nature portfolio

Corresponding author(s): Yafei Wang & Murray Scown

Last updated by author(s): Jul 17, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Sample points used for predicting land subsidence rates were extracted from figures in previously published, peer-reviewed journal articles using WebPlotDigitizer v4.6 (https://automeris.io/WebPlotDigitizer), a validated tool for digitizing graphical data. The original sources from which data were extracted are properly cited in the Supplementary Data 2. For the logistic regression modeling of land suitability, sample points were generated and processed using CLUMondo v1.4.0, a spatially explicit land-use allocation model. The configuration files, model parameters, and derived datasets used in the suitability analysis are available via Figshare at https://doi.org/10.6084/m9.figshare.29263130.
Data analysis	The CLUMondo model is available on GitHub (https://github.com/VUEG/CLUMondo). Python scripts used for projecting land subsidence and generating figures can be accessed via Figshare at https://doi.org/10.6084/m9.figshare.29263130. The improved bathtub model, which incorporates hydrological connectivity and attenuation, is available on GitHub (https://github.com/geoye/attenuated_bathtub). ArcGIS Pro is used for analyzing spatial data and creating maps. Additional code supporting the findings of this study is available from the corresponding author upon reasonable request.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The projected land system maps for five policy scenarios, along with associated validation datasets and sampling points used for projecting land subsidence, are available on Figshare: https://doi.org/10.6084/m9.figshare.29263130. Input datasets for CLUMondo simulations are cited throughout the manuscript, with full documentation provided in Supplementary Note 4 and Supplementary Table S4. Sea-level rise (SLR) projections were obtained from the IPCC AR6 database (Zenodo: https://zenodo.org/records/6382554), and tide and surge data were sourced from the CoDEC dataset (Zenodo: https://zenodo.org/records/3660927). Coastal elevation data (CoastalDEM) were acquired from Climate Central (https://go.climatecentral.org), and mean dynamic topography (MDT) data were obtained from AVISO (https://doi.org/10.24400/527896/a01-2023.003). Sources for the explanatory variables used in modeling land subsidence are detailed in Supplementary Table S7. All source data supporting the findings of this study are provided with the paper or are publicly available from the cited repositories.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This study addresses three key applications of sampling techniques in spatial analysis: (1) Integrating sampling with machine learning to improve long-term land use classification accuracy using remote sensing and historical land-use data; (2) Combining sampling and machine learning approaches to model spatially explicit land subsidence rates, based on observed subsidence points and explanatory environmental variables; (3) Applying spatial sampling methods to assess local land-use suitability within a logistic regression framework in the CLUMondo land system model, enabling policy-driven scenario projections. Sampling strategies in each application were tailored to the spatial resolution, data availability, and modeling requirements, and are described in detail in the Methods and Supplementary Information.
Research sample	The study employed distinct sampling strategies for three core components: (1) Land use classification: A total of 4,700 sample points were manually identified to document annual land cover types across China's coastal zone from remote sensing imagery. The sampling ensured temporal and spatial representativeness across diverse land systems. (2) Land subsidence modeling: A total of 6,203 sample points were compiled from published sources and field records, comprising 4,417 subsidence points and 1,786 non-subsidence points. These were used to train and validate machine learning models predicting spatial patterns of subsidence. (3) Land suitability analysis (CLUMondo): For logistic regression modeling of land system suitability, a random sample of 30% of grid cells within the study region was used to reduce computational burden while maintaining statistical power and geographic coverage. Full details on the spatial resolution, source data, and sampling rationale are provided in the Methods section and Supplementary Information.
Sampling strategy	(1) Land use classification: 80% of the samples were used to train a Random Forest classifier, while the remaining 20% were reserved for accuracy assessment. To mitigate sampling bias, random iterations were performed 20 times, and the iteration with the highest

accuracy was selected for the final prediction.(2) Land subsidence rate modeling: A nested cross-validation (CV) approach was applied. In the inner loop, a 5-fold GridSearchCV (from Python's scikit-learn package) was used to optimize hyperparameters and minimize overfitting. In the outer loop, a 10-fold random CV provided an unbiased estimate of the model's generalization performance. Each Random Forest regressor was executed 20 times with varying training, testing, and validation sample distributions. (3) Suitability regression in CLUMondo: A stratified sample comprising 30% of all cells in the study region was used for logistic regression analysis to quantify relationships between specific land system types and explanatory factors. To reduce spatial autocorrelation, a minimum separation of 2 km (2 pixels) was enforced between sampled cells. The logistic regression dataset maintained a balanced sample, including an equal number of presence and absence points for each land use type.

Data collection

Land use classification data were collected from two primary sources: (1) Field surveys conducted in ten Chinese coastal provinces (Liaoning, Tianjin, Hebei, Shandong, Jiangsu, Shanghai, Zhejiang, Fujian, Guangdong, and Guangxi); and (2) Visual interpretation of high-resolution historical imagery from Google Earth. These sample points (n = 4,700) were used to train a Random Forest classifier in Google Earth Engine for generating annual land use maps.Land subsidence data (n = 6,203) were digitized from 40 peer-reviewed journal articles and international datasets using WebPlotDigitizer (v4.6). The extracted data were georeferenced in ArcMap 10.8 based on geographic coordinates or map overlays provided in the original publications.Land suitability data for use in the CLUMondo model were derived by randomly sampling 30% of grid cells across the study region. Environmental and socio-economic input layers used in the logistic regression were compiled from authoritative, publicly available datasets, with full source information provided in Supplementary Table S4.

Timing and spatial scale

Temporal scope: The study includes both historical and future projections. Historical land use classification was conducted using Landsat satellite imagery spanning 1990–2020. Future land system changes were projected at decadal intervals from 2020 to 2100 under three Shared Socioeconomic Pathways (SSP126, SSP245, and SSP585) and five additional policy-driven scenarios. Spatial scope: The spatial extent of the study includes all coastal prefecture-level cities and municipalities in mainland China, including Hong Kong, Macao, and Hainan Province.Land system simulations were performed at a spatial resolution of 1 km^2, while coastal flooding and inundation analyses were conducted at a finer resolution of 1 ha (100 m × 100 m) to capture localized hydrological impacts.

Data exclusions

Pixels labeled as 'Nodata' were excluded from all spatial analyses.In land subsidence modeling, the following exclusions and adjustments were applied to ensure data quality and minimize spatial or statistical bias: (1) Non-subsiding locations (subsidence rates < 0 mm/year) were assigned a value of 0 mm/year to avoid introducing negative values into the regression model. (2) Duplicate or overlapping sampling points within 1 km grid cells were removed to reduce spatial redundancy and avoid overrepresentation of specific locations. (3) The number of zero-subsidence points was downsampled to address class imbalance and prevent model bias toward stable (non-subsiding) areas.

Reproducibility

To ensure reproducibility across all analyses: (1) Land system classification: The seed parameter was set in the Random Forest classifier within Google Earth Engine to control data splits and ensure consistent classification results across repeated runs. (2) Land subsidence rate modeling: The random_state parameter was applied during cross-validation and regression procedures in Python (scikit-learn) to maintain reproducibility of results. (3) Suitability regression in CLUMondo: The randomly selected training samples used in logistic regression were saved to disk within CLUMondo v1.4.0 to ensure consistency across model calibration and future re-runs. (4) Land system model calibration and validation: The CLUMondo model was calibrated using logistic regression on historical land system data and validated by comparing simulated outcomes with observed land system maps from 2010 to 2020. The model achieved an overall accuracy of 72.18% and a Figure of Merit (FoM) of 0.278, demonstrating good agreement with real-world land system changes.

Randomization

(1) Land use classification: Training data were randomly sampled 20 times to minimize the influence of sample distribution on classification outcomes and to ensure the robustness of the Random Forest classifier implemented in Google Earth Engine. (2) Land subsidence rate modeling: A nested cross-validation procedure with 20 repeated experiments was used in the machine learning workflow to generate unbiased, generalizable predictions and reduce overfitting. (3) Suitability regression in CLUMondo: A random selection of 30% of grid cells across the study region was used to fit logistic regression models quantifying relationships between land system types and explanatory variables. To further reduce spatial sampling bias, a minimum distance of 2 km between samples was enforced, and balanced sampling across land system types was applied to avoid overrepresentation of dominant classes.

Blinding

This study used fully objective, data-driven methods with no subjective assessments; blinding was not applicable.

Did the study involve field work? ☒ Yes ☐ No

Field work, collection and transport

Field conditions

Fieldwork was conducted in the coastal zones of China, regions characterized by diverse environmental conditions. These areas are influenced by various climatic factors, such as monsoonal weather patterns and occasional extreme events like typhoons. The coastal zones also experience significant human activity, including urbanization, industrial development, and agricultural or aquacultural practices. During the fieldwork, sample points for land use classification were collected through photography to document land cover types, as well as interviews to gather information on land use changes and local perceptions of coastal flooding risks.

Location

Fieldwork was conducted across ten coastal provinces and municipalities in China: Liaoning, Tianjin, Hebei, Shandong, Jiangsu, Shanghai, Zhejiang, Fujian, Guangdong, and Guangxi. These regions were selected due to their significant coastal land system transformations and high exposure to sea-level rise and coastal flooding, offering a representative gradient of coastal environmental and socio-economic conditions from north to south.Field sampling and semi-structured interviews were conducted in specific locations within these areas, including but not limited to: (1) Yingkou and Panjin (Liaoning); (2) Tianjin Municipality; (3) Caofeidian (Tangshan) and Huanghua (Cangzhou), Hebei; (4) Qingdao (Shandong); (5) Ganyu (Lianyungang) and Dafeng (Yancheng), Jiangsu; (6) Zhenhai (Ningbo) and Wenling (Taizhou), Zhejiang; (7) Xiamen (Fujian); (8) Shantou and Guangzhou (Guangdong).These sites cover diverse coastal geomorphologies, land use pressures, and development patterns, forming a robust empirical foundation for the study.

Access & import/export	No restricted or sensitive areas were accessed, and no materials subject to import/export controls were involved.
Disturbance	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	☐	Antibodies
☒	☐	Eukaryotic cell lines
☒	☐	Palaeontology and archaeology
☒	☐	Animals and other organisms
☒	☐	Clinical data
☒	☐	Dual use research of concern
☒	☐	Plants

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☒	☐	Flow cytometry
☒	☐	MRI-based neuroimaging

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A