simply redistribute measured values without leveraging terrain characteristics, geological context, or subsurface parameters. While basin-scale studies found limited utility in geomorphometric variables due to sparse data (Trevisani et al., 2021), our regional analysis with 4400 measurements demonstrates that physical predictors substantially enhance *fo* mapping reliability beyond what interpolation alone can achieve.

Our analysis reveals a clear predictor hierarchy governing seismic site response. Out of twenty predictors, valley bottom identification (MRVBF) emerges as the strongest predictor (Shapley importance ≈ 0.18), reflecting the critical role of sediment accumulation zones in controlling impedance contrasts. Geological formation characteristics, particularly Precambrian basement rocks (≈ 0.17), rank second, followed by elevation (≈ 0.15) and bedrock depth. This hierarchy directly reflects the physical controls of depositional environments and basement geology on natural frequency patterns. The two-stage regression kriging framework achieves $R^2 = 0.516$ and RMSE = 0.634 log units, with variogram analysis revealing that 52 % of residual variance exhibits spatial correlation extending 7.3 km. This structured spatial component validates combining machine learning predictions with geostatistical interpolation, demonstrating that each stage captures complementary aspects of *fo* variation.

The methodology provides practical benefits for earthquake-prone developing regions where conventional geotechnical site characterization remains logistically challenging and cost-prohibitive. Generated *fo* maps for Delhi, Kathmandu, and Dhaka successfully delineate contrasting site response zones from low-frequency deep sedimentary basins (<1.0 Hz) to high-frequency bedrock-controlled areas (>3.0 Hz) using globally available satellite data. Unlike interpolation methods requiring dense measurement networks, the predictive approach extends reliable mapping to undersampled areas through learned terrain-frequency relationships, offering significant operational advantages for regional seismic zonation and preliminary hazard assessment.

However, critical limitations define appropriate application boundaries. The moderate model performance ($R^2 \approx 0.51$) suits regional planning purposes but proves insufficient for site-specific engineering design. The 49 % unexplained variance reflects local subsurface complexities such as buried paleo-channels, weathered zones, anthropogenic fills etc. that surface-derived parameters cannot capture. Consequently, field validation through direct *fo* measurements remains essential for infrastructure applications. Additionally, effective prediction resolution is constrained by the coarsest input dataset (250 m bedrock depth), not the 50 m visualization grid. Georeferencing uncertainties, multi-resolution data integration across three orders of magnitude, and urban measurement concentration further limit extrapolation confidence to undersampled regions. The predicted *fo* values represent

characteristic site frequencies from HVSR measurements rather than event-specific earthquake resonance, which varies with source mechanisms and incident wavefield properties.

Future advances for the *fo* prediction improvement will require systematic field campaigns targeting elevation gradients, diverse geological contexts, and currently data sparse regions to improve model transferability beyond the present scope. Methodological enhancements through physics-based composite indices combining terrain parameters, integration of gravity anomalies constraining basin geometry, higher-resolution bedrock depth datasets, and region-specific calibration accounting for local geological conditions may offer pathways toward improved accuracy. The demonstrated capability of DEM-based predictive modelling to generate regional *fo* maps from globally available data represents meaningful progress for seismic site characterization in data-constrained regions, provided practitioners recognize the methodology's appropriate scope for planning applications and the continued necessity of field validation for engineering design.

CRediT authorship contribution statement

Harish Thakur: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **P. Anbazhagan:** Writing – review & editing, Supervision, Conceptualization.

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 thank Ministry of Jal Shakti (MoJS), Government of India for the project entitled “Integrated Investigation for Risk Assessment of the Dam” under grant R-24011/57/2023-Pen Riv Section-MOWR; and M/s Secon Pvt. Ltd. for funding the project “Effect of shear wave velocity calibration on Amplification of shallow and deep soil sites” grant No: SECON/IISc/MoES/WO/07-18/0079.

Data availability

Data will be made available on a reasonable request.

References

Abbasnejadfard, M., Bastami, M., Jafari, M.K., Azadi, A., 2023. Spatial correlation models of VS30 values: A case study of the Tehran region. Eng. Geol. 325, 107300. https://doi.org/10.1016/j.enggeo.2023.107300.

Akin, M.K., Kramer, S.L., Topal, T., 2011. Empirical correlations of shear wave velocity (Vs) and penetration resistance (SPT-N) for different soils in an earthquake-prone area (Erbaa-Turkey). Eng. Geol. 119 (1–2), 1–17. https://doi.org/10.1016/j.enggeo.2011.01.007.

Allen, T.I., Wald, D.J., 2009. On the use of high-resolution topographic data as a proxy for seismic site conditions (VS 30). Bull. Seismol. Soc. Am. 99 (2A), 935–943. https://doi.org/10.1785/0120080255.

Amatulli, G., Domisch, S., Tuanmu, M.N., Parmentier, B., Ranipeta, A., Malczyk, J., Jetz, W., 2018. A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. Sci. Data 5 (1), 1–15. https://doi.org/10.1038/sdata.2018.40.

Ansary, M.A., Arefin, M.R., 2020. Assessment of predominant frequencies in Dhaka city, Bangladesh using ambient vibration. Asian J. Civ. Eng. 21 (1), 91–104. https://doi.org/10.1007/s42107-019-00194-2.

Ansary, M.A., Helaly, A.L., Khair, A., Saha, S., Helaly, S.M., 2022. Assessment of soil-building resonant effect for DMDP area in Bangladesh. Natl. Hazards Rev. 23 (4), 04022024. https://doi.org/10.1061/(ASCE)NH.1527-6996.0000576.

Brando, G., Pagliaroli, A., Cocco, G., Di Buccio, F., 2020. Site effects and damage scenarios: the case study of two historic centers following the 2016 Central Italy earthquake. Eng. Geol. 272, 105647. https://doi.org/10.1016/j.enggeo.2020.105647.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24 (2), 123–140. https://doi.org/10.1007/BF00058655.

BSSC, 2015. NEHRP Recommended Seismic Provisions for New Buildings and Other Structures. Washington, D.C.. Accessed using : https://www.fema.gov/sites/default/files/2020-07/fema_nehrp-seismic-provisions-new-buildings_p-1050-1_2015.pdf.

Chamlagain, D., Acharya, S., Neupane, P., Dhakal, N.R., 2025. Study of ambient noise for seismic-microzonation of the Eastern part of Hetauda Dun Valley, Central Nepal. Nat. Hazards 121 (4), 3805–3828. https://doi.org/10.1007/s11069-024-06938-x.

Chen, T.H.K., Kincey, M.E., Rosser, N.J., Seto, K.C., 2024. Identifying recurrent and persistent landslides using satellite imagery and deep learning: A 30-year analysis of the Himalaya. Sci. Total Environ. 922, 171161. https://doi.org/10.1016/j.scitotenv.2024.171161.

Chowdhuri, S.N., Singh, O.P., Majumdar, R.K., 2011. Site response studies in Agartala Urban agglomeration. Nat. Hazards 59 (1), 329–345. https://doi.org/10.1007/s11069-011-9759-y.

Delgado, J., Lopez Casado, C., Giner, J., Estevez, A., Cuenca, A., Molina, S., 2000. Microtremors as a geophysical exploration tool: applications and Limitations. Pure Appl. Geophys. 157 (9), 1445–1462. https://doi.org/10.1007/PL00001128.

Di Martino, A., Sgattoni, G., Purri, F., Amorosi, A., 2024. Seismic amplification of late Quaternary paleovalley systems: 2D seismic response analysis of the Pescara paleovalley (Central Italy). Eng. Geol. 341, 107697. https://doi.org/10.1016/j.enggeo.2024.107697.

Dong, L., Shan, J., 2013. A comprehensive review of earthquake-induced building damage detection with remote sensing techniques. ISPRS J. Photogramm. Remote Sens. 84, 85–99. https://doi.org/10.1016/j.isprsjprs.2013.06.011.

DST, 2007. Seismic Microzonation Atlas of Guwahati Region. Govt. of India Publication. Accessed using : https://asdma.assam.gov.in/sites/default/files/Seismic_Microzonation.pdf.

European Space Agency, 2024. Copernicus Global Digital Elevation Model. Distributed by OpenTopography. https://doi.org/10.5069/G9028PQB.

Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of online learning and an application to boosting. J. Comput. Syst. Sci. 55 (1), 119–139. https://doi.org/10.1006/jcss.1997.1504.

Fukushima, Y., Bonilla, L.F., Scotti, O., Douglas, J., 2007. Site classification using horizontal-to-vertical response spectral ratios and its impact when deriving empirical ground-motion prediction equations. J. Earthq. Eng. 11 (5), 712–724. https://doi.org/10.1080/13632460701457116.

Gallant, J.C., Dowling, T.I., 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resour. Res. 39 (12). https://doi.org/10.1029/2002WR001426.

Geyin, M., Maurer, B.W., 2023. U.S. National Models and Maps Informed by Remote Sensing and Machine Learning. Seismol. Res. Lett. 94 (3), 1467–1477. https://doi.org/10.1785/0220220181.

Gupta, I.D., Kumari, N., 2023. Characterization of strong-motion recording sites in the Himalayas and adjoining areas. Earthquake Eng. Res. 2 (2), 226–258. https://doi.org/10.1002/eer2.44.

Gupta, S.V., Parvez, I.A., Khan, P.K., 2023. Site response analysis beneath the Kashmir basin (NW Himalaya) using ambient noise. Earthquake Spectra 39 (4), 2436–2469. https://doi.org/10.1177/87552930231195640.

Haghshenas, E., Bard, P.-Y., Theodulidis, N., Sesame WP04 Team, 2008. Empirical evaluation of microtremor H/V spectral ratio. Bull. Earthq. Eng. 6 (1), 75–108. https://doi.org/10.1007/s10518-007-9058-x.

Hengl, T., MacMillan, R.A., 2019. Predictive Soil Mapping with R. OpenGeoHub foundation, Wageningen. http://www.soilmapper.org/. ISBN: 978–0–359-30635-0.

Hassani, B., Atkinson, G.M., 2018. Site-effects model for central and eastern North America based on peak frequency and average shear-wave velocity. Bulletin of the Seismological Society of America 108 (1), 338–350. https://doi.org/10.1785/0120170061.

Hengl, T., Heuvelink, G.B.M., Rossiter, D.G., 2007. About regression-kriging: from equations to case studies. Comput. Geosci. 33 (10), 1301–1315. https://doi.org/10.1016/j.cageo.2007.05.001.

Ho, T.K., 1995. Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition (Vol. 1, pp. 278-282). IEEE. https://doi.org/10.1109/ICDAR.1995.598994.

Jasiewicz, J., Stepinski, T.F., 2013. Geomorphons—a pattern recognition approach to classification and mapping of landforms. Geomorphology 182, 147–156. https://doi.org/10.1016/j.geomorph.2012.11.005.

Jenness, J., 2006. Topographic position index (TPI) v. 1.2. Jenness Enterprises.

JRA, 2019. Specification for Highway Bridges, Part V: Seismic Design. Japan Road Association, Tokyo, Japan. Accessed at: https://iisee.kenken.go.jp/worldlist/31_Japan/Japan_2_HighwayBridge_2017%20Part_V_Seismic_Design.pdf.

Kim, H.S., Sun, C.G., Lee, M.G., Cho, H.I., 2021. Multivariate geotechnical zonation of seismic site effects with clustering-blended model for a city area, South Korea. Eng. Geol. 294, 106365. https://doi.org/10.1016/j.enggeo.2021.106365.

Kramer, S.L., Stewart, J.P., 2024. Geotechnical Earthquake Engineering. CRC Press. ISBN 9781032842745.

Kuldeep, Shekar B., Mohan, G., Singh, S.K., 2022. Assessment of seismic vulnerability using the ambient noise recordings in Cachar Fold Belt, Assam. J. Geol. Soc. India 98 (6), 795–804. https://doi.org/10.1007/s12594-022-2070-6.

Kumar, P., Kamal, Kumar A., 2025. Site classification of locations of installed sensors in the Kumaon Region of the Himalayas using empirical approaches. Arab. J. Geosci. 18 (1), 27. https://doi.org/10.1007/s12517-024-12154-5.

Kumar, P., Mahajan, A.K., Sharma, M., 2023. Site effect assessment and vulnerability analysis using multi-geophysical methods for Kangra city, NW Himalaya, India. J. Earth Syst. Sci. 132 (1), 14. https://doi.org/10.1007/s12040-022-02032-7.

Kundu, P., Pain, A., Das, J., Kumar, A., 2024. Study of Seismic Site Effects and Development of Soil-Building Resonance Map of Noida City, the Industrial Hub of India. Indian Geotech. J. 1-18. https://doi.org/10.1007/s40098-024-01134-x.

Laouami, N., 2020. Proposal for a new site classification tool using microtremor data. Bull. Earthq. Eng. 18 (10), 4681–4704. https://doi.org/10.1007/s10518-020-00882-4.
Lehner, B., Grill, G., 2013. Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. Hydrol. Process. 27 (15), 2171–2186. https://doi.org/10.1002/hyp.9740.
Lin, A., Wotherspoon, L., Bradley, B., Motha, J., 2021. Evaluation and modification of geospatial liquefaction models using land damage observational data from the 2010–2011 Canterbury Earthquake Sequence. Eng. Geol. 287, 106099. https://doi.org/10.1016/j.enggeo.2021.106099.
Liu, W., Chen, Q., Wang, C., Juang, C.H., Chen, G., 2017. Spatially correlated multiscale Vs30 mapping and a case study of the Suzhou site. Eng. Geol. 220, 110–122. https://doi.org/10.1016/j.enggeo.2017.01.026.
Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Proces. Syst. 30.
Mahajan, A.K., Kumar, P., Kumar, P., 2021. Near-surface seismic site characterization using Nakamura-based HVSR technique in the geological complex region of Kangra Valley, northwest Himalaya, India. Arab. J. Geosci. 14 (10), 826. https://doi.org/10.1007/s12517-021-07136-w.
Maklad, M., Yokoi, T., Hayashida, T., ElGabry, M.N., Hassan, H.M., Hussein, H.M., Fattah, T.A., Rashed, M., 2020. Site characterization in Ismailia, Egypt using seismic ambient vibration array. Eng. Geol. 279, 105874. https://doi.org/10.1016/j.enggeo.2020.105874.
Martínez-Segura, M.A., García-Nieto, M.C., Navarro, M., Vásconez-Maza, M.D., Oda, Y., García-Jerez, A., Enomoto, T., 2024. Seismic characterisation of the subsoil under a historic building: Cathedral Church of Saint Mary in Murcia case study. Eng. Geol. 335, 107529. https://doi.org/10.1016/j.enggeo.2024.107529.
MATLAB, 2023. Version R2023b. The MathWorks Inc., Natick, Massachusetts. URL: https://www.mathworks.com.
Minár, J., Evans, I.S., Jenčo, M., 2020. A comprehensive system of definitions of land surface (topographic) curvatures, with implications for their application in geoscience modelling and prediction. Earth Sci. Rev. 211, 103414. https://doi.org/10.1016/j.earscirev.2020.103414.
Mishra, O.P., Singh, P., Ram, B., Gera, S.K., Singh, O.P., Mukherjee, K.K., Chakrabortty, G.K., Chandrasekhar, S.V.N., Selinraj, A., Som, S.K., 2020. Seismic site specific study for seismic microzonation: A way forward for risk resiliency of vital infrastructure in Sikkim, India. Int. J. Geosci. 11 (3), 125–144. https://doi.org/10.4236/ijg.2020.113008.
Mori, F., Mendicelli, A., Moscatelli, M., Romagnoli, G., Peronace, E., Naso, G., 2020. A new Vs30 map for Italy based on the seismic microzonation dataset. Eng. Geol. 275, 105745. https://doi.org/10.1016/j.enggeo.2020.105745.
Mundepi, A.K., Galiana-Merino, J.J., Lindholm, C., 2010. Soil characteristics and site effect assessment in the city of Delhi (India) using H/V and f-k methods. Soil Dyn. Earthq. Eng. 30 (7), 591–599. https://doi.org/10.1016/j.soildyn.2010.01.016.
Mundepi, A.K., Galiana-Merino, J.J., Asthana, A.K.L., Rosa-Cintas, S., 2015. Soil characteristics in Doon Valley (north west Himalaya, India) by inversion of H/V spectral ratios from ambient noise measurements. Soil Dyn. Earthq. Eng. 77, 309–320. https://doi.org/10.1016/j.soildyn.2015.06.006.
Nakamura, Y., 1989. A method for dynamic characteristics estimation of subsurface using microtremor on the ground surface. Quarterly Rep. Railway Technical Res. 30 (1), 25–33.
NASA JPL, 2013. NASA shuttle Radar Topography Mission Global 1 arc second [Data set]. NASA Land Processes Distributed Active Archive Center. https://doi.org/10.5067/MEASURES/SRTM/SRTMGL1.003.
Nath, S.K., Adhikari, M.D., Devaraj, N., Maiti, S.K., 2015. Seismic vulnerability and risk assessment of Kolkata City, India. Nat. Hazards Earth Syst. Sci. 15 (6), 1103–1121. https://doi.org/10.5194/nhess-15-1103-2015.
NCS, 2016. A report on Seismic Hazard Microzonation of NCT Delhi on 1:10,000 scale. Govt. of India Publication. Accessed using: https://seismo.gov.in/sites/default/files/publication/Delhi_microzonation_report-2015.pdf.
Panzera, F., Lombardo, G., Imposa, S., Grassi, S., Gresta, S., Catalano, S., Di Maio, E., 2018. Correlation between earthquake damage and seismic site effects: the study case of Lentini and Carlentini, Italy. Eng. Geol. 240, 149–162. https://doi.org/10.1016/j.enggeo.2018.04.014.
Paudyal, Y.R., Yatabe, R., Bhandary, N.P., Dahal, R.K., 2013. Basement topography of the Kathmandu Basin using microtremor observation. J. Asian Earth Sci. 62, 627–637. https://doi.org/10.1016/j.jseaes.2012.11.011.
Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Comput. Geosci. 30 (7), 683–691. https://doi.org/10.1016/j.cageo.2004.03.012.
QGIS Development Team, 2023. QGIS Geographic Information System. Open Source Geospatial Foundation. URL. http://qgis.org.
R Core Team, 2024. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Rahman, M.Z., Hossain, M.S., Kamal, A.M., Siddiqua, S., Mustahid, F., Farazi, A.H., 2018. Seismic site characterization for Moulvibazar town, Bangladesh. Bull. Eng. Geol. Environ. 77 (4), 1451–1471. https://doi.org/10.1007/s10064-017-1031-6.
Riley, S.J., DeGloria, S.D., Elliot, R., 1999. Index that quantifies topographic heterogeneity. Intermountain J. Sci. 5 (1–4), 23–27.
Sandhu, M., Yadav, R.B.S., Kumar, D., Abhishek, 2022. Estimation of Site Amplification factor and Predominant Frequency in and around Panchkula City, Haryana, India. In: Symposium in Earthquake Engineering. Springer, pp. 599–611. https://doi.org/10.1007/978-981-99-1459-3_48.
Sappington, J.M., Longshore, K.M., Thompson, D.B., 2007. Quantifying landscape ruggedness for animal habitat analysis: a case study using bighorn sheep in the Mojave Desert. J. Wildl. Manag. 71 (5), 1419–1426. https://doi.org/10.2193/2005-723.
Senkaya, M., Silahtar, A., Erkan, E.F., Karaaslan, H., 2024. Prediction of local site influence on seismic vulnerability using machine learning: A study of the 6 February 2023 Türkiye earthquakes. Eng. Geol. 337, 107605. https://doi.org/10.1016/j.enggeo.2024.107605.
SESAME Project, 2004. Guidelines for the Implementation of the H/V Spectral Ratio Technique on Ambient Vibrations Measurements, Processing and Interpretation. http://sesame-fp5.obs.ujf-grenoble.fr/Papers/HV_User_Guidelines.pdf.
Shangguan, W., Hengl, T., Mendes de Jesus, J., Yuan, H., Dai, Y., 2017. Mapping the global depth to bedrock for land surface modeling. J. Adv. Model. Earth Syst. 9 (1), 65–88. https://doi.org/10.1002/2016MS000686.
Shankar, U., Yadav, P.K., Singh, A.P., Gupta, A.K., 2021a. Evaluation of site-specific characteristics using microtremor measurements in the Gorakhpur city of Uttar Pradesh, India. J. Earth Syst. Sci. 130 (4), 188. https://doi.org/10.1007/s12040-021-01690-3.
Shankar, U., Kumari, S., Yadav, P.K., Singh, A.P., Gupta, A.K., 2021b. Microtremor measurements in the India's holy city, Varanasi for assessment of site characteristics. Quat. Int. 585, 143–151. https://doi.org/10.1016/j.quaint.2021.01.008.
Singh, A.P., Kumar, M.R., Pandey, A., Roy, K.S., 2019. Investigation of spatial and temporal variability of site response in the Arunachal Himalaya using ambient seismic noise and earthquake waveforms. Near Surface Geophys. 17 (4), 427–445. https://doi.org/10.1002/nsg.12053.
Sörensen, R., Zinko, U., Seibert, J., 2006. On the calculation of the topographic wetness index: evaluation of different methods based on field observations. Hydrol. Earth Syst. Sci. 10 (1), 101–112. https://doi.org/10.5194/hess-10-101-2006.
Thakur, H., Anbazhagan, P., 2025. Geology, geomorphology and Vs30 based site classification of the Himalayan region using a stacked model. Eng. Geol., 108229 https://doi.org/10.1016/j.enggeo.2025.108229.
Trevisani, S., Boaga, J., Agostini, L., Galgaro, A., 2017. Insights into bedrock surface morphology using low-cost passive seismic surveys and integrated geostatistical analysis. Sci. Total Environ. 578, 186–202. https://doi.org/10.1016/j.scitotenv.2016.11.041.
Trevisani, S., Pettenati, F., Paudyal, S., Sandron, D., 2021. Mapping long-period soil resonances in the Kathmandu basin using microtremors. Environ. Earth Sci. 80 (7), 265. https://doi.org/10.1007/s12665-021-09532-7.
Vilalta, R., Drissi, Y., 2002. A perspective view and survey of meta-learning. Artif. Intell. Rev. 18, 77–95. https://doi.org/10.1023/A:1019956318069.
Walling, M.Y., Mohanty, W.K., Nath, S.K., Mitra, S., John, A., 2009. Microtremor survey in Talchir, India to ascertain its basin characteristics in terms of predominant frequency by Nakamura's ratio technique. Eng. Geol. 106 (3–4), 123–132. https://doi.org/10.1016/j.enggeo.2009.03.013.
Wandrey, C.J., 1998. Geologic Map of South Asia (geo8ag) [Data Set]. U.S. Geological Survey. https://doi.org/10.5066/P9YC1C8G.
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. https://doi.org/10.1016/j.isprsjprs.2018.02.017.
Yilar, E., Baise, L.G., Ebel, J.E., 2017. Using H/V measurements to determine depth to bedrock and Vs30 in Boston, Massachusetts. Eng. Geol. 217, 12–22. https://doi.org/10.1016/j.enggeo.2016.12.002.
Zahoor, F., Rao, K.S., Mir, B.A., Satyam, N., 2023. Geophysical surveys in the Kashmir valley (J&K Himalayas) part I: estimation of dynamic parameters for soils and investigation of the deep basin structure. Soil Dyn. Earthq. Eng. 174, 108155. https://doi.org/10.1016/j.soildyn.2023.108155.
Zhao, J.X., Irikura, K., Zhang, J., Fukushima, Y., Somerville, P.G., Asano, A., Ohno, Y., Oouchi, T., Takahashi, T., Ogawa, H., 2006. An empirical site-classification method for strong-motion stations in Japan using H/V response spectral ratio. Bull. Seismol. Soc. Am. 96 (3), 914–925. https://doi.org/10.1785/0120050124.
Yao Xiaobai, A., 2020. Georeferencing and Geocoding. In: Kobayashi, Audrey (Ed.), International Encyclopedia of Human Geography. https://doi.org/10.1016/B978-0-08-102295-5.10548-7.