Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Machine Learning Based National Mapping of Archaeological Occurrence and Landscape Pressure in Tajikistan

Chenglin Lin
Shanghai Jiao Tong University
Marco Nebbia
University College London
Federica Cilio
University College London
Patrick Cuthbertson
University College London
Rui Pang
University College London
Gaygysyz Jorayev
Macao University of Tourism
Tim D. Williams
University College London
Hang Xu
Shanghai Jiao Tong University
Qiang Sun
qiang.sun@rmit.edu.au

RMIT University

Article

Keywords:

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11654966/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Machine learning based national mapping of archaeological occurrence and landscape pressure in Tajikistan

Chenglin Li[1], Marco Nebbia[2], Federica Cilio[2], Patrick Cuthbertson[2], Rui Pang[2], Gaygysyz Jorayev[2,3], Tim D. Williams[2], Hang Xu[1,†], and Qiang Sun[4,*]

[1]State Key Lab of Ocean Engineering, School of Ocean and Civil Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

[2]Institute of Archaeology, University College London, 31-34 Gordon Square, London WC1H 0PY, United Kingdom

[3]Macao University of Tourism, Colina de Mong-Há, Macau, China

[4]School of Science, RMIT University, Melbourne, VIC 3001, Australia

[†]hangxu@sjtu.edu.cn

[*]qiang.sun@rmit.edu.au

Abstract

National maps combining archaeological occurrence and landscape pressure can inform heritage survey; their outputs depend on methodological choices. We examined four such choices for Tajikistan, combining a Maximum Entropy (MaxEnt) occurrence model from 15,728 records with cultivation persistence, built-surface growth and soil loss. The national model discriminated sites from random background with a spatially blocked area under the receiver operating characteristic curve (AUC) of 0.85; the default MaxEnt value of 0.68 largely reflected its evaluation background. Regional models discriminated better in all four regions, and the national surface assigned many Pamir sites to low classes. Integration covered two-thirds of the country, excluding 42% of Pamir records. High occurrence and high pressure coincided over 75–248 km^2 across eight configurations, and a stable 75 km^2 core with few recorded sites is a field-survey priority. These results favour regionally calibrated heritage-risk assessment, and the framework is adaptable to other countries and regions.

Introduction

Many archaeologically rich landscapes of Central Asia are undergoing rapid transformation. For millennia, environmental gradients linking irrigated valleys, foothills and high plateaus shaped patterns of farming, herding and movement along the Silk Roads. The same valleys now concentrate agricultural intensification, irrigation, urban growth and new infrastructure.[1–3] Decisions on where to survey and where to intervene first therefore require evidence at national scale, in addition to assessments of individual monuments.

The Central Asian Archaeological Landscapes (CAAL) project has developed a sub-continental spatial database of heritage resources across Central Asia. The database draws primarily on national inventories, largely derived from the Soviet-era 'Monument Passports'.[4,5] A major contribution of the project is the systematic mapping of archaeological features from satellite imagery across the region, which records sites and monuments as well as more ephemeral, partly preserved traces of past activity. In addition to compiling and managing these data in a Geographic Information System (GIS), CAAL aims to develop quantitative methods for assessing the condition of, and risks to, the cultural heritage of the region.[3] As shown in Fig. 1, Tajikistan was the first country for which this systematic mapping was completed, and it therefore serves as the pilot for national-scale assessment. The country is also a suitable test case: it contains some of the most important archaeological sites in Central Asia, lies at the centre of the historical Silk Roads and spans a wide range of environments, from irrigated lowlands at about 300 m above sea level to the Pamirs above 7,000 m.

Archaeological documentation across Tajikistan is, however, spatially uneven.[6] Historically compiled inventories are concentrated in accessible and agriculturally productive areas, whereas high-elevation and remote areas are under-represented.[7,8] Concurrently, agricultural intensification,[1] irrigation expansion,[9] peri-urban sprawl[2] and infrastructure development[10] are reshaping the valley systems and settlement corridors where most sites are recorded, posing a threat to both documented and undocumented deposits. Other countries in the region share this combination of uneven documentation and rapid change, which suggests that a method developed for Tajikistan is potentially transferable.

Archaeological predictive modelling (APM) provides one approach to uneven documentation.[11] By relating known site locations to environmental variables, predictive models generate continuous surfaces of relative archaeological occurrence likelihood.[12,13] Presence-only methods such as Maximum Entropy (MaxEnt) are suited to settings in which only presence records are reliable.[13–15] Two properties of these models are directly relevant to heritage planning. First, model evaluation depends on the definition of background locations, so the area under the receiver operating characteristic curve (AUC) is specific to the study extent and setting.[16,17] Second, recorded sites are spatially clustered, and spatially blocked validation therefore provides a more realistic estimate than random cross-validation of predictive performance in new areas.[18,19] Both properties bear on a practical question for heritage agencies, namely whether a single national model is adequate or whether regionally calibrated models are required.

Open remote-sensing products now enable consistent quantification of land-use change at national scale.[20–22] Multi-year vegetation-index time series can separate persistent cultivation from short-lived greening,[23] global settlement layers record the growth of built surfaces,[24,25] and global soil-erosion models provide a consistent proxy for geomorphic instability.[26] Previous work in southern Tajikistan combined such information with site records to map heritage risk at regional scale,[3] and studies elsewhere have linked land-use change to the preservation of individual sites.[27] Extending this approach to national scale, by coupling occurrence modelling with explicit pressure indicators, is the logical next step for Tajikistan. Such a synthesis involves a sequence of methodological choices, each of which can alter the spatial distribution of priority:

the evaluation of the occurrence model, the use of one national model or several regional models, the proportion of the country covered by the pressure data, and the conversion of continuous outputs into priority classes. These choices warrant joint examination.

In this study, we develop such a framework for Tajikistan, combining archaeological occurrence modelling with land-use and erosion pressure indicators to support survey prioritisation, and we examine each of these choices in turn. Four questions are addressed. (1) What proportion of the country, and of its recorded archaeology, can be assessed with open pressure data? (2) To what extent do the evaluation background and validation design affect the apparent performance of the occurrence model? (3) Does a single national model serve all regions, or do regionally calibrated models discriminate better? (4) Which areas of coincident high occurrence and high pressure remain stable when the occurrence model and its classification change? The overall workflow is shown in Fig. 2. The objective is landscape-scale stratification for survey and preventive planning.

Results

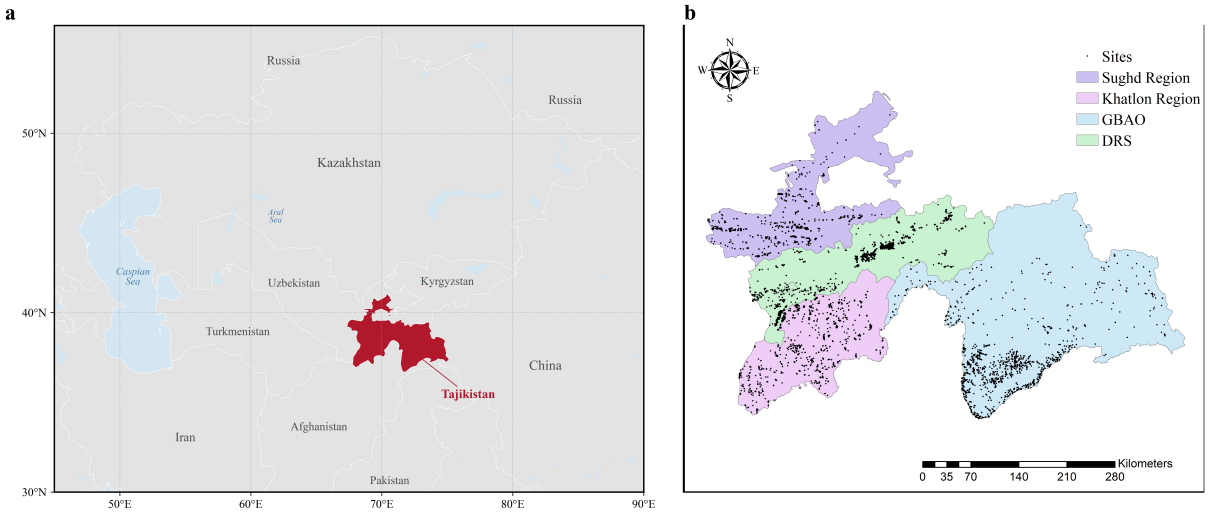


Fig. 1: Study area and archaeological records. (a) Location of Tajikistan in Central Asia; country boundaries from Natural Earth (1:50m). (b) The four regional units used for regional modelling: Sughd Region, Khatlon Region, Gorno-Badakhshan Autonomous Region (GBAO) and the Districts of Republican Subordination (DRS). Black dots: recorded archaeological sites.

Recorded sites and analytical coverage

Approximately two-thirds of Tajikistan could be assessed within the integrated framework, and the remaining third contains much of the recorded archaeology of the Pamirs. As shown in Fig. 1b, the dataset comprises 15,728 archaeological records. After retention of one record per 100 m predictor cell with complete predictor values, 10,918 presence cells remained for the national model, and 5,246, 1,460, 1,942 and 2,269 for the regional models of the Districts of Republican Subordination (DRS), Gorno-Badakhshan Autonomous Region (GBAO), Khatlon and Sughd, respectively, as presented in Table 1.

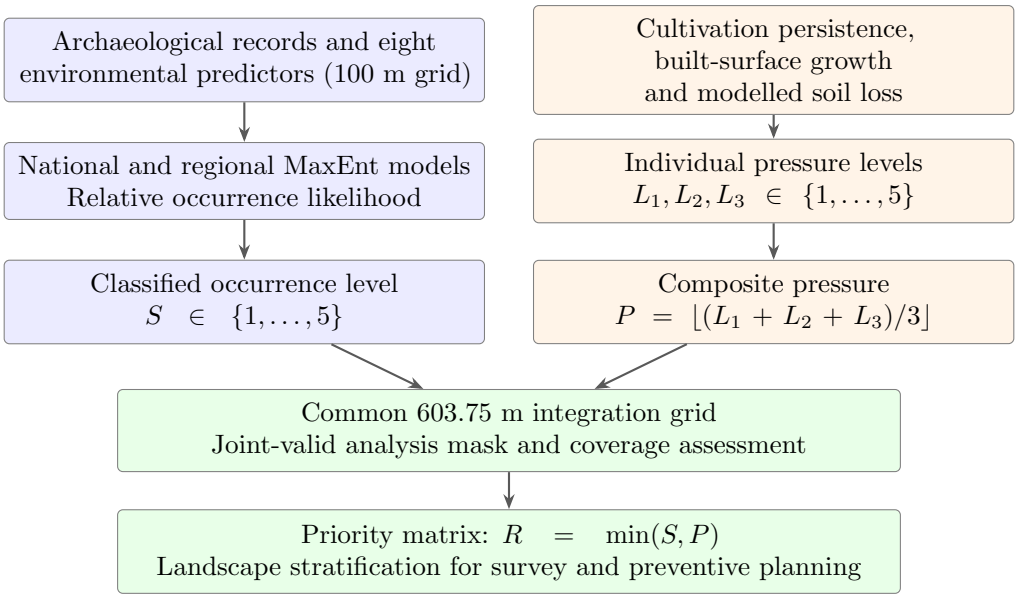


Fig. 2: Workflow. Archaeological occurrence modelling (MaxEnt), land-use pressure modelling (agricultural, urban-expansion and soil-erosion pressure) and their integration into an archaeological risk-priority surface through a priority matrix.

The original *S*-*P* cross-tabulation covers 94,243 km^2. An independent reconstruction of the analysed area, based on valid predictors and composite pressure, covers 94,431 km^2, approximately two-thirds of the national territory, as shown in Fig. 3; the coverage values and record counts reported below refer to this reconstructed mask. As presented in Table 1, coverage was almost complete in Khatlon (98.8%) and Sughd (93.6%), high in DRS (85.4%) and 33.0% in GBAO. The excluded area of approximately 49,400 km^2 lacked either a composite-pressure value (46,500 km^2) or at least one occurrence predictor (2,900 km^2), and 87% of this area lies in GBAO. It consists almost entirely of high-mountain terrain: 93% lies above 3,500 m (median elevation 4,364 m), and ESA WorldCover classifies 65% of it as bare or sparsely vegetated ground and 22% as snow and ice. At the national level, 14,990 of the 15,728 records (95.3%) fall within the analysed area, compared with 939 of 1,625 (57.8%) in GBAO. A considerable proportion of the recorded archaeology of the Pamirs is therefore included in the occurrence models and falls outside the integrated assessment.

Table 1: Analytical coverage and archaeological records by region. Areas are cells of the 603.75 m integration grid whose centre falls within each region. The reconstructed analysed area comprises cells with a composite-pressure value and complete occurrence predictors at the cell centre. Its total (94,431 km^2) differs by 0.2% from the original *S*-*P* cross-tabulation (94,243 km^2). The reconstruction serves as a sensitivity baseline, as described in Methods. Presence cells are the 100 m cells with at least one record and complete predictors used by each model.

Region	Area (km^2)	Analysed area (km^2)	Analysed area (%)	Records	Records in analysed area (%)	Presence cells
DRS	28,759	24,548	85.4	7,709	99.7	5,246
GBAO	64,473	21,266	33.0	1,625	57.8	1,460
Khatlon	24,045	23,747	98.8	2,662	99.2	1,942
Sughd	26,572	24,871	93.6	3,732	99.8	2,269
Tajikistan	143,850	94,431	65.6	15,728	95.3	10,918

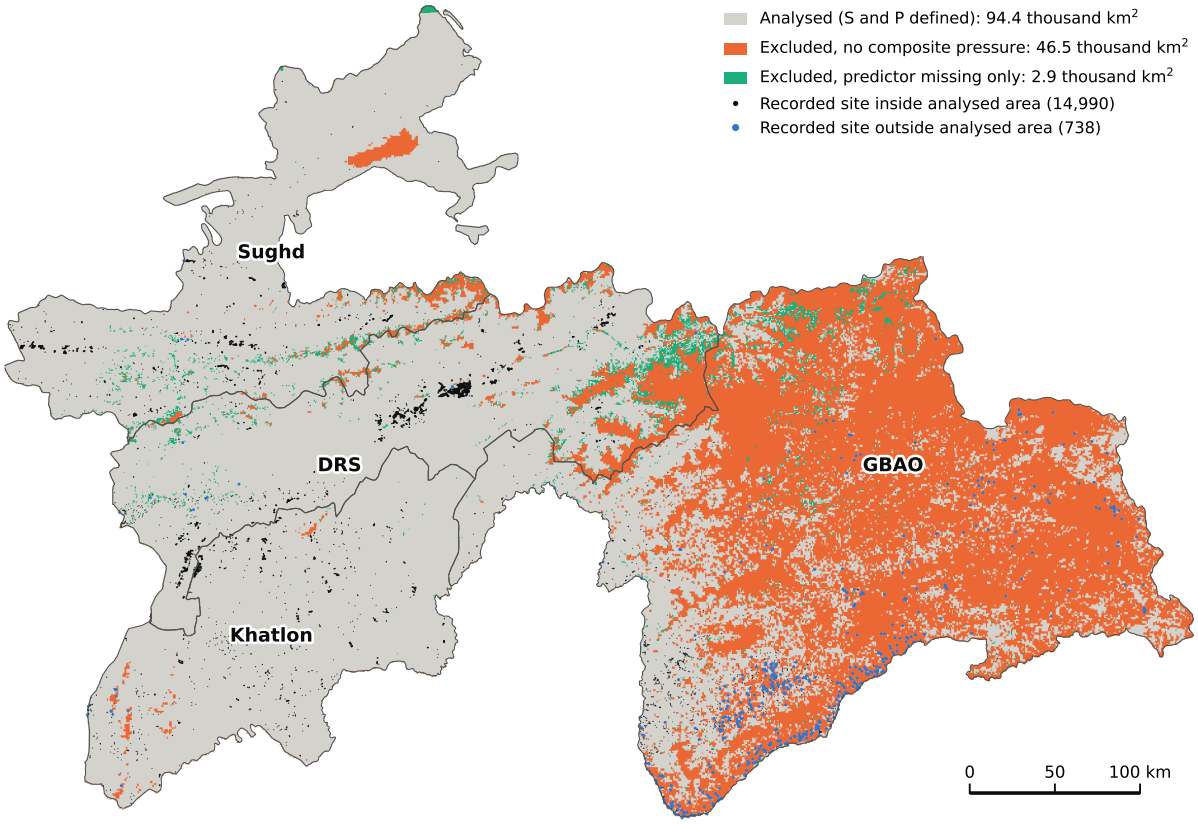


Fig. 3: Coverage of the integrated analysis. Grey: reconstructed analysis cells with a composite-pressure value and complete occurrence predictors at the cell centre. Orange: cells inside Tajikistan without a composite-pressure value. Green: cells with a pressure value and at least one missing occurrence predictor. Dots: recorded archaeological sites inside (black) and outside (blue) the analysed area. Lines: boundaries of the four regional modelling units. Projection: WGS 84 / UTM zone 41N.

National occurrence model

The national model discriminated recorded sites well from the random landscape background, and this performance becomes apparent when the evaluation background is defined explicitly. The classified national surface was fitted with eight terrain, hydrological, land-surface and night-time-light predictors, as listed in Supplementary Table S1 and mapped in Supplementary Fig. S1. As shown in Fig. 4, the surface displays a pronounced spatial contrast: high occurrence levels cover the valley systems, foothill-plain transitions and inter-montane basins of western and southwestern Tajikistan, whereas low levels predominate in the high ranges and on the Pamir plateau.

As shown in Fig. 5 and presented in Table 2, the national model evaluated against random background locations achieved a mean test AUC of 0.864 ± 0.005 (mean ± s.d.) under random ten-fold cross-validation and 0.846 ± 0.035 under spatially blocked five-fold cross-validation; per-fold values are presented in Supplementary Tables S2 and S3. For the same ten models, the AUC reported by MaxEnt is considerably lower (0.683 ± 0.008). This difference reflects the evaluation background. Under default settings, MaxEnt adds the training presences to the background; with 10,918 presence cells and 10,000 random background points, approximately

half of the points used to compute the AUC were therefore presence cells (Methods). A refit with 100,000 random background points and no added presences supported the higher value (0.863 ± 0.007). Blocked AUC was similar with 25 km blocks (0.857 ± 0.024), and the exclusion of training data within 10 km or 25 km of the held-out blocks reduced it slightly (0.839 ± 0.035 and 0.833 ± 0.026, on average 0.007 and 0.013 lower than for the same folds without a buffer). AUC decreased when 100 km or 200 km blocks were held out (0.777 ± 0.068 and 0.812 ± 0.086), as presented in Supplementary Table S7. Because these folds remove large contiguous parts of the country, the resulting values approximate transfer between regions.

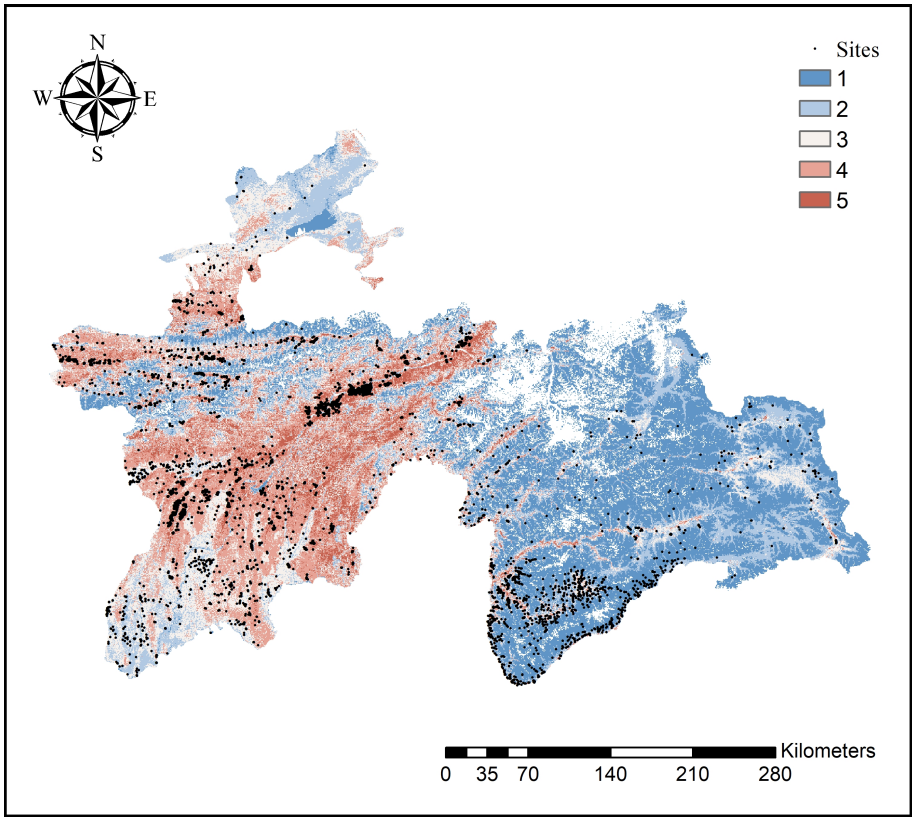


Fig. 4: National classified archaeological occurrence surface. Levels 1-5 are equal-interval classes of the MaxEnt cloglog output, from low to high relative occurrence likelihood. Black dots are the recorded sites used in modelling, shown for visual comparison. White: outside Tajikistan or no predictor data.

As presented in Table 3, elevation was the most influential predictor (mean permutation importance 42.6%), followed by slope (17.0%), land cover (13.9%) and NDVI (13.8%). Aspect and distance to rivers each contributed about 6%, and TWI and night-time light each less than 1%. The response curves shown in Fig. 6 describe the variation of relative occurrence likelihood with each predictor. Likelihood increased from the lowlands to a broad maximum at about 1,500-2,500 m, declined steadily above about 2,700 m and approached zero above about 5,000 m. It peaked on moderate slopes of about 10-20° and decreased on flat ground and on steeper terrain. Likelihood was higher on north-facing than on east- to west-facing slopes, peaked at intermediate TWI values and at about 2 km from high-order rivers, and declined with greater distance from rivers. It increased steeply once NDVI exceeded zero and peaked at about 0.5. Among land-cover classes, it was highest for grassland and cropland, followed by built-up land, and it remained

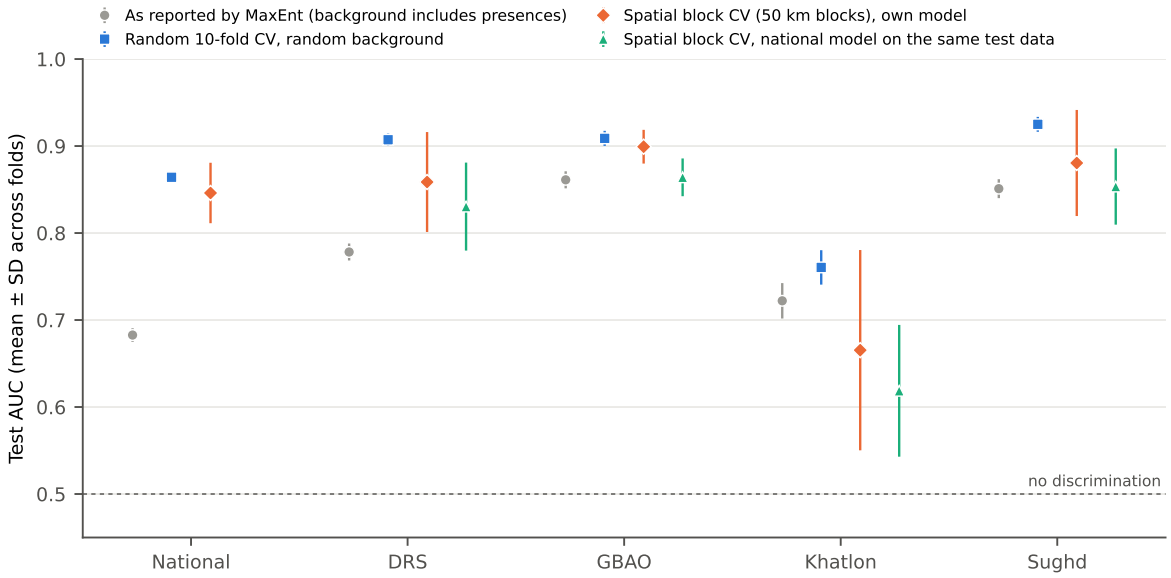


Fig. 5: Discrimination of the national and regional occurrence models. Test AUC (mean ± s.d. across folds) under four evaluation designs. Grey: AUC as reported by MaxEnt with default settings, in which training presences are added to the background (ten-fold random cross-validation). Blue: the same ten models evaluated against a random sample of 60,000 background cells of the same domain. Orange: spatially blocked five-fold cross-validation (50 km blocks), with each model evaluated on held-out presences and background of its own domain. Green: the national model of the same spatial fold, evaluated at the same held-out regional locations as the orange symbol, using national predictor values (Methods). The dashed line marks no discrimination (AUC = 0.5).

nearly constant across the range of night-time light. A model restricted to the five terrain and hydrological predictors achieved comparable discrimination (random five-fold cross-validation 0.849 ± 0.002, compared with 0.865 ± 0.002 for the eight-predictor model under the same design; spatial blocks 0.830 ± 0.037). Under these evaluation designs, terrain and hydrology therefore appear to account for most of the discriminatory signal.

The national surface distributes its classes unevenly among regions. On the 100 m cells where both national and regional predictions are valid, the two highest national levels cover 50.0% of DRS and contain 96.1% of its presence cells. In GBAO, the same levels cover 4.7% of the area and contain 36.4% of the presence cells, whereas 31.4% of GBAO presence cells fall in the two lowest levels, as presented in Supplementary Table S4. Approximately three-quarters of the GBAO records (74%) lie in the southwestern Pamirs, south of 38°N and west of 73°E, and the median elevation of all GBAO records is 3,531 m. At these elevations the national elevation response is already declining, and the national classification consequently assigns low levels to many recorded presence cells. The Pamirs contain a substantial archaeological record in environments that differ from those of the western valleys, which dominate the national calibration.

Regional calibration

Regionally calibrated models outperformed the national model in every region, including when both used the same inputs. As shown in Fig. 5 and presented in Table 2, the regional models discriminated recorded sites well within DRS, GBAO and Sughd, and to a lesser degree within Khatlon. Under random cross-validation against a random background, test AUC was 0.907 in

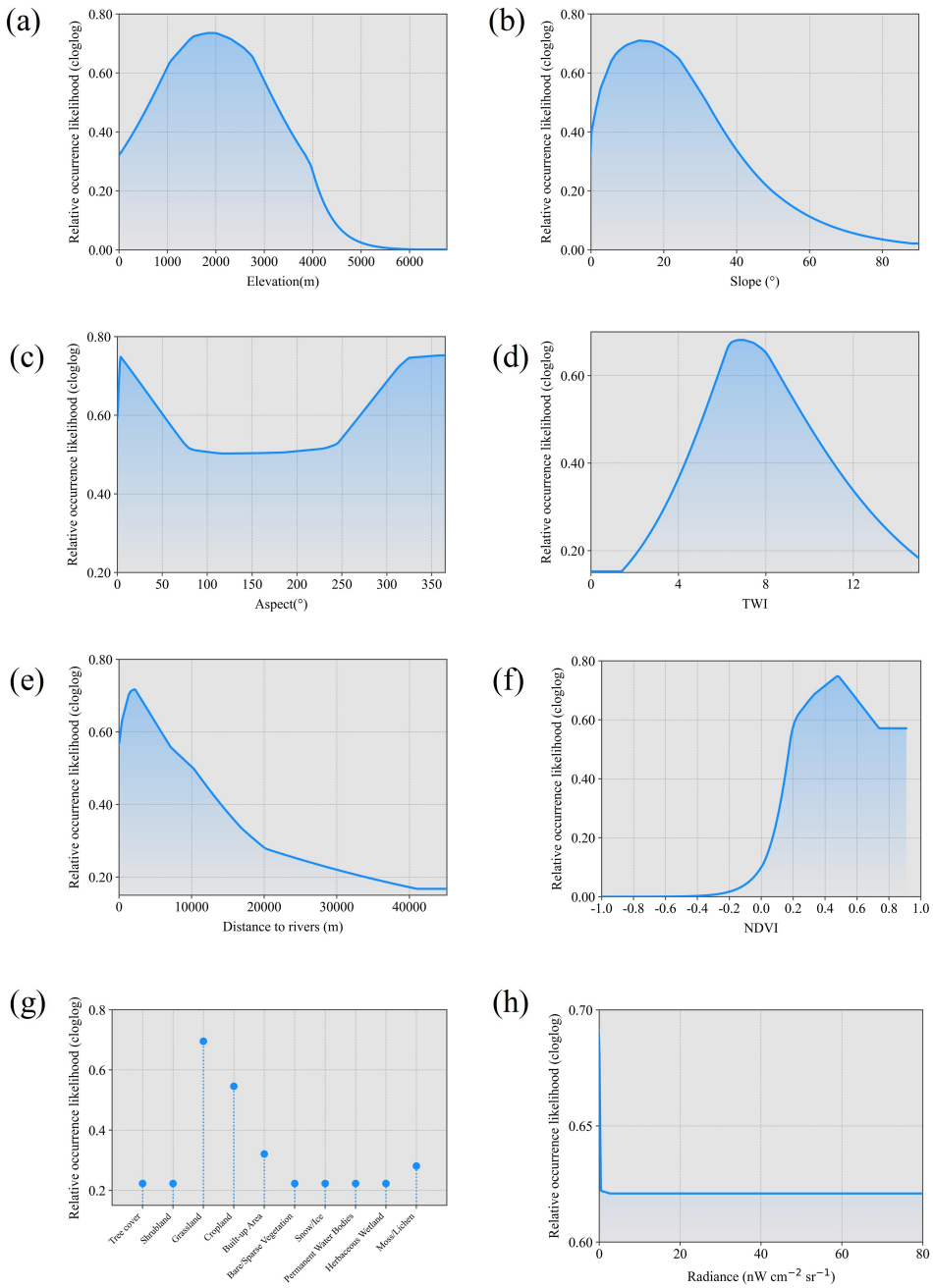


Fig. 6: Response curves of the national MaxEnt model. Marginal response of relative occurrence likelihood (cloglog output) to (a) elevation, (b) slope, (c) aspect, (d) topographic wetness index (TWI), (e) distance to high-order rivers, (f) NDVI, (g) land-cover class (ESA WorldCover) and (h) night-time light radiance, with the other predictors held at their mean values.

DRS, 0.909 in GBAO, 0.925 in Sughd and 0.760 in Khatlon; spatial blocking reduced all values and increased their variability (DRS 0.859, GBAO 0.899, Khatlon 0.665, Sughd 0.881). To ensure a consistent comparison, both models were evaluated at the same held-out locations: in each spatial fold, the regional and national models were scored on the same held-out presence cells and the same background cells of the region. Mean regional-model AUC was higher in all four regions, by 0.028 in DRS, 0.035 in GBAO, 0.047 in Khatlon and 0.027 in Sughd, and higher in 16 of the 20 region-fold comparisons. Because the national and regional slope layers differ, the regional models were also refitted with the national predictor rasters. With the same inputs, mean regional AUC remained higher by 0.023 in DRS, 0.032 in GBAO, 0.045 in Khatlon and 0.030 in Sughd, again in 16 of the 20 comparisons, as presented in Supplementary Table S6. A national-type model trained on the other three regions and applied to the fourth reached 0.793 in DRS, 0.833 in GBAO, 0.809 in Sughd and 0.603 in Khatlon, as also presented in Table 2. The regional models were also more selective. In DRS, Khatlon and Sughd, they concentrated recorded sites in considerably smaller top classes; in Sughd, for example, the two highest regional levels cover 7.2% of the common valid area and contain 71.6% of its presence cells, compared with 24.6% and 84.2% for the national surface. In GBAO, the regional top levels cover a share of the common valid area similar to that of the national levels (6.9% and 4.7%, respectively) and contain a considerably larger proportion of its presence cells (60.7% and 36.4%). These shares are presented in Supplementary Table S4.

Table 2: Test AUC of the national and regional MaxEnt models (mean ± s.d. across folds). 'MaxEnt default': AUC reported by MaxEnt when training presences are added to the background (the setting used for the original occurrence maps). 'Random': the same ten-fold cross-validation models scored against random background cells. 'Spatial blocks': five-fold cross-validation with 50 km blocks. 'National model': the national model of the same spatial fold, scored on the same held-out regional presences and background. 'Other regions': a model trained on the other three regions and scored on all presence and background cells of the region.

	Presence	Random ten-fold CV		Spatial blocks (five-fold)		Other
Model	cells	MaxEnt default	Random	Own model	National model	regions
Tajikistan (national)	10,918	0.683 ± 0.008	0.864 ± 0.005	0.846 ± 0.035	-	-
DRS	5,246	0.778 ± 0.010	0.907 ± 0.007	0.859 ± 0.057	0.830 ± 0.051	0.793
GBAO	1,460	0.861 ± 0.010	0.909 ± 0.009	0.899 ± 0.019	0.864 ± 0.022	0.833
Khatlon	1,942	0.722 ± 0.020	0.760 ± 0.020	0.665 ± 0.115	0.619 ± 0.076	0.603
Sughd	2,269	0.851 ± 0.011	0.925 ± 0.009	0.881 ± 0.061	0.853 ± 0.044	0.809

As presented in Table 3, the relative importance of the predictors also varied among regions. Elevation dominated in GBAO (60.5%), Sughd (56.9%) and DRS (47.5%). In Khatlon, NDVI was the leading predictor (32.2%), followed by elevation (25.4%) and land cover (19.9%), which is consistent with a lowland region in which most of the landscape is environmentally suitable and land-use differences carry more of the signal. Distance to rivers was more important in Sughd (11.1%), and aspect in DRS (14.0%), than in the other regions. The regional surfaces are shown in Fig. 7.

Table 3: Permutation importance (%) of the predictors in the national and regional MaxEnt models (mean of the ten cross-validation replicates).

Predictor	Tajikistan	DRS	GBAO	Khatlon	Sughd
Elevation (DEM)	42.6	47.5	60.5	25.4	56.9
Slope	17.0	16.7	21.9	8.9	6.0
Aspect	6.1	14.0	1.6	4.4	9.3
TWI	0.5	0.7	0.3	1.4	0.6
Distance to rivers	5.5	4.3	1.0	6.7	11.1
NDVI	13.8	6.5	13.9	32.2	9.5
Land cover (WorldCover)	13.9	9.7	0.3	19.9	6.2
Night-time light	0.7	0.7	0.6	1.0	0.4

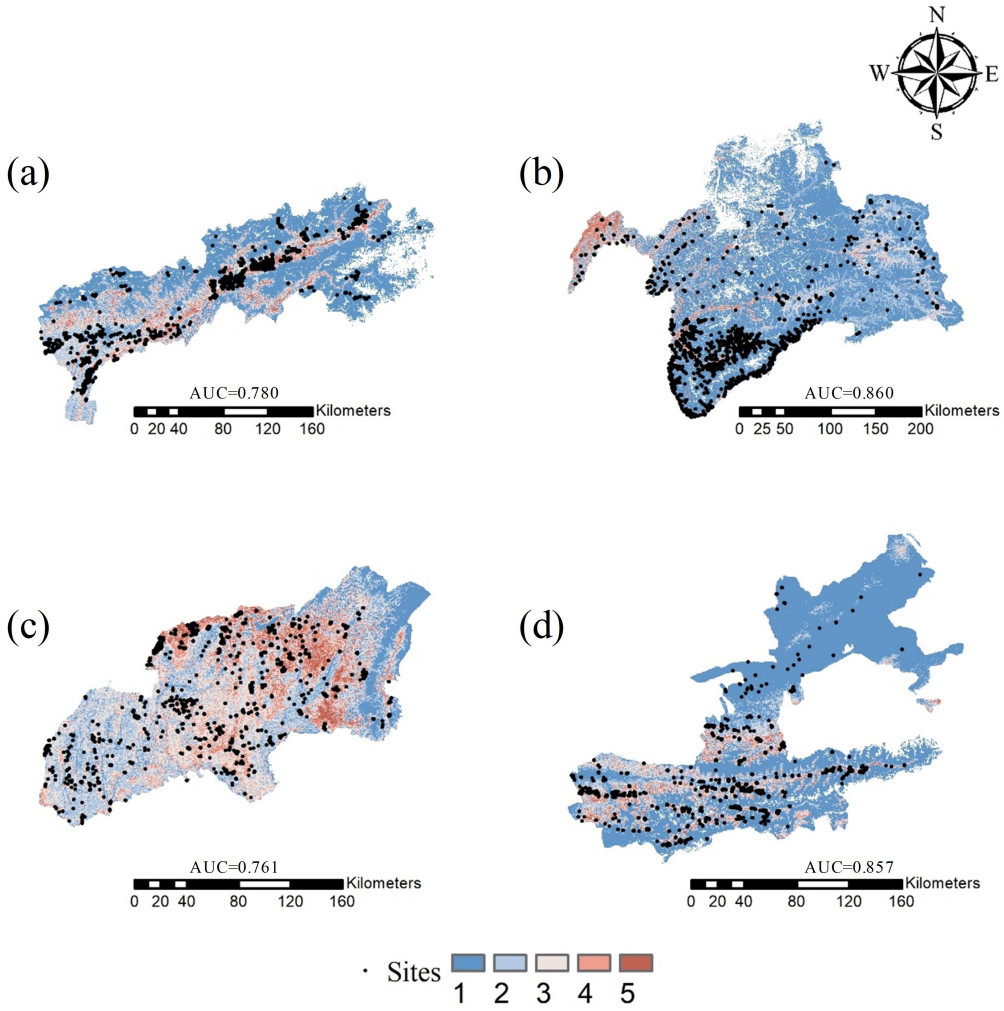


Fig. 7: Regional classified occurrence surfaces. (a) DRS, (b) GBAO, (c) Khatlon and (d) Sughd. Levels 1-5 are equal-interval classes of the cloglog output of each regional model. Black dots: recorded sites.

Land-use and erosion pressure

As shown in Fig. 8, the three pressure indicators display distinct spatial patterns, and high composite pressure is confined to less than 1% of the analysed area. Cultivation persistence, mapped in Fig. 8a, is concentrated in the Sughd lowlands, the western and southwestern cultivated plains, and the Hissar and Vakhsh valley corridors. High values form continuous belts along river-fed valley floors and decrease markedly on the adjacent uplands, where arable land is restricted to low-relief, water-accessible terrain. Built-surface growth, mapped in Fig. 8b, is considerably more localised. It forms clustered patches, peripheral belts and linear fringes around towns and along valley corridors, consistent with edge expansion, corridor development and fragmented peri-urban infill. Its footprint is much smaller than that of cultivation, although it represents a largely irreversible form of land conversion. Erosion pressure, mapped in Fig. 8c, varies at a broader regional scale; it is highest in western and southwestern Tajikistan and in parts of the north, and generally low in the east.

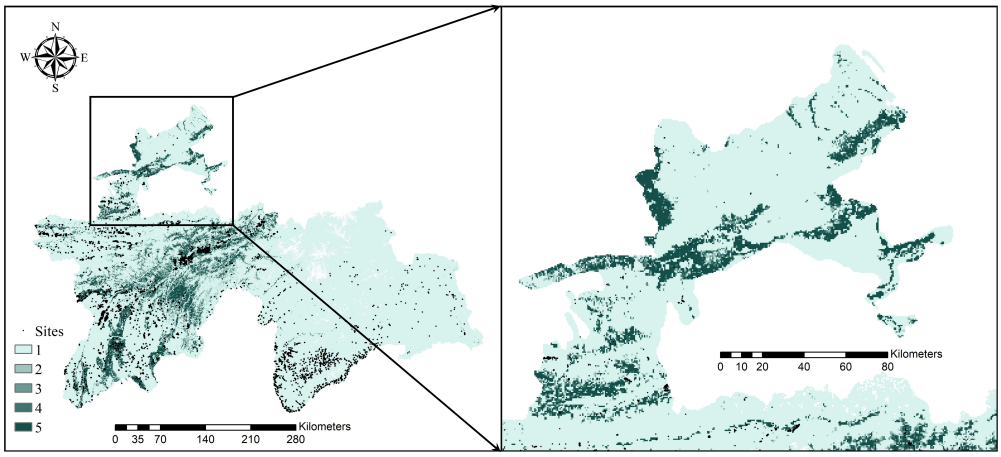
As shown in Fig. 9, low levels predominate in the composite pressure surface. Within the analysed area, levels 1 and 2 cover 56.9% and 33.8%, level 3 covers 8.6%, and levels 4 and 5 together cover 0.8% (771 km^2). The highest values occur where several pressures overlap in accessible, long-occupied terrain, namely the irrigated valleys, basin margins, foothill belts and mountain-plain transitions of western, southwestern and northern Tajikistan, and they extend beyond the most urbanised areas. Most recorded sites lie under low to moderate composite pressure: of the records with a pressure value, 7,592 are at level 1, 6,544 at level 2, 791 at level 3 and 81 at levels 4 or 5. A similar pattern was reported for southern Tajikistan, where elevated risk was likewise confined to limited areas.[3]

Integrated risk-priority surface

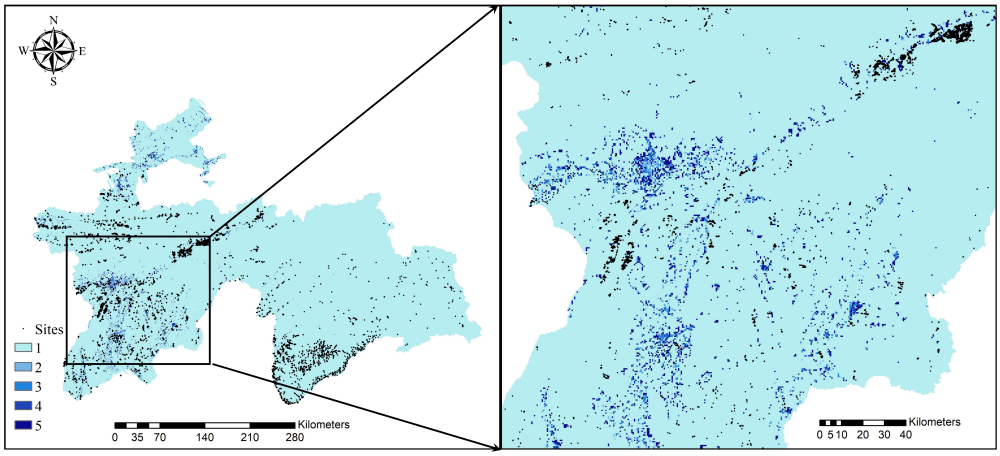
High-priority areas account for a small proportion of the analysed area. As shown in Fig. 10 and presented in Table 4, the integrated surface under the original classification is dominated by low and moderate priority: level 1 covers 59.40% of the analysed area, level 2 32.88% and level 3 7.52%, whereas level 4 covers 0.20% and level 5 0.001%. A large part of this low priority results from elevated occurrence under low pressure; the $S = 4$ row alone accounts for 27.20% of the analysed area, mostly at $P = 1$ (14.10%) or $P = 2$ (11.48%). Cells with both high occurrence and high pressure ($S \geq 4$, $P \geq 4$) cover 192.10 km^2, or 0.20% of the analysed area. These cells are concentrated in the irrigated valleys and expanding settlement corridors of western Tajikistan, where they form discontinuous belts and small patches. Because occurrence levels in the west are predominantly high, the priority surface there largely follows composite pressure. The dependence of this extent on the occurrence model and its classification is examined in the following section.

Stable core and configuration-dependent fringe

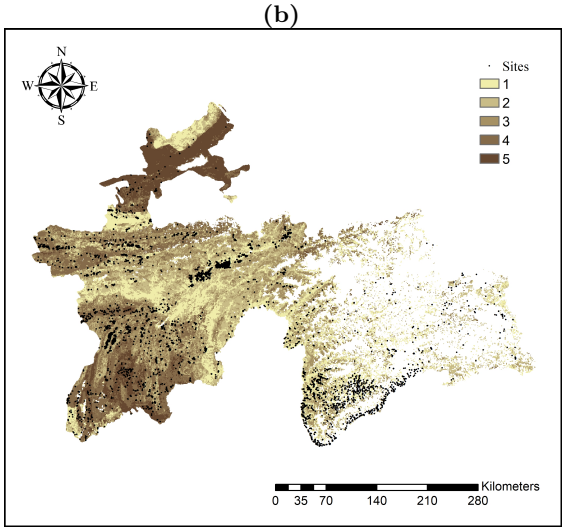
The extent of the high-high class varies with modelling and classification choices, whereas a stable core persists across all of them. S was reconstructed from the re-run models (Methods),



(a)



(b)



(c)

Fig. 8: Pressure indicators. (a) Agricultural pressure L_1 (cultivation persistence, Sentinel-2 NDVI; valid years 2020, 2021 and 2024), with an enlarged view of the Sughd valleys. (b) Urban-expansion pressure L_2 (annualised growth of GHSL built surface, 1990-2020), with an enlarged view. (c) Soil-erosion pressure L_3 (GloSEM soil loss, 2012). Levels 1-5 from low to high; black dots are recorded sites; white areas are outside Tajikistan or lack data.

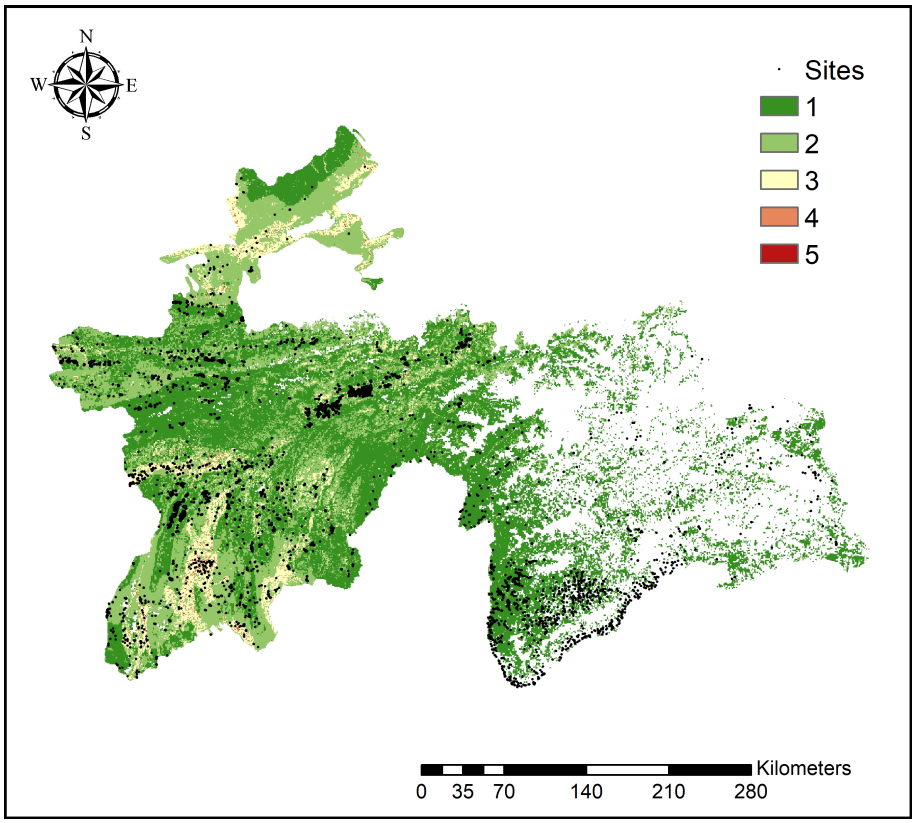


Fig. 9: Composite land-use pressure P. Equal-weight mean of the three ordinal pressure levels, truncated to an integer (Methods). Levels 1-5 from low to high; black dots: recorded sites; white: outside Tajikistan or no data, as shown in Fig. 3.

Table 4: Area (km^2) and share of the analysed area (94,243.34 km^2) of each combination of classified occurrence level S and composite pressure level P. Cells without data in either surface are excluded, as shown in Fig. 3. Areas are planimetric areas on the UTM 41N grid. This table is retained from the original analysis; its classified occurrence raster and processing settings remain to be verified.

S \ P	$P = 1$	$P = 2$	$P = 3$	$P = 4$	$P = 5$
$S = 1$	13,736.30 (14.58%)	2,341.63 (2.48%)	45.20 (0.05%)	1.46 (0.00%)	–
$S = 2$	11,288.96 (11.98%)	5,949.21 (6.31%)	1,363.64 (1.45%)	115.92 (0.12%)	29.53 (0.03%)
$S = 3$	11,156.28 (11.84%)	9,233.47 (9.80%)	4,503.56 (4.78%)	335.72 (0.36%)	96.60 (0.10%)
$S = 4$	13,289.41 (14.10%)	10,819.83 (11.48%)	1,405.56 (1.49%)	111.91 (0.12%)	10.57 (0.01%)
$S = 5$	4,117.17 (4.37%)	3,479.64 (3.69%)	742.15 (0.79%)	68.89 (0.07%)	0.73 (0.00%)

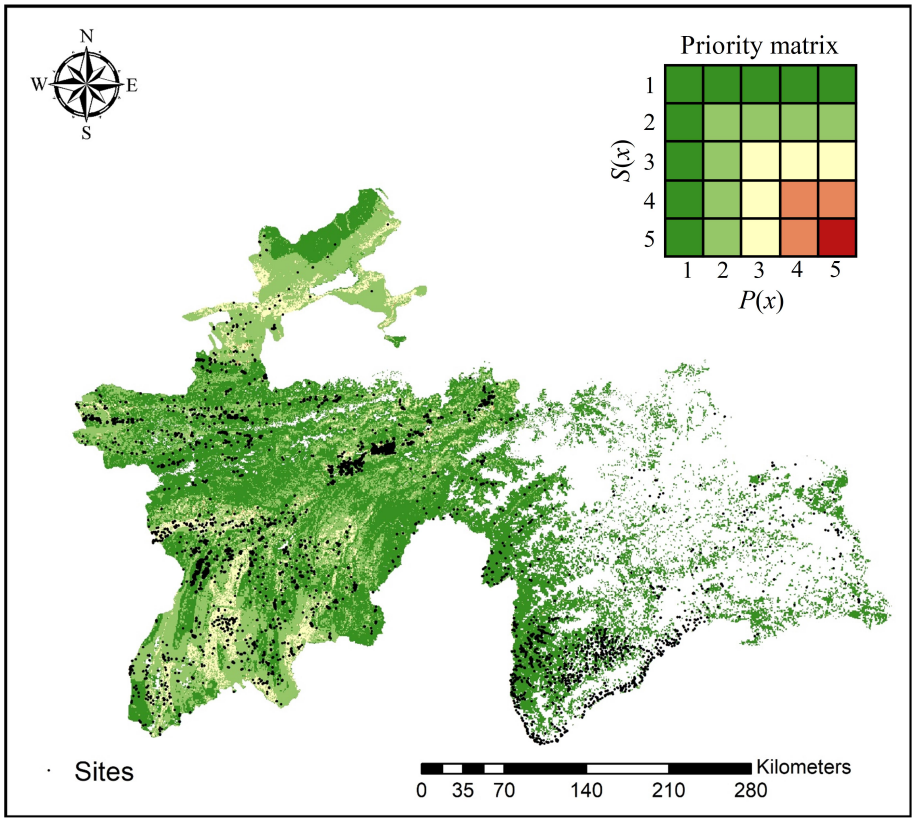


Fig. 10: Archaeological risk-priority surface. Priority level $R = \min(S, P)$ from the classified occurrence level S and composite pressure level P; the inset matrix shows how each S-P combination maps to a priority level. Black dots: recorded sites; white: outside Tajikistan or no data, as shown in Fig. 3.

and four occurrence surfaces were compared: the ten-fold average under default settings, the ten-fold average refitted with 100,000 background points and no added presences, and two matched full-data models with all eight predictors and with only the five terrain and hydrological predictors. Each surface was classified with equal intervals and with quintiles, with the pressure raster held fixed; the comparisons are presented in Supplementary Table S5. Continuous rankings were similar (Spearman correlation 0.89–0.998), whereas the high-high extent of the eight surface-classification combinations ranged from 75 to 248 km^2, compared with 180 km^2 for the default reconstruction and 192 km^2 in the original map presented in Table 4. The largest difference resulted from the refit with 100,000 background points: its equal-interval levels 4–5 covered 12,782 km^2, compared with 32,472 km^2 for the default reconstruction, and almost all of its 75 km^2 of high-high cells lay within the default high-high set. Removal of NDVI, land cover and night-time light affected the composition of the high-high set more than its size (Jaccard index 60% under equal intervals and 57% under quintiles).

Across all eight combinations, 75 km^2 (205 cells) were classified as high-high in every case, and a further 86 km^2 in six or seven combinations, as mapped in Supplementary Fig. S3. The remaining 181 km^2, classified as high-high in five or fewer combinations, constitutes a configuration-dependent fringe. The stable core lies mainly in DRS (49 km^2) and Sughd (15 km^2), with 7 km^2 in Khatlon and 4 km^2 in GBAO. Only 10 of its 205 cells contain a recorded site. Because the pressure raster was held fixed, all of these areas lie within its levels 4–5 (774 km^2, 0.8% of the analysed area), which define the upper limit of the high-high extent; the comparison therefore isolates the effect of the occurrence model and its classification.

Discussion

A national map of archaeological occurrence and land-use pressure can inform survey planning in Tajikistan, and its outputs depend on four methodological choices. The available open pressure data restrict the integrated assessment to approximately two-thirds of the country, leaving much of the recorded archaeology of the Pamirs outside it. The definition of the evaluation background alone changes the reported AUC of the same fitted model from 0.68 to 0.86. A single national calibration obscures regional patterns, particularly in the Pamirs. Finally, the extent of the highest priority class varies with the occurrence model and its classification, whereas a core of approximately 75 km^2, located mainly in the valleys of DRS and Sughd, is classified as high-high under every configuration tested. Few cells of this core contain recorded sites. Because cultivation and construction can also obscure surface traces on imagery, the core constitutes a primary target for field survey, through new programmes or the extension of existing ones, and the surrounding configuration-dependent fringe constitutes a secondary target. Considerably larger areas combine elevated occurrence with low present pressure: the $S = 4$ row alone covers 27% of the analysed area, mostly under the two lowest pressure levels. Low values of these pressure proxies may indicate opportunities for documentation before further change, especially where cultivation or built-up land is expanding into foothills, with site-specific urgency to be assessed in the field. These results support a stratified approach to heritage management.

Reported model performance depends on the reference against which a model is evaluated. The national model discriminates recorded sites from the random landscape background with an AUC of 0.86, or 0.85 under the 50 km spatial-block design, whereas the default MaxEnt AUC of about 0.68 is computed against a background that also contains roughly 9,800 training presence cells in addition to 10,000 random points. The two values are based on different comparison distributions and describe the same fitted model. Studies with large presence datasets should therefore specify the evaluation background and report spatially blocked performance.[17–19,28] Discrimination also decreased when larger contiguous areas were held out (100-200 km blocks or whole regions), indicating that the national surface is more suitable for well-recorded areas than for extrapolation into large, sparsely recorded areas. Regional models achieved higher mean AUC in all four matched regional evaluations, with improvements in 16 of 20 region-fold comparisons, and this advantage persisted with the same predictor rasters, indicating that it arises from regional calibration. Together with the low national classes assigned to many recorded GBAO cells, these results show that a national classification can obscure regional archaeological patterns. Regional models should therefore be evaluated alongside national surfaces, in combination with local validation. Khatlon warrants particular attention. It showed the weakest spatial discrimination (AUC 0.665 ± 0.115 for the regional model), and the regional model improved on the national model in three of five folds. Survey history, preservation and surface visibility are possible explanations and represent topics for further research. In this region, pressure-led prioritisation and systematic field survey could provide complementary evidence.

The occurrence surface should be interpreted according to its definition. It represents the relative likelihood of occurrence in environments similar to those of recorded sites, a quantity distinct from both the probability of presence and heritage value. Low occurrence levels indicate environments that differ from those of recorded sites, and such areas may still contain undocumented remains. Preservation and surface visibility influence what satellite imagery can reveal even under systematic coverage, as illustrated by the GBAO records in the lowest national levels. The framework is an evidence-based prioritisation tool that complements archaeological prospection, and its value is expected to increase as ground observations and digital records accumulate. The announcement by the Government of Tajikistan of a new national digital platform for historical and cultural monuments strengthens this foundation,[29] and the contribution of CAAL to the development of a national heritage database will support this initiative.

The pressure indicators are proxies, and they determine the geographical limits of the integrated result. Cultivation persistence is a rule-based NDVI proxy; validation of its thresholds against independent cropland or field data in Tajikistan, and separation of cultivation from seasonal greening of irrigated pastures or meadows, are priorities for refinement. Recurrent cultivation also represents a pressure on buried and surface deposits, which is distinct from observed damage. Cultivation frequency was calculated from the valid observation years 2020, 2021 and 2024, and extension of the record to consecutive years would broaden the temporal coverage of the indicator. The urban indicator measures relative growth of built surface between 1990 and 2020, so established towns with little recent growth receive low scores. The erosion layer is model-based, refers to 2012 and has a resolution of approximately 25 km, coarser than that of

the other two indicators. Because the three indicators cover different periods, their composite represents a summary of several multi-period pressure proxies. This has two implications. First, the composite pressure layer covers western and central Tajikistan and approximately one-third of GBAO, leaving much bare, rocky and glaciated terrain for future assessment; the contribution of the individual input masks to this pattern requires further analysis. The recorded sites of the Pamirs outside this area require a pressure set suited to high mountains, including grazing, road construction, mining and periglacial processes. Second, the composite rule attenuates single strong pressures. A cell with very high urban expansion and low cultivation and erosion (levels 1, 5 and 1) receives $P = 2$, and the minimum rule then retains a low priority even where occurrence is high. Planners should therefore interpret the priority surface together with the individual pressure layers and adjust the weights where one process dominates. Finally, priority classes respond to classification choices even when continuous rankings are similar. Reporting the cells that remain high-high across configurations, as in this study, distinguishes a stable core from a configuration-dependent fringe and provides a robust basis for initial prioritisation; extension of the analysis to pressure-layer uncertainty is a logical next step. The occurrence AUCs assess the discrimination of recorded locations, and validation of the pressure proxies, damage to sites and management outcomes requires field evidence.

More broadly, the results relate to a shift in archaeological heritage practice towards understanding cultural landscapes through the relationships among places, routes, land-use practices and material traces, beyond collections of discrete monuments.[30,31] This shift is particularly relevant in Central Asia, where the evidence derives from heterogeneous survey histories. The present framework shares the area-based, GIS-enabled logic of Historic Landscape Characterisation (HLC), which characterises pattern, time depth and change across whole landscapes,[32] while remaining a distinct, prioritisation-oriented approach. Development initiatives in Tajikistan increasingly link heritage with tourism, rural livelihoods and infrastructure,[33,34] and such development entails both opportunities and risks for heritage. Spatial evidence of the type produced here can identify where development, landscape transformation and archaeological sensitivity converge, and where information gaps should be addressed before irreversible decisions are made. Such evidence complements field survey, local expertise, community knowledge and site-specific assessment.

The usefulness of a national heritage-priority map depends on the transparency of the methodological choices on which it is based. For Tajikistan, the national model discriminated recorded sites from the random landscape background under all evaluation designs tested, the regional models performed better in matched comparisons, and the combination with open pressure indicators identified a stable core of approximately 75 km^2 in western and northern Tajikistan, where high occurrence likelihood and high land-use pressure coincide under every configuration, surrounded by a wider fringe whose classification depends on modelling choices. Application to other countries with incomplete inventories and ongoing land transformation would require local evaluation of model performance, pressure proxies and data coverage. Future work should include field verification, beginning with the stable core; condition assessments that link pressure to damage at recorded sites;[27] pressure indicators suited to high mountains; and

regular updates as national digital records expand. The results are also intended to inform communication between CAAL and national heritage authorities, as a reproducible basis for discussing survey priorities, documentation gaps and national heritage planning.

Methods

Study area

Tajikistan covers about 142,100 km^2 in Central Asia (36°40′-41°05′N, 67°31′-75°14′E), as shown in Fig. 1a. Elevation rises from about 300 m in the western river valleys and alluvial fans to almost 7,500 m in the Pamirs, and about 90% of the territory lies above 600 m and is considered mountainous.[35,36] Productive settlement landscapes are concentrated in foothill-valley transitions of the Amu Darya and Syr Darya basins. These rivers are fed mainly by mountain runoff, and Tajikistan is a major source region for the Amu Darya.[37] Precipitation in the Western Pamirs ranges from 90 to 400 mm per year and peaks in winter.[38] Irrigation, which expanded substantially in the second half of the twentieth century, is central to arable farming,[39] and in the lowlands irrigation and agrochemicals are associated with declining ecological condition.[35] Pasture accounted for more than 75% of agricultural land between 1960 and 2006.[39] The high plateaus and steep mountain belts are sparsely populated and geomorphically dynamic, and deforestation and over-grazing are the main threats to their environments.[35,40] For regional analysis we used four administrative units: Sughd Region, Khatlon Region, Gorno-Badakhshan Autonomous Region (GBAO) and the Districts of Republican Subordination (DRS), as shown in Fig. 1b.

Archaeological records

Site locations were compiled from national inventories and from systematic mapping of archaeological features on satellite imagery, including Google Satellite, Bing Aerial, ESRI World Imagery and CORONA historical imagery.[3] Records from national inventories are biased by the historically uneven discovery of sites. The remote-sensing component, which makes up most of the dataset, was produced by a single researcher who inspected the whole territory of Tajikistan systematically, so remote-sensing coverage is consistent across the country, and the uneven distribution of detected sites reflects factors such as settlement history, preservation and visibility. Preservation and surface visibility nevertheless shape what can be detected on imagery, so locations without detected features enter the analysis as background locations. Each record is a georeferenced point without attributes of site type or period, so the models describe the combined distribution of all recorded archaeological occurrences. Records were supplied as points in WGS 84 / UTM zone 41N (EPSG:32641); the released point and raster files carry no projection metadata, and their coordinates align with the georeferenced composite-pressure raster on this projection; the four regional files partition the national file. For modelling, records were thinned to one per 100 m predictor cell, as in the default duplicate removal of MaxEnt, and records in cells with a missing predictor were dropped, as presented in Table 1.

Environmental predictors

Eight predictors, listed in Supplementary Table S1 and mapped in Supplementary Fig. S1, were grouped by the aspect of settlement logic they represent. Terrain setting was described by elevation, slope and aspect, which relate to accessibility, construction feasibility, microclimate, erosion sensitivity and long-term preservation.[41,42] Hydrological accessibility was described by distance to high-order rivers, a proxy for stable water, floodplain environments and movement corridors, and by the topographic wetness index (TWI), a terrain-based proxy for moisture retention and agricultural suitability.[43] Surface context was described by NDVI and land cover (ESA WorldCover), which characterise the present agro-ecological background and ground visibility. They are interpreted as proxies for present landscape context. Night-time light represents present-day infrastructure density and disturbance, and was included to capture modern filtering of the archaeological record. All predictors were provided on a common 99.99 m grid (approximately 100 m) in UTM zone 41N (6,980 × 4,766 cells nationally), with regional subsets on the same grid. This is the modelling resolution, which can differ from the native resolution of the source data. Supplementary Table S1 summarises the available predictor information and refers to the preprocessed modelling inputs distributed in release v1.0.0 of the project repository (Data availability). For the re-analysis, seven of the eight regional predictor layers matched their national counterparts on overlapping cells, whereas the slope layers differed (correlations of 0.89–0.91).

Occurrence modelling

Occurrence was modelled with MaxEnt version 3.4.4,[15,44,45] a presence-background method that estimates a relative suitability surface by maximising entropy subject to constraints on the environmental features of presence locations, with regularisation to limit over-fitting. Default settings were used: automatic feature classes, a regularisation multiplier of 1, 10,000 random background cells, addition of presence cells to the background, and at most 500 iterations. WorldCover was treated as categorical. Outputs used the cloglog transform. Each model was run with ten-fold random cross-validation. Each original occurrence map was derived from a single replicate of the ten-fold run. For the re-analysis, mapped predictions were averaged across the ten replicate models, and permutation importance was also averaged across replicates. The matched eight- and five-predictor full-data models used for the additional mapping comparisons are described below. The four regional models used the same settings with regional rasters and record files.

We evaluated the models in five ways, reporting test AUC as the mean and standard deviation across folds. (i) AUC as reported by MaxEnt. With default settings, MaxEnt computes this AUC against a background that includes the training presences it has added. For the national model these made up about half of the roughly 19,800 background points, so the comparison distribution differs from a purely random landscape background. (ii) The same ten models, scored on their held-out presence cells against 60,000 random cells with complete predictors drawn from the same domain. Predictions were computed from the models' `.lambdas` files with

an independent implementation of the MaxEnt cloglog transform, which reproduced the sample predictions of MaxEnt itself to within 3×10^{-16} for all 60 ten-fold replicate models and all 111 samples-with-data models (validation logs in the re-analysis package). (iii) Spatially blocked five-fold cross-validation.[18,19] The national grid was divided into 50 km square blocks. Blocks containing presence cells were assigned greedily to reduce imbalance in regional presence counts across folds; blocks without records were assigned pseudo-randomly. For each fold, a model with the same MaxEnt settings was fitted in samples-with-data mode, using the presence cells outside the held-out blocks and 10,000 background cells from the training blocks. It was scored on the held-out presence cells against all pool cells in the held-out blocks. A random five-fold version of the same design reproduced the values of (ii) to within 0.003. In this design, adjacent blocks could belong to different folds and no separation buffer was imposed. To examine sensitivity across evaluation designs, the national evaluation was repeated with 25, 100 and 200 km blocks (same assignment rule), and with 50 km blocks after removing training presences and background within 10 km or 25 km of any held-out block. The buffered runs used the same folds, held-out presences and held-out background as the main 50 km design. (iv) National versus regional models at the same test locations. In each spatial fold, the national fold model (using national predictors) and the regional fold model (using regional predictors) were scored at the same held-out presence and background locations in the region. Predictor values were extracted from the rasters of each model. Because the regional slope layers differ from the national one, the regional fold models were also refitted with the national predictor rasters (same training cells, background sample and folds) and compared with the national fold models on the same inputs. (v) Transfer between regions. A model trained on the presence cells of three regions, with 10,000 background cells from the national pool outside the fourth region, was scored on all presence and pool cells of the fourth. AUC was computed as the Mann-Whitney statistic, with ties counted as one half. Two sensitivity models were also fitted. The first was a national model with 100,000 background points and no addition of presences to the background, run with ten-fold cross-validation. The second was a model with only the five terrain and hydrological predictors, evaluated with the random and spatially blocked five-fold designs of (iii). For mapping comparisons, full-data models with eight and with five predictors were fitted to the same 10,918 presence cells and the same 10,000 background cells.

Classified occurrence levels

The continuous national surface was processed in ArcGIS for Desktop 10.2 (Esri), harmonised to the common 603.75 m analytical grid, and reclassified into five equal-interval levels based on the full valid value range of the Tajikistan raster, $S(x) \in \{1, 2, 3, 4, 5\}$, with higher levels indicating greater relative occurrence likelihood. $S(x)$ denotes the classified level, distinct from the continuous output. For the sensitivity analysis, we reconstructed S from the re-run ten-fold mean surface by nearest-neighbour sampling at the centres of the 603.75 m cells. Equal-interval breaks were calculated from the full valid value range of each compared surface on its 100 m grid. The reconstructed analysed area differs from Table 4 by 0.2% and the high-high area by 6.1%, differences consistent with the resampling and classification settings of the original ArcGIS

processing. The reconstruction is therefore treated as a separate sensitivity baseline, and the interpretation of high-priority areas rests on the stable core defined below, which is common to all compared configurations.

Quintile classes, used to compare configurations, have breaks at the 20th, 40th, 60th and 80th percentiles of the same 259,062 integration-grid cells with complete predictors and valid pressure. For each ten-fold surface, each replicate was evaluated at cell centres using its saved coefficients, then predictions were averaged before classification. The matched full-data models were evaluated directly at the same cell centres. For each pair of surfaces we report the share of the baseline high-high cells retained, the share of the alternative high-high cells lying within the baseline set, and the Jaccard index (intersection over union); quintile and equal-interval comparisons use their own baselines. Finally, for each integration cell we counted in how many of the eight combinations of the four surfaces and two classification schemes it was high-high ($S \geq 4$ and $P \geq 4$); cells that were high-high in all eight form the stable core. This count reproduced the high-high cell numbers presented in Supplementary Table S5. Records were assigned to integration cells by their coordinates.

Agricultural pressure

For each pressure type k, an ordinal field $L_k(x) \in \{1, \ldots, 5\}$ describes pressure intensity at location x, with $k = 1, 2, 3$ denoting agricultural, urban-expansion and soil-erosion pressure. Agricultural pressure L_1 represents recurrent cultivation across multiple valid observation years within 2020-2024, because repeated ploughing, irrigation and land levelling progressively affect shallow deposits and surface visibility.[1] Sentinel-2 surface reflectance (collection `COPERNICUS/S2_SR_HARMONIZED` in Google Earth Engine[46]) was filtered to scenes with less than 30% cloudy pixels. Clouds were masked with the QA60 opaque and cirrus bits (10 and 11) and Scene Classification Layer classes 3, 8, 9, 10 and 11, and NDVI was calculated from bands B8 and B4. For each year t, let $\mathcal{G}_t(x)$ be the valid observations from April to October and $\mathcal{A}_t(x)$ all valid observations of that year. A year was treated as valid ($O^{(t)} = 1$) if $\mathcal{G}_t(x)$ contained at least one observation, and crop presence was recorded when all four criteria held:

$$\begin{aligned} C^{(t)}(x) = 1 \iff & \max_{\mathcal{G}_t} \mathrm{NDVI} > 0.45, \quad \overline{\mathrm{NDVI}}_{\mathcal{G}_t} > 0.35, \\ & \max_{\mathcal{A}_t} \mathrm{NDVI} - \min_{\mathcal{A}_t} \mathrm{NDVI} > 0.25, \quad \frac{N^{(t)}_{\mathrm{NDVI}>0.35}}{|\mathcal{G}_t|} > 0.40. \end{aligned} \tag{1}$$

Invalid years were treated as missing values. Where at least three of the five years were valid ($Y_{\mathrm{valid}} = \sum_t O^{(t)} \geq 3$), cultivation frequency was

$$F(x) = 100\, \frac{\sum_{t=2020}^{2024} O^{(t)}(x)\, C^{(t)}(x)}{Y_{\mathrm{valid}}(x)}, \tag{2}$$

and $L_1 = 1, 2, 3, 4, 5$ for F in $[0, 25]$, $(25, 40]$, $(40, 60]$, $(60, 80]$ and $(80, 100]$. F is the share of adequately observed years in which the crop rule was met, that is, a measure of cultivation persistence. The thresholds define the proxy used here and are specific to this application.[23]

The layer was exported at 500 m (EPSG:4326) and harmonised to the analytical grid in ArcGIS. Because QA60 values are masked in this collection from early 2022 to early 2024, the joint cloud mask removed the observations of 2022 and 2023. Cultivation frequency was therefore calculated from the valid observation years 2020, 2021 and 2024, and the requirement $Y_{\text{valid}} \geq 3$ meant that all three years had to be valid; missing years were excluded from the denominator.

Urban-expansion pressure

Urban-expansion pressure L_2 represents the intensity of built-surface growth, so that places undergoing active land take and construction are separated from established settlements with little recent growth.[2] Built-surface values of the Global Human Settlement Layer (GHS-BUILT-S R2023A;[24] m^2 of built surface per 100 m cell) for $t_s = 1990$ and $t_e = 2020$ were combined into a regularised annualised growth index,

$$C(x) = \left[\left(\frac{b_e(x)}{b_s(x) + \epsilon} \right)^{1/30} - 1 \right] \times 100\%, \qquad \epsilon = 0.1 \text{ m}^2, \tag{3}$$

where ϵ keeps the index defined where the initial built surface is zero. $C(x)$ is an operational index of built-surface change. Levels were $L_2 = 1$ for $C \leq 0$ (no growth, decline, or no built surface in 2020), 2 for $0 < C \leq 1\%$, 3 for $1 < C \leq 3\%$, 4 for $3 < C \leq 5\%$ and 5 for $C > 5\%$. The layer was computed at 100 m in Google Earth Engine and harmonised to the analytical grid after clipping to the study area. Equal positive built-surface values in both years yield a slightly negative regularised index and are included in level 1.

Soil-erosion pressure

Soil-erosion pressure L_3 was derived from the 2012 soil-loss estimates of version 1.1 of the Global Soil Erosion Modelling platform (GloSEM), which implements the Revised Universal Soil Loss Equation $A = R \times K \times LS \times C_{\text{f}} \times P_{\text{f}}$ with rainfall erosivity R, soil erodibility K, slope length and steepness LS, cover-management factor C_{f} and support-practice factor P_{f}.[26] Annual soil loss (t ha^{-1} yr^{-1}) was used as a proxy for geomorphic instability, which truncates stratigraphy and exposes buried remains more gradually than urban expansion. The input was the publicly distributed 2012 GloSEM v1.1 soil-loss raster at approximately 25 km resolution, derived from an underlying model originally produced at approximately 250 m resolution. To ensure spatial consistency with the other pressure indicators, the 25 km raster was resampled to the common 603.75 m analytical grid in ArcGIS and classified into five equal-interval levels over its valid range. The effective spatial resolution of the erosion information therefore remains that of the 25 km source raster.

Composite pressure and integration

The three levels were combined with equal weights and truncated to an integer, equivalent to the `Int` operation used in ArcGIS:

$$P(x) = \left\lfloor \frac{L_1(x) + L_2(x) + L_3(x)}{3} \right\rfloor, \qquad P(x) \in \{1, \ldots, 5\}. \tag{4}$$

Under this rule, $P \geq 4$ requires the three levels to sum to at least 12, which requires at least two pressures at level 4 or above; a single very high pressure is damped. Equal weights were chosen for transparency and comparability. The framework allows weights to be adapted to regional conditions, site types or management objectives. Occurrence and pressure were integrated with a conservative matrix rule,

$$R(x) = \min\big(S(x), P(x)\big), \tag{5}$$

so that the risk-priority level $R(x) \in \{1, \ldots, 5\}$ is high only where both occurrence and pressure are high. Strong pressure on a landscape of low occurrence likelihood, or high likelihood under weak present pressure, both yield a low level. R is a strategic priority surface for survey and preventive conservation.

Spatial harmonisation, coverage and area statistics

The agricultural and urban-expansion pressure indicators were computed in Google Earth Engine, while the GloSEM soil-loss raster was obtained as an existing 25 km product. Occurrence models were fitted on the 100 m predictor grid. The classified occurrence surface and the pressure layers were then harmonised in ArcGIS to a common grid of 603.75 m (1,156 × 790 cells, WGS 84 / UTM zone 41N) that shares its origin with the predictor grid, and this grid was used for classification, integration, area statistics and mapping. Areas are planimetric areas on this grid (0.3645 km^2 per cell). Tajikistan lies 4-12° east of the central meridian of the zone, so grid areas exceed ellipsoidal areas by 0.3-2.8%: by 1.2% for the national territory (142,121 km^2 ellipsoidal) and 0.9% for the analysed area. The mask-level area summary is included in the re-analysis package (Data availability). For the coverage analysis, the national domain comprises the cells whose centre falls on a valid national elevation cell, and regional membership was taken from the regional elevation rasters. The analysed area comprises cells where P is defined and all eight predictors are valid at the cell centre. Records were located by their coordinates. The re-analysis was carried out in Python (NumPy, SciPy, pandas) with MaxEnt 3.4.4.

Data availability

The modelling inputs (national and regional predictor rasters and site-location files), the classified occurrence, pressure and risk-priority rasters, the re-analysis package (scripts, MaxEnt model files, result tables and validation logs) and the source code are available in the Zenodo record of the project (`https://doi.org/10.5281/zenodo.23879254`). The modelling inputs and source code are also provided in release v1.0.0 of the project repository (`https://github.com/sq-m`

odelling/tajikistan-archaeological-risk-mapping). The CAAL database is described at https://uclcaal.org/database/. Its core is openly accessible without registration at map scales up to approximately 1:320,000, and records classed as sensitive are available to registered researchers on request to CAAL (caal@ucl.ac.uk), with a statement of the data required, the research purpose and the academic affiliation. Sentinel-2, GHSL and GloSEM data are available from their providers.

References

[1] Plekhov, D., Gosner, L. R. & Nowlin, J. Effects of land-use intensity on archaeological survey. *Advances in Archaeological Practice* **13**, 496–513 (2025).

[2] Agapiou, A. *et al.* Impact of urban sprawl to cultural heritage monuments: The case study of Paphos area in Cyprus. *Journal of Cultural Heritage* **16**, 671–680 (2015).

[3] Nebbia, M., Cilio, F. & Bobomulloev, B. Spatial risk assessment and the protection of cultural heritage in southern Tajikistan. *Journal of Cultural Heritage* **49**, 183–196 (2021).

[4] Sdykov, M. Historical and Archaeological Research of the Historical and Cultural Heritage of the West Kazakhstan Region. [Istoricheskie i arkheologicheskie issledovaniia istoriko-kul'turnogo naslediia Zapadno-Kazakhstanskoi oblasti]. *Ülken Altai älemi – Mir Bol'shogo Altaia – World of Great Altay* 1075–1096 (2020).

[5] Kharkovenko, R. History of the development of registration documentation for objects of cultural heritage in Ukraine. *International Journal of Legal Studies (IJOLS)* **2**, 327–342 (2017). Available at https://ijols.com/gicid/01.3001.0012.2257.

[6] Zaidner, Y. & Kurbanov, S. Soii Havzak: a new Palaeolithic sequence in Zeravshan valley, central Tajikistan. *Antiquity* **98**, e31 (2024).

[7] Shnaider, S., Taylor, W. T., Abdykanova, A., Kolobova, K. & Krivoshapkin, A. Evidence for early human occupation at high altitudes in western Central Asia: the Alay site. *Antiquity* **92**, e1 (2018).

[8] Boxleitner, K. *et al.* Archaeobotanical investigations at high-elevation sites of the Pamir mountains and Fergana foothills. *iScience* **29**, 114349 (2026).

[9] Lindsay, I. & Mkrtchyan, A. Free and low-cost aerial remote sensing in archaeology: An overview of data sources and recent applications in the south caucasus. *Advances in Archaeological Practice* **11**, 164–183 (2023).

[10] De Laet, V., van Loon, G., Van der Perre, A., Deliever, I. & Willems, H. Integrated remote sensing investigations of ancient quarries and road systems in the Greater Dayr al-Barshā region, middle Egypt: a study of logistics. *Journal of Archaeological Science* **55**, 286–300 (2015).

[11] Verhagen, P. & Whitley, T. G. Integrating archaeological theory and predictive modeling: a live report from the scene. *Journal of Archaeological Method and Theory* **19**, 49–100 (2012).

[12] Wachtel, I., Zidon, R., Garti, S. & Shelach-Lavi, G. Predictive modeling for archaeological site locations: Comparing logistic regression and maximal entropy in north israel and north-east china. *Journal of Archaeological Science* **92**, 28–36 (2018).

[13] Yaworsky, P. M., Vernon, K. B., Spangler, J. D., Brewer, S. C. & Codding, B. F. Advancing predictive modeling in archaeology: An evaluation of regression and machine learning methods on the grand staircase-escalante national monument. *PLOS ONE* **15**, 1–22 (2020).

[14] Noviello, M., Cafarelli, B., Calculli, C., Sarris, A. & Mairota, P. Investigating the distribution of archaeological sites: Multiparametric vs probability models and potentials for remote sensing data. *Applied Geography* **95**, 34–44 (2018).

[15] Phillips, S. J., Anderson, R. P. & Schapire, R. E. Maximum entropy modeling of species geographic distributions. *Ecological Modelling* **190**, 231–259 (2006).

[16] Lobo, J. M., Jiménez-Valverde, A. & Real, R. AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography* **17**, 145–151 (2008).

[17] Merow, C., Smith, M. J. & Silander, J. A. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography* **36**, 1058–1069 (2013).

[18] Roberts, D. R. *et al.* Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40**, 913–929 (2017).

[19] Valavi, R., Elith, J., Lahoz-Monfort, J. J. & Guillera-Arroita, G. blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution* **10**, 225–232 (2019).

[20] Zhu, Z., Qiu, S. & Ye, S. Remote sensing of land change: A multifaceted perspective. *Remote Sensing of Environment* **282**, 113266 (2022).

[21] Hansen, M. C. & Loveland, T. R. A review of large area monitoring of land cover change using Landsat data. *Remote Sensing of Environment* **122**, 66–74 (2012). Landsat Legacy Special Issue.

[22] Xu, S. & Fina, S. National-scale imperviousness mapping and detection of urban land changes. *ISPRS Journal of Photogrammetry and Remote Sensing* **202**, 369–384 (2023).

[23] Recuero, L. *et al.* Fallowing temporal patterns assessment in rainfed agricultural areas based on NDVI time series autocorrelation values. *International Journal of Applied Earth Observation and Geoinformation* **82**, 101890 (2019).

[24] Pesaresi, M. & Politis, P. GHS-BUILT-S R2023A – GHS built-up surface grid, derived from Sentinel2 composite and Landsat, multitemporal (1975–2030). European Commission, Joint Research Centre (JRC) (2023).

[25] Liu, Z., Huang, S., Fang, C., Guan, L. & Liu, M. Global urban and rural settlement dataset from 2000 to 2020. *Scientific Data* **11**, 1359 (2024).

[26] Borrelli, P. *et al.* An assessment of the global impact of 21st century land use change on soil erosion. *Nature Communications* **8**, 2013 (2017).

[27] Wu, M. *et al.* Anthropogenic influencing mechanisms on earthen sites preservation under the spatial-temporal variation of land utilization. *npj Heritage Science* **13**, 293 (2025).

[28] Barbet-Massin, M., Jiguet, F., Albert, C. H. & Thuiller, W. Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution* **3**, 327–338 (2012).

[29] TGO News Service. Tajikistan launches digital platform to preserve cultural heritage and historic monuments. The Gulf Observer (2026). Available at `https://thegulfobserver.com/tajikistan-launches-digital-platform-to-preserve-cultural-heritage-and-historic-monuments/`. Accessed 26 July 2026.

[30] Sargent, L. Documenting cultural landscapes. In Brown, S. & Goetcheus, C. (eds.) *Routledge Handbook of Cultural Landscape Practice*, 216–230 (Routledge, 2023).

[31] Brown, S. Documenting pastoral landscapes. *Heritage & Society* **4**, 33–58 (2011).

[32] Historic England. Historic landscape characterisation: a tool for understanding and managing whole areas (n.d.). Available at `https://historicengland.org.uk/research/methods/characterisation/historic-landscape-characterisation/`. Accessed 26 July 2026.

[33] World Bank. Building resilience across Tajikistan for a more inclusive economy (2024). Available at `https://www.worldbank.org/en/results/2024/04/05/building-resilience-across-tajikistan-for-a-more-inclusive-economy`. Accessed 26 July 2026.

[34] Rijabekov, N. C. & Sultonbekova, Z. V. Development of tourism along the Great Silk Road of the Republic of Tajikistan. International University of Tourism and Entrepreneurship of Tajikistan (2025). Available at `https://iutet.tj/en/development-of-tourism-along-the-great-silk-road-of-the-republic-of-tajikistan/`. Accessed 26 July 2026.

[35] Aliev, N. N. *et al.* Assessment of agro-ecological condition of the territory of the Republic of Tajikistan. *Research Journal of Agricultural Science* **53**, 3–9 (2021).

[36] Aliev, N. N. Monitoring of irrigated lands of the Hissar valley of the Republic of Tajikistan. In Sultan, M., Imran, M. & Ahmad, F. (eds.) *Irrigation Systems and Applications* (IntechOpen, 2024).

[37] Viviroli, D. & Weingartner, R. The hydrological significance of mountains: From regional to global scale. *Hydrology and Earth System Sciences* **8**, 1017–1030 (2004).

[38] Vanselow, K. A., Zandler, H. & Samimi, C. Time series analysis of land cover change in dry mountains: Insights from the Tajik Pamirs. *Remote Sensing* **13**, 3951 (2021).

[39] Lerman, Z. Tajikistan: An overview of land and farm structure reforms. Discussion Paper 2.08, Hebrew University of Jerusalem, Department of Agricultural Economics and Management (2008).

[40] Breckle, S.-W. & Wucherer, W. Vegetation of the Pamir (Tajikistan): Land use and desertification problems. In Spehn, E. M., Liberman, M. & Körner, C. (eds.) *Land Use Change and Mountain Biodiversity*, 225–237 (CRC Press, Boca Raton, FL, 2006).

[41] Espa, G., Benedetti, R., De Meo, A., Ricci, U. & Espa, S. GIS based models and estimation methods for the probability of archaeological site location. *Journal of Cultural Heritage* **7**, 147–155 (2006).

[42] Nsanziyera, A. F., Rhinane, H., Oujaa, A. & Mubea, K. GIS and remote-sensing application in archaeological site mapping in the Awsard area (Morocco). *Geosciences* **8**, 207 (2018).

[43] Böhner, J. & Selige, T. Spatial prediction of soil attributes using terrain analysis and climate regionalisation. In *SAGA-Analyses and modelling applications*, 13–28 (Goltze, 2006).

[44] Phillips, S. J. & Dudík, M. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography* **31**, 161–175 (2008).

[45] Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E. & Blair, M. E. Opening the black box: an open-source release of Maxent. *Ecography* **40**, 887–893 (2017).

[46] Gorelick, N. *et al.* Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment* **202**, 18–27 (2017).

Acknowledgements

This work was supported by the Shanghai Jiao Tong University-University College London (SJTU-UCL) Joint Research Grant (2023-2024).

Author contributions

C.L. developed the methodology and software, performed the analyses and visualisations. C.L., M.N. and Q.S. wrote the original draft. M.N. conceived the study and curated the CAAL data. C.L., M.N. and Q.S. performed formal analysis and investigation; F.C., P.C., R.P. and G.J. curated the archaeological records. T.D.W. conceived and directed the CAAL project. H.X. and Q.S. conceived and supervised the study and acquired funding. All authors reviewed and edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information accompanies this paper: Supplementary Tables S1-S7 and Supplementary Figs. S1-S3.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SI.pdf