

Filling subnational sanitation data gaps in low- and middle-income countries

Esther Greenwood

Esther.Greenwood@eawag.ch

Eawag, Swiss Federal Institute of Aquatic Science and Technology <https://orcid.org/0000-0001-6376-3134>

Nour El-Ajou

Eawag, Swiss Federal Institute of Aquatic Science and Technology

Andreas Scheidegger

Eawag Swiss Federal Institute of Aquatic Science and Technology <https://orcid.org/0000-0003-2575-2172>

Johan van den Hoogen

Department of Environmental Systems Science, ETH Zurich, Swiss Federal Institute of Technology

Richard Johnston

WHO

Ayca Donmez

UNICEF Office of Strategy and Evidence – Innocenti

Linda Strande

Eawag, Swiss Federal Institute of Aquatic Science and Technology <https://orcid.org/0000-0003-4477-6268>

Qiao Wang

The World Bank

Yi Hoo

The World Bank

Timothy Julian

Eawag, Swiss Federal Institute of Aquatic Science and Technology <https://orcid.org/0000-0003-1000-0306>

George Joseph

The World Bank

Karin Gallandat

Eawag, Swiss Federal Institute of Aquatic Science and Technology

Keywords:

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11485081/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: **Yes** there is potential Competing Interest. QW, YRH, and GJ are employees of The World Bank. Part of this work was conducted under a consultancy contract funded by The World Bank. The authors declare no other competing interests.

Filling subnational sanitation data gaps in low- and middle-income countries

Authors:

Esther E. Greenwood^{1*}, Nour El-Ajou¹, Andreas Scheidegger¹, Johan van den Hoogen²,
Richard Johnston⁴, Ayca Donmez⁵, Linda Strande¹, Qiao Wang⁶, Yi Rong Hoo⁶, Timothy R.
Julian^{1,7,8}, George Joseph⁶, Karin Gallandat¹

Affiliations:

¹ Eawag, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland

² Department of Environmental Systems Science, ETH Zurich, Swiss Federal Institute of
Technology, Zurich, Switzerland

⁴ Department of Environment, Climate Change, One Health and Migration, World Health
Organization, Geneva, Switzerland

⁵ UNICEF Office of Strategy and Evidence – Innocenti, Florence, Italy

⁶ World Bank, Washington, District of Columbia, United States of America

⁷ University of Basel, Basel, Switzerland

⁸ Swiss Tropical and Public Health Institute, Allschwil, Switzerland

*Corresponding author: Esther E. Greenwood (esther.greenwood@eawag.ch)

Abstract

Access to safe sanitation is a human right, yet as of 2024, close to 1.5 billion people lacked access to basic sanitation services globally. Here, we address subnational data gaps, applying a Tabular Prior-Fitted Networks model that leverages subnational survey estimates derived from 1.8 million households to predict basic sanitation and open defecation at first-level administrative units in 130 low- and middle-income countries worldwide. Our results highlight inequalities between country income groups, with median basic sanitation coverage ranging from 36% (90% prediction interval=17–56%) in low-income to 94% (76–99%) in upper-middle-income countries. Open defecation shows the reverse gradient (8% [1–29%] to <1% [0–1%]). Fragile, low-income sub-Saharan African settings had the poorest sanitation outcomes, highest uncertainties and greatest data gaps. Our work illustrates a transparent approach to filling data gaps, provides clear indications of uncertainties and highlights the vulnerability of low-income and crisis-affected populations.

Introduction

Access to safe sanitation services is a human right and essential for public health, yet progress towards global sanitation targets has been lagging (1). United Nations' (UN) member states have committed to achieving adequate and equitable sanitation and hygiene for all and ending open defecation by 2030 as reflected in Sustainable Development target 6.2 (2). Insufficient investment, underutilisation of available financial resources, and persistent inequalities in public spending contribute to many UN member states being off track (3). In 2024, it was estimated that almost 1.5 billion people still lacked access to basic sanitation, including 354 million who practiced open defecation (1). To achieve universal access to basic sanitation services for all by 2030, a threefold increase in the speed of progress from 2024 would be needed compared to the progress observed between 2015-2024.

Over the past decades, geographic targeting has been increasingly applied to enhance the efficiency of development programmes by directing resources to areas of greatest need (4, 5). These approaches require spatially resolved information on development outcomes. In the sanitation sector, the lack of up-to-date, granular information on service coverage is a major bottleneck in progress monitoring and in identifying priority needs.

The WHO/UNICEF Joint Monitoring Programme for Water Supply, Sanitation and Hygiene (JMP) is the custodian agency for Sustainable Development Goal (SDG) target 6.2. As such, the JMP maintains the largest database on household sanitation services globally harmonizing nationally representative data from over 190 countries (1). Progress is tracked using a sanitation service ladder that classifies household sanitation into five ordinal service tiers: open defecation (as the lowest tier), unimproved facilities, followed by limited, basic, and safely managed sanitation services (6). In alignment with the SDG ambition to “leave no one behind”, the JMP has disaggregated data on sanitation indicators for approximately 100 low- and middle-income countries (LMICs) as of 2026 (7, 8) based on cross-sectional household surveys. As the primary data source underpinning these estimates, such surveys are time- and resource-intensive and have limited ability to provide updated subnational estimates for several LMICs (1, 9). While JMP efforts to update and map subnational sanitation estimates continue, there remain immediate challenges with uneven data coverage at subnational levels.

Geospatial modelling offers data-driven approaches to address these gaps, and its application to water and sanitation has expanded alongside the increased availability of satellite-derived data products (10–13). Satellite-derived products capturing built-up area extent, nighttime

light intensity, population sums and environmental factors have been shown to explain substantial spatial variation in drinking water and other infrastructure outcomes (13–15). Traditional machine learning models such as random forest are increasingly applied for survey data gap filling yet face practical challenges on small datasets as they are susceptible to overfitting and may generalise poorly for out-of-sample predictions (13, 16).

Recent advances in machine learning include deep learning models pre-trained on large amounts of synthetic data that can be applied to new prediction tasks with limited labelled data. Tabular Prior-Fitted Networks (TabPFN) are an example (17). TabPFN has demonstrated a stronger predictive performance when using datasets of up to 10,000 samples and 500 predictors in benchmarks compared to traditional machine learning approaches (17). A further advantage is that the probabilistic outputs from TabPFN enable direct quantification of prediction intervals, allowing estimates to be accompanied by maps that reflect the spatial distribution of model uncertainties.

In this study, we predict subnational sanitation coverage at administrative level 1 across 130 LMICs using sanitation service definitions from the JMP service ladder framework (6). We model the proportion of the population using at least basic sanitation services, defined as use of improved facilities not shared with other households, and the proportion practicing open defecation, which involves disposal of human faeces in open spaces such as fields, forests, or open bodies of water (6). We focus on these outcomes because they represent key milestones in achieving SDG target 6.2 and can be reliably measured from household survey data. We combine data from the most recent Demographic and Health Surveys (DHS) and Multiple Indicator Cluster Surveys (MICS) gathered between 2015 and 2024 from 1.8 million households across 1205 administrative areas and 89 countries (Fig. 1, Data S1) with geospatial covariates representing environmental and socio-economic conditions at national and subnational levels (Tables S1 and S2, Data S2). We compare TabPFN and traditional random forest models on their ability to extrapolate to countries outside the training sample using spatial cross-validation and select the best-performing model to predict subnational sanitation coverage in 2024 and map uncertainties. We further discuss the challenges of extending this approach to safely managed sanitation services and outline priorities for future data collection and modelling efforts. As such, we aim to provide a transparent, uncertainty-aware approach to filling subnational sanitation data gaps.

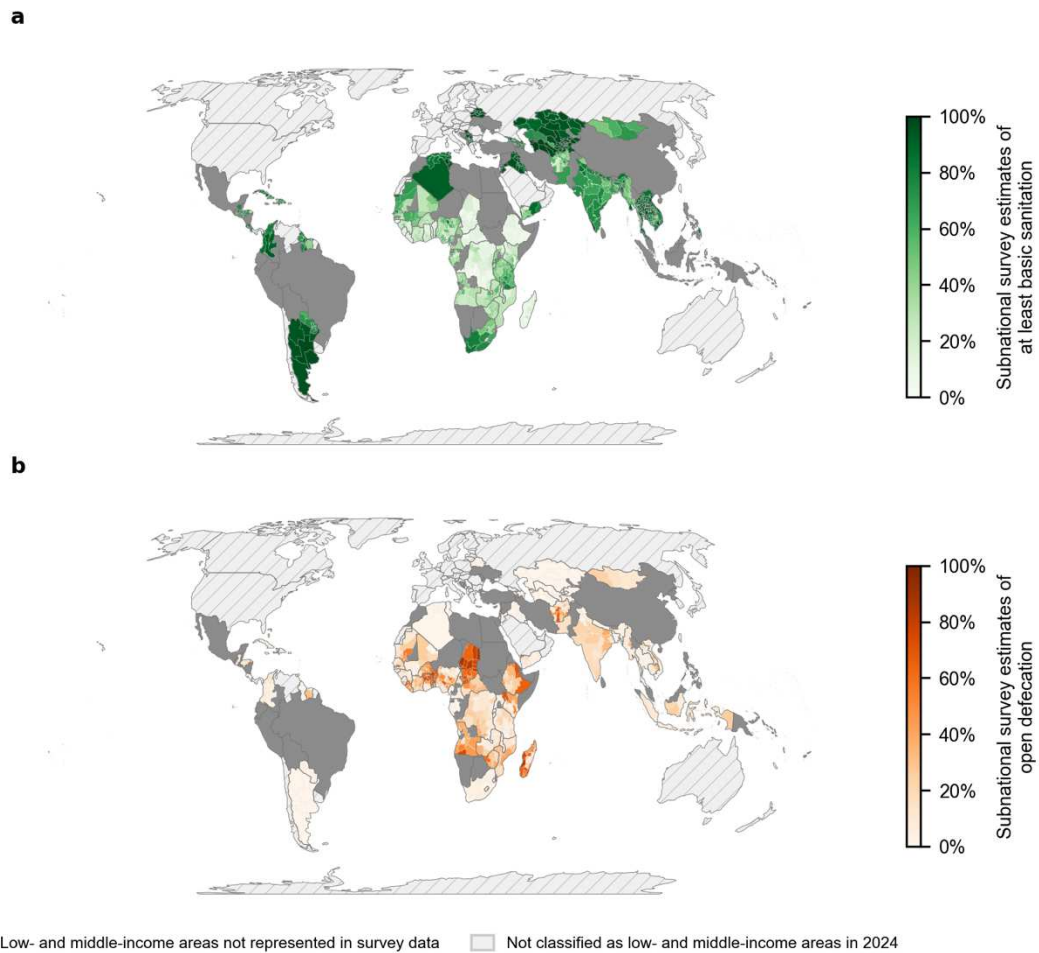


Figure 1. Subnational survey estimates of sanitation outcomes used for modelling (Data S1). Estimates are from Demographic and Health Surveys (DHS) and Multiple Indicator Cluster Surveys (MICS) across 88 countries for basic sanitation (a) and 89 countries for open defecation (b) gathered between 2015 and 2024. Values of subnational estimates represent the observed proportion of the population using at least basic sanitation (a) and practising open defecation (b). Dark grey areas represent low- and middle-income areas not included in the survey data; hatched areas were not classified as low- and middle-income in the calendar year of 2024 (18) .

Results

Spatial variation of subnational sanitation predictions

TabPFN model predictions of populations using at least basic sanitation services and practicing open defecation across 2,310 admin-1 units in 130 LMICs varied substantially within and between SDG regions, as well as across income groups and fragility status (Fig. 2a, 2c, Data S2).

The lowest basic sanitation coverage rates, highest open defecation rates and greatest within-region variability were predicted in sub-Saharan Africa and Oceania (Fig. 3a, 4a). Median predictions of at least basic sanitation ranged from 39% in sub-Saharan Africa (n=662 admin-1 units, absolute range of admin-1 values =7 – 97%) and 49% in Oceania (n=74, range=21–97%) to 97% in Europe and Northern America (n=223, range=86–99%). Median predictions from all other SDG regions lay between 89% and 95%, spanning 1,351 admin-1 units with individual predictions ranging from 36–99% (Fig. 2, Fig. 3). The highest median open defecation rates were 8% in sub-Saharan Africa (n=662, range=0–73%) and 1% in Oceania (n=74, range=0–46%), with near-zero median predictions in Northern Africa, Western Asia, Europe, and Northern America (Fig. 2, Fig. 4). These predictions were derived from the best-performing TabPFN models, which in leave-one-country-out cross-validation captured 78% and 55% of the subnational variation in basic sanitation and open defecation, respectively (Table 1).

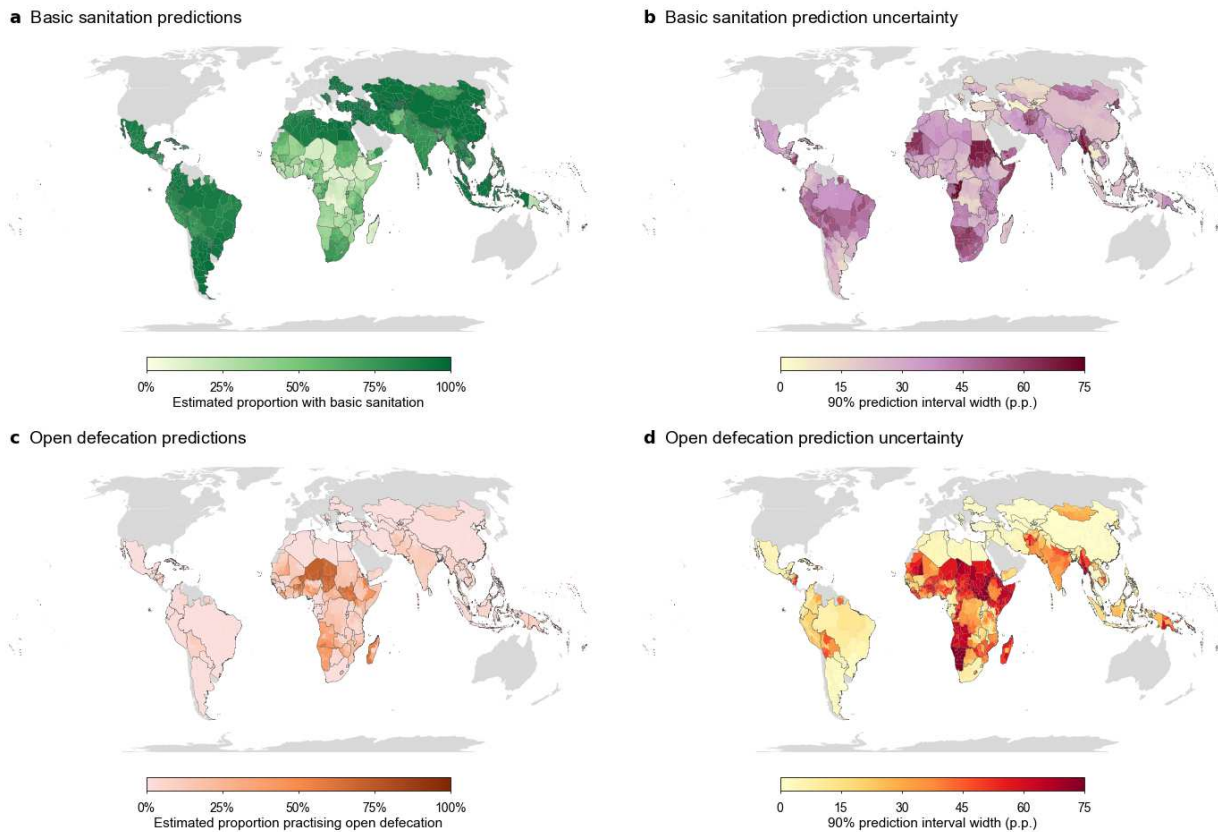


Figure 2. Maps of median prediction values for basic sanitation (a) and open defecation (c) and the corresponding 90% prediction interval width in percentage points (p.p.) (b, d) across 2,310 admin-1 units and 130 low- and middle-income countries (LMICs) for 2024 based on the best-performing Tabular Prior-Fitted Networks (TabPFN) model (Data S2).

Low-income countries and countries categorised as fragile or extremely fragile by the Organisation for Economic Cooperation and Development (OECD) in 2022, as provided in the WHO/UNICEF JMP 2025 world dataset (7, 19) showed the lowest prediction values for populations using at least basic sanitation services and highest predictions of populations practicing open defecation (Fig. 3a, Fig. 4a). Median basic sanitation coverage increased from 36% in low-income countries (n=406 regions, range=7–97%) to 78% in lower-middle-income (n=812, range=14–99%) and 94% in upper-middle-income countries (n=1,092, range=46–99%), with a corresponding decline in open defecation prevalence from 8% (range=0–73%) in low-income countries to 3% (range=0–67%) in lower-middle-income and less than 1% (range=0–11%)

in upper-middle-income countries. Fragile or extremely fragile contexts had substantially lower basic sanitation coverage (median=48%, range=7–99%) and higher open defecation prevalence (median=7%, range=0–73%) compared to other contexts (basic sanitation: median=93%, range=16–99%; open defecation: median < 1%, range=0–59%). The model outputs also suggested greater subnational variability in low- and lower-middle-income countries and fragile settings compared to upper-middle-income and non-fragile contexts (Fig. 3a, Fig. 4a).

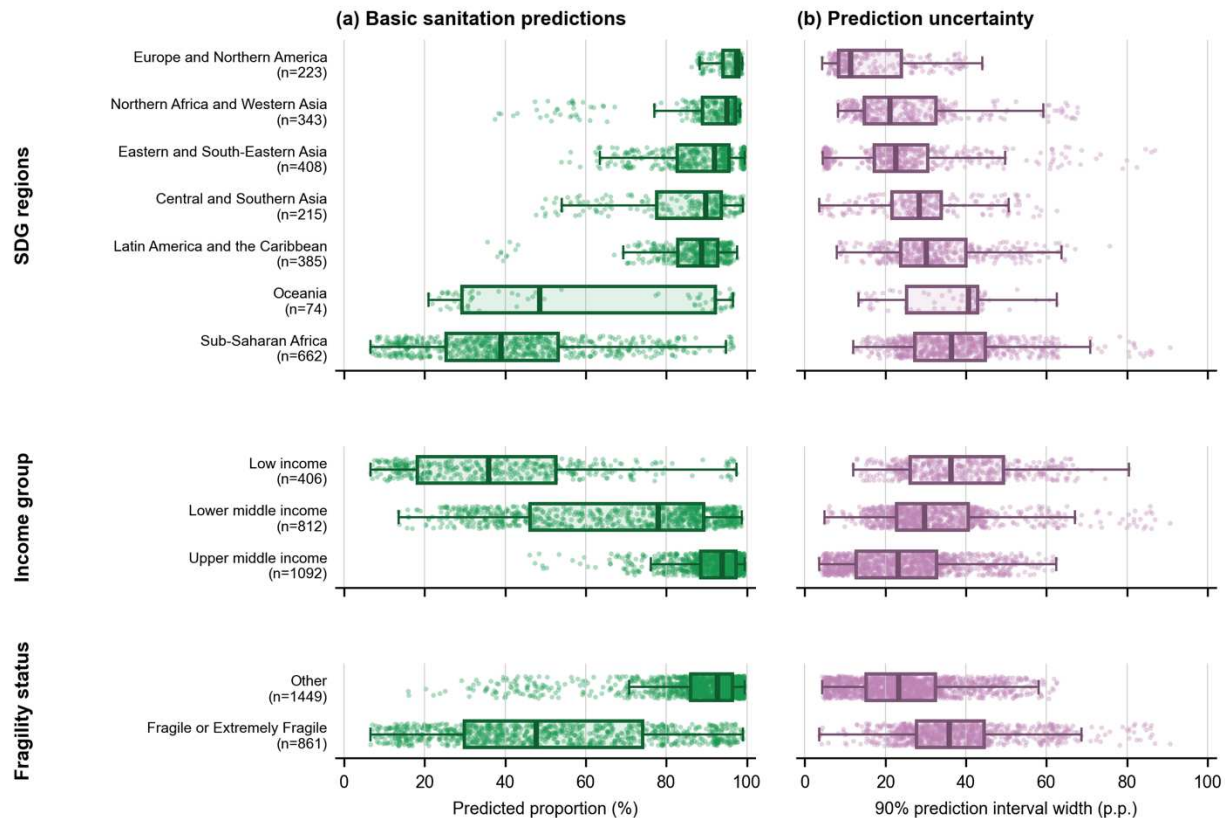


Figure 3. The distribution of (a) predictions of subnational populations using at least basic sanitation in admin-1 units (n=2,310, left) and (b) the width of corresponding 90% prediction intervals, in percentage points (p.p., right) across SDG regions, income groups and fragility status.

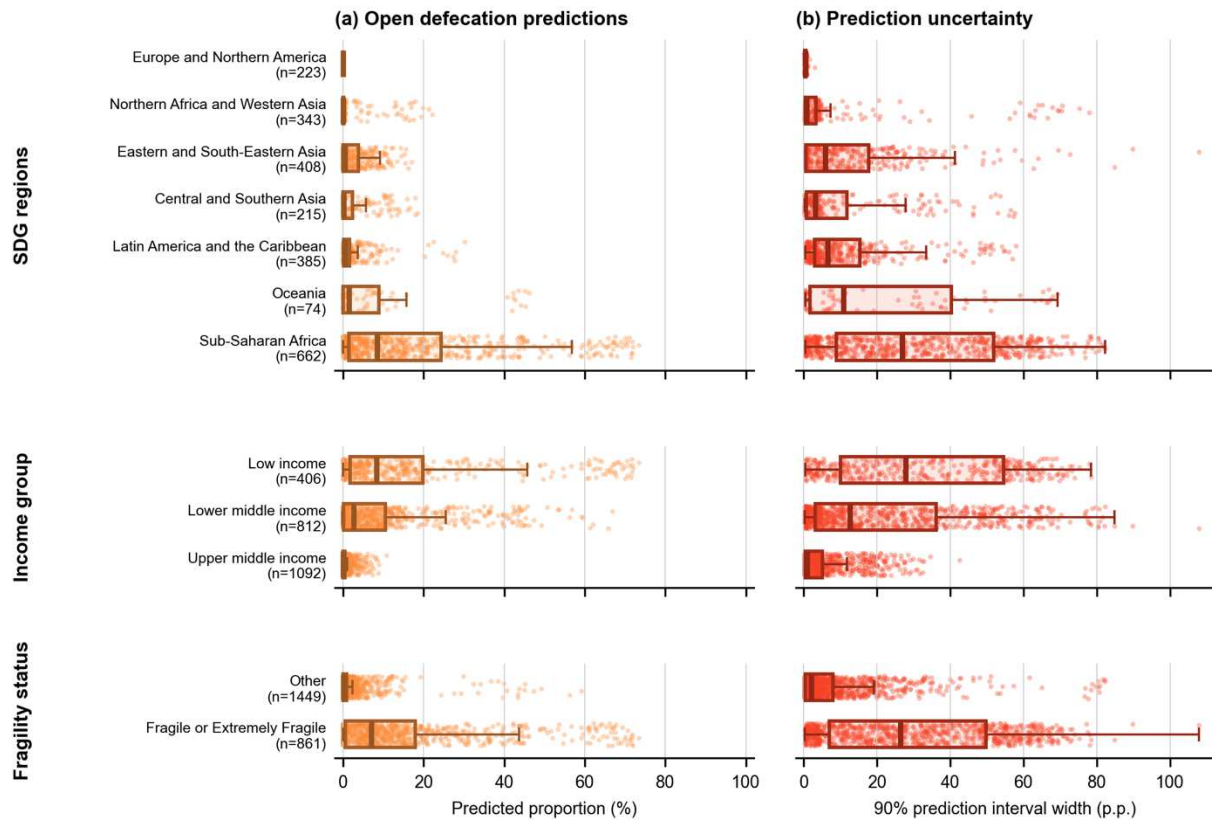


Figure 4. The distribution of (a) predictions of subnational populations practicing open defecation in admin-1 units (n=2,310, left) and (b) the width of corresponding 90% prediction intervals, in percentage points (p.p., right) across SDG regions, income groups and fragility status.

At the country level, the lowest predicted use of at least basic sanitation was in the Central African Republic, South Sudan, Chad, Guinea-Bissau and Madagascar, and the highest predicted open defecation in Niger, Chad, South Sudan, Togo and Benin (Fig. S1, Data S3). All of these are OECD-classified fragile settings, and all but Benin (lower-middle-income) are low-income countries. Where these countries had survey data, the raw estimates gathered between 2015 and 2024 confirmed their ranking (Data S1).

Subnational predictions revealed pronounced within-country inequalities in sanitation service use, especially in sub-Saharan Africa, echoing the patterns in the raw survey estimates (Fig. 1, Data S1) but at a smaller magnitude. Intraclass correlation coefficients (ICC) of raw survey data indicated that, although between-country variation exceeded within-country variation,

subnational differences remained substantial (ICC = 0.87 for basic sanitation, 0.61 for open defecation; median within-country range 32 percentage points for basic sanitation and 13 percentage points for open defecation). Predicted within-country range was greatest in Nigeria, Yemen, Tanzania, Mozambique and Malawi for basic sanitation, and in Niger, Chad, Benin, Angola and Madagascar for open defecation (Fig. S1, Data S3). Within country variability may be considerably influenced by the number of subdivisions within the country.

Model performance

The best-performing TabPFN models used 45 environmental and socio-economic predictors for basic sanitation and 90 for open defecation, capturing 78% (mean absolute error (MAE) = 0.08) and 55% (MAE=0.07) of the subnational spatial variation in basic sanitation and open defecation, respectively, based on leave-one-country-out cross-validation (Tables 1, S1 and S2, Fig. S2, S3). Model performance did not meaningfully decrease by adding additional predictors beyond 45 and 90, showing robustness to the inclusion of redundant features (Fig. S2).

TabPFN models performed as well as or better than the Distributed Random Forest models with moderate to large predictor subsets (30-100 predictors) (Fig. S2) (20). Model performance with small subsets of predictors (5-20 predictors) resulted in substantially lower performance and higher cross-fold variability for both models. The largest TabPFN versus random forest performance gains were observed for open defecation (coefficient of determination (R^2) = 0.55 vs. 0.45).

While residuals of median prediction values of the basic sanitation model were distributed evenly and remained close to zero, residuals of open defecation pointed towards a tendency to underestimate prevalence in areas where high rates of open defecation were reported (Fig. S4). For basic sanitation, the 90% prediction intervals were well calibrated and contained the observed survey-based estimate in 89% of cases. For open defecation, intervals were moderately overconfident, capturing the observed value in 82% of cases (Fig. S5). There is evidence that responses to household-level survey questions may underestimate the extent of open defecation in LMICs, which may partly explain lower model performance and potential underestimation of this outcome (21, 22).

To understand how the 2024 predictions differed from the original survey estimates gathered between 2015 and 2024, we compared the two at overlapping subnational areas (Fig. S6, Data S4). Survey and prediction-unit boundaries were closely aligned (median intersection-over-union, IoU = 0.96). Agreement between survey derived estimates and predictions was high

but decreased with survey age: MAE rose from 0.04 and 0.03 at recent surveys (2022–2024) to 0.10 and 0.05 at the oldest (2015–2018), for basic sanitation and open defecation respectively. Observed discrepancies were directional, with predicted basic sanitation in 2024 generally greater than observed values from earlier years, consistent with expected improvements in services since data collection. For open defecation, predictions were generally lower than observed, partly reflecting potential declines in open defecation since older surveys. However, a gap persisted even at recent surveys, indicating the model's underestimation of reported high open-defecation prevalence. Both basic sanitation and open defecation predictions mirrored which countries had the greatest within-country inequality but systematically underestimated the magnitude of within-country ranges (Fig. S7, Data S4).

At the country level, the raw survey estimates (2015–2024) confirmed the predicted rankings where data were available: the countries predicted to have the lowest basic sanitation and the highest open defecation were the same in the observed surveys (Data S1). Two surveyed countries were exceptions, with observed conditions more severe than predictions: Sierra Leone for basic sanitation (surveyed 2019) and Burkina Faso for open defecation (surveyed 2021), where median survey estimates were 12–13 percentage points worse than predictions across their admin-1 units. Trend analysis of historic survey data in Burkina Faso shows declines in open defecation of approximately 1.6 percentage points per year between 2003 and 2010 and 2.4 between 2010 and 2021 (23). These trends suggest that genuine progress since data collection may account for a substantial part of this discrepancy with the rest potentially due to residual model underestimation. More recent data would be required to confirm this balance.

Table 1 Leave-one-country-out cross-validation model performance metrics using the final feature set, selected based on the best model performance using a spatially stratified cross-validation approach

Outcome	Model	Observations (admin-1 units)	Countries	Predictors (n)	Mean absolute error (MAE)	Coefficient of determination (R ²)
Basic sanitation	Random Forest	1168	88	45	0.09	0.78
Basic sanitation	TabPFN	1168	88	45	0.08	0.78
Open defecation	Random Forest	1205	89	90	0.08	0.45
Open defecation	TabPFN	1205	89	90	0.07	0.55

Predictors and prediction uncertainty

Analysis of predictor importance, based on each predictor's mean absolute SHapley Additive exPlanation (SHAP) value, revealed complementary contributions from national and subnational covariates (Data S5). Aggregated by spatial scale, national predictors accounted for 54% (n=5) of the total predictor importance for basic sanitation and 41% (n=13) for open defecation, with the remaining contributions from subnational covariates (46% [n=40] for basic sanitation and 59% [n=77] for open defecation). National sanitation estimates were the single most important predictor group (50% for basic sanitation, 31% for open defecation), dominated by the national JMP sanitation estimates for 2024 (38% and 25%, respectively) and supported by additional national wastewater indicators (Fig. S8, Data S5). National JMP sanitation estimates are themselves derived by the JMP by fitting linear time trends to nationally representative household surveys and censuses, separately for urban and rural populations, and projecting to the target year (6). The model's subnational spatial structure predominantly came from indicators of human presence (20% of total predictor importance for basic sanitation and 24% for open defecation), terrain and vegetation (11% and 14%), climate (8% and 11%) and hydrology and soil (5% and 7%). The remaining importance was distributed across national socio-economic and governance factors, including GDP, education and governance quality (4% and 11%) and subnational region size (2% and 3%).

For both basic sanitation and open defecation, prediction uncertainty (90% prediction-interval (PI) width) was greatest in low-income and fragile sub-Saharan African settings (Fig. 2-4). In contrast, the proportion of admin-1 units with the highest certainty predictions (90% PI width below 25 percentage points) were in Europe and Northern America (including 76% of admin-1 units for basic sanitation and 100% for open defecation), Eastern and South-Eastern Asia (64% and 87%) and Northern Africa and Western Asia (59% and 93%) (Fig. S9, Fig S10). Uncertainty was highest at intermediate predicted coverage for basic sanitation (Fig. 5a) and generally greater where recent subnational survey data (2015–2024) or a 2024 national JMP estimate were absent (Fig. 5c–h). For open defecation uncertainty generally increased with predicted prevalence (Fig. 5b). However open defecation prevalence rarely exceeded 75 percentage points across admin-1 units, so the relationship at higher prevalence levels could not be characterised. Among the ten countries with the most uncertain predictions (median PI width = 59–83%) the majority lacked recent subnational survey data and/or 2024 national JMP estimates (Fig. S1). We identified four priority countries for data collection where subnational survey data and JMP national estimates are lacking and high prediction uncertainties and population numbers coincide: the Republic of Congo, Eritrea, Nicaragua and North Korea, all of which are classified by the OECD as fragile settings (Fig. S11).

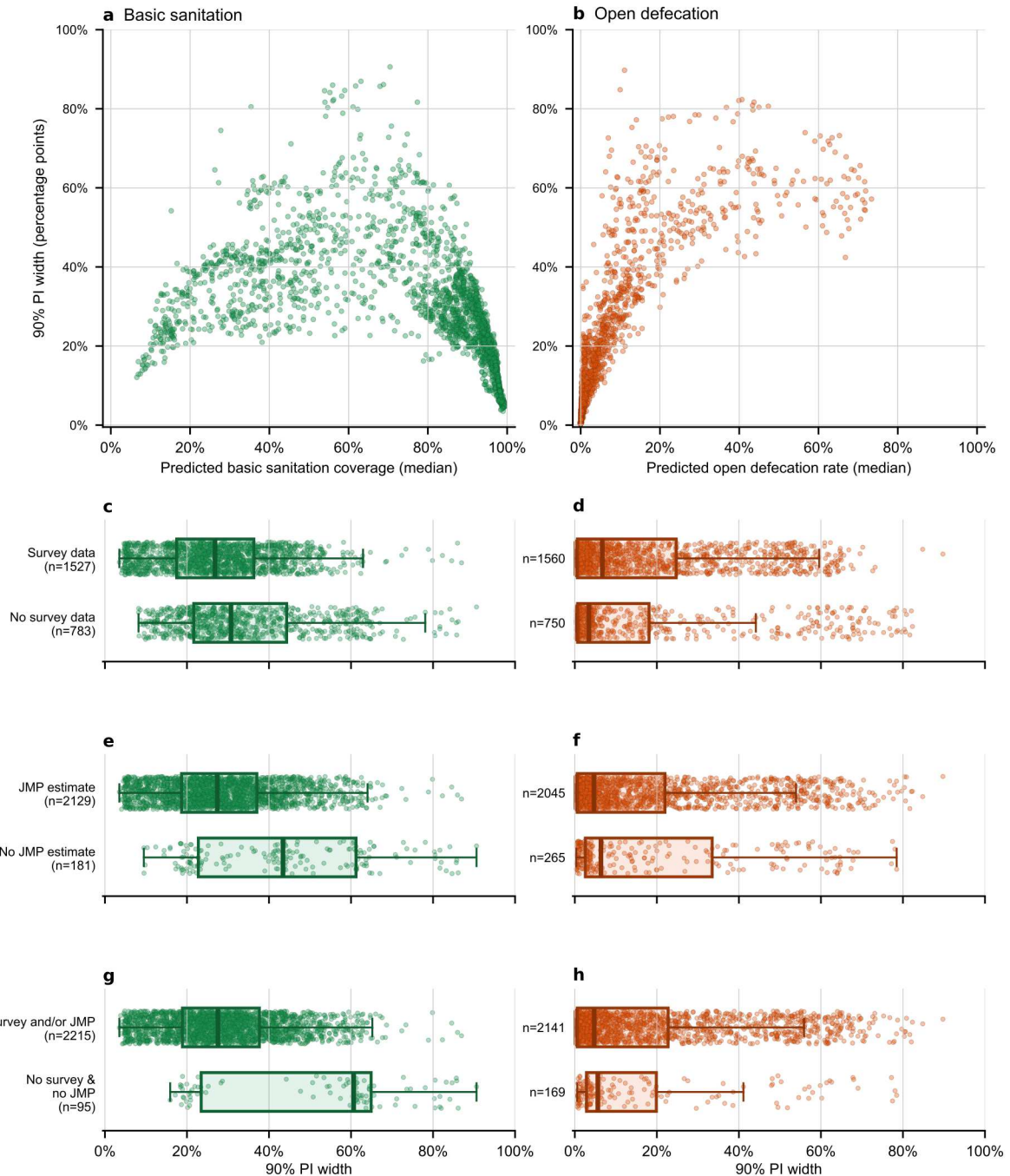


Figure 5. Predictive uncertainty in relation to predicted sanitation outcomes and data availability across admin-1 districts. The x-axis of a, b shows the median prediction of (a) at least basic sanitation coverage or (b) open defecation rate, and the y-axis the width of the 90% prediction interval (percentage points) with each point representing a subnational region. Figures c-h show the distribution of the prediction interval width by survey data (c,d) and JMP estimate (e,f) and

combined survey data and/or JMP estimate availability for basic sanitation (left, green) and open defecation (right, orange) across admin-1 districts (g,h).

Discussion

We applied a transformer-based deep learning model (TabPFN), leveraging household survey and geospatial data from 89 countries to predict basic sanitation use and open defecation at subnational level across 130 LMICs in 2024 (17). TabPFN matched or outperformed conventional random-forest models, capturing 78% of the subnational variation of basic sanitation and 55% open defecation using leave-one-country-out cross-validation. Median basic sanitation coverage predictions ranged from 36% (90% PI=17–56%) in low-income to 94% (PI=76–99%) in upper-middle-income countries with the lowest coverage in fragile, sub-Saharan African settings. Open defecation rates showed the reverse gradient, ranging from 8% (PI=1–29%) in low-income to less than 1% (PI=0–1%) in upper-middle-income countries. Pronounced sanitation service inequalities persisted within countries, particularly in sub-Saharan Africa. Estimated prediction intervals helped identify pockets of uncertainty where future on-the-ground data collection is needed (Fig. S9-S11). Here, we discuss (i) the practical and policy value of granular, harmonized subnational sanitation coverage estimates; (ii) modelling approaches, uncertainties and limitations; and (iii) recommendations for future work addressing higher sanitation service levels.

Gaps in granular data coverage remain a fundamental barrier in the definition of sanitation improvement strategies and priorities. Despite a long-standing recognition of the usefulness of geographic targeting in development programs (5) water and sanitation data gaps remain substantial at subnational levels. Our subnational basic sanitation predictions for 2024 show similar spatial trends to subnational estimates of improved sanitation access in 2017, with the lowest coverage in the sub-Saharan African region (24). The persistence of the lowest sanitation coverage in sub-Saharan Africa across both periods indicates slow progress in reducing inequalities consistent with the JMP progress report from 2024 (1). Many of the areas most affected in our study were identified in low-income countries for which the JMP has estimated that as of 2024, an 18-fold increase in rates of progress in basic sanitation would be needed, on average, to achieve universal access by 2030 (1). Inadequate sanitation and open defecation remain important drivers of diarrhoeal disease, enteric infection and child stunting and mortality (25), burdens concentrated in the same low-income and fragile settings our estimates flag as furthest behind. Critically, low-income, fragile and sub-Saharan African settings that are furthest

from SDG target 6.2 are also where subnational surveys and national JMP estimates can be absent, contributing to high prediction uncertainties. By putting predictions and their uncertainties for these countries on the map, our results highlight areas where service and data needs are the greatest.

Modelling sanitation coverage across LMICs is inherently constrained by sparse and heterogeneous training data, conditions under which TabPFN demonstrates advantages over traditional machine learning approaches such as random forests (17). TabPFN leverages meta-learning across synthetic datasets making it less susceptible to overfitting than other traditional tree-based machine learning models. It is also better suited to analyses in which data are limited and heterogeneous because of its flexibility with complex covariate structures and native missing value imputations. SHAP contributions showed that our models relied heavily on national JMP estimates, using subnational covariates such as night-time lights and population density to add finer spatial structure. This anchoring keeps predictions coherent with national monitoring but means they are limited by the availability and are less certain where a 2024 JMP value is unavailable. Out-of-sample validation methods, such as leave-one-country-out cross-validation, are therefore important to assess how well predictions generalise to data-sparse areas (26).

The opportunities which subnational and higher resolution geospatial and machine learning provide can come at a cost of increased uncertainty compared to national estimates (27, 28). Methods to leverage Earth Observation have only recently matured, making it possible to map SDG targets at subnational level across the globe (15). However, communication of the trade-offs and spatial uncertainties is critical (29). A key methodological advantage of using TabPFN (14) is that it produces distributions of predicted outcomes rather than single point predictions. This enables the generation of prediction intervals which can be mapped across regions and subgroups. Comparing predicted with observed within-country ranges over the same surveyed regions, the model reflected which countries were most internally unequal but compressed the magnitude of local differences, especially when surveys were already several years old (Fig. S7). This compression is consistent with a common challenge in poverty mapping using machine learning. Spatial heterogeneity in the relationship between observable geospatial characteristics and socioeconomic conditions can limit models' ability to distinguish local socioeconomic variation (30, 31). Machine learning applications have been shown to overestimate welfare in poorer rural areas and underestimate it in wealthier urban areas (31). Our

maps therefore may understate the sharpest within-country inequalities and should complement rather than replace local existing data when targeting the worst-served areas.

Aligning definitions of geospatial modelling outcomes with SDG target indicators can help inform sanitation improvement strategies (19, 27). Additional research is needed to develop subnational predictions that support the SDG 6 vision of universal access to adequate and equitable sanitation services, requiring higher levels of services. A specific challenge of estimating safely managed sanitation use is the need to combine information from multiple sources covering the entire sanitation service chain, from user to treatment or disposal (32, 33). Household-level data can only be used to assess the first elements of the service chain, while volumes of wastewater collected and treated are typically measured or estimated at system-level (34). Future research could explore the application of geospatial machine learning approaches as presented herein to predict the first few elements of the sanitation chain (user interface, containment, emptying), and combine it with modelling approaches to downscale existing estimates of wastewater collection and treatment (35). However, relying on the current global indicator definition of safely managed sanitation may lead to overestimating the proportion of safe sanitation systems. This is because global indicators exclude several public-health-relevant dimensions of on-site sanitation safety, such as groundwater contamination, and unsafe or overdue emptying (36). Integrating a more comprehensive set of indicators into global monitoring would help identify sanitation-related health risks but would introduce further data challenges.

Our study has several limitations. First, there are intrinsic limitations to household survey data, such as underreporting of unsafe practices like open defecation (37) which could affect the reliability of our estimates as well as JMP sanitation estimates. Second, data driven methods like TabPFN make a conceptual interpretation challenging. Performance-based interpretation methods such as SHAP-values offer only an approximation of variable relevance based on predictive relationships, limiting our ability to draw causal conclusions about the specific drivers of sanitation access. Third, due to limited data, we were not able to separate an independent test set for 2024 which suitably represented the characteristics across LMICs without losing valuable data for model training. Separating a substantial amount of data from the training data set to validate final model performance is best practice in machine learning contexts when large amounts of data are available. We use leave-one-country-out cross-validation as well as a stratified regional cross-validation sampling approach to approximate this and evaluate the performance of models on out-of-sample areas which provides a robust alternative for spatial validation in this data-scarce context. Temporal out-of-sample validation for 2024 would be

valuable to conduct when more surveys for this time period become available. Fourth, our study is limited to predictions of household use of sanitation services at admin-1 and does not capture inequalities in data coverage and service access below this scale. While our results enable the detection of geographic disparities in household sanitation use between policy relevant administrative areas within countries, they do not capture inequalities at a smaller spatial scale (12) nor do they disaggregate estimates for different population groups (e.g., for example by gender, wealth, disability, ethnicity). As limited sample sizes and cluster densities would likely increase uncertainties, future work leveraging existing georeferenced survey microdata for finer scale predictions should be accompanied by uncertainty quantifications at the scale they are to be used (38) .

Enabling the well-informed prioritization and efficient targeting of investments in data collection and sanitation service improvements is essential in a context of drastic sectoral funding cuts and increasing needs (39). Our work illustrates a transparent approach to filling data gaps, provides clear indications of uncertainties and highlights the vulnerability of low-income and crisis-affected populations. While further research is needed to address higher levels of sanitation service and more granular spatial resolutions, the open-access dataset and code provided here offer a resource for the evidence-based prioritization of data collection and sanitation improvements.

Methods

Sanitation service data and outcome definitions

Variables on sanitation service were derived from nationally representative DHS and MICS surveys which were obtained from the official program websites (40, 41). We selected the most recent DHS and MICS surveys gathered between 2015 and 2024 per country. The geographic scope of our analysis included all LMICs based on the World Bank's income classifications for the calendar year 2024, as well as Ethiopia for which the income group from 2023 was used (18). Countries were further categorised as fragile or extremely fragile following the OECD fragility framework, which identifies contexts facing high economic, environmental, political, security, societal and human risks, and where capacities are insufficient to cope with, manage or absorb those risks (19).

Survey responses on sanitation facility types were extracted and harmonized across DHS (phase 7-9) and MICS surveys (round 6-7) consistent with JMP classifications of improved,

unimproved and open defecation (6). Basic sanitation was defined as the use of an improved sanitation facility which is not shared with other households according to the JMP service ladder definition.

MICS and DHS surveys are designed to be representative of populations at the national and regional level (i.e.: departments, states) using a two-stage sampling approach (42, 43). In the first stage primary sampling units (PSU), usually census enumeration areas, are selected with probability proportional to their size, ensuring that the sample reflects the distribution of the population within a geographic area. In the second step a fixed number of households are randomly selected within each PSU. Because households are selected with unequal probabilities across areas, sampling weights are applied to each household to correct for varying response rates and differences in selection probability during the first sampling stage, ensuring estimates are representative of the target population. We calculated the proportion of the subnational population using basic sanitation services and the proportion practising open defecation by weighting each household's indicator value by its survey household weight and the number of household members. The number of PSUs per subnational region was extracted as a measure of survey precision and applied as an observation weight during model training and evaluation, giving greater influence to regions with more robust survey coverage.

To characterise how sanitation outcomes varied within compared with between countries, we fitted, for each outcome, an intercept-only linear mixed model to the subnational survey estimates, with a random intercept for country and observations weighted by PSU survey weights (lme4 R package). The intraclass correlation coefficient (ICC) was computed as the proportion of total variance attributable to between-country differences:

$$ICC = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_e^2} \text{ (equation 1)}$$

where σ_c^2 is the between-country (random-intercept) variance and σ_e^2 the within-country (residual) variance; values near 1 indicate that variation lies predominantly between countries, and lower values indicate greater subnational variation. We additionally summarised within-country spread as the range (maximum – minimum) of admin-1 estimates within each country.

Linking household surveys to geographic boundaries

To obtain geospatial covariates at subnational level the corresponding geographic boundary was required for each survey region. When available, official DHS and MICS survey boundaries were

downloaded as shapefiles from the program websites for the included surveys (40, 41). For MICS surveys, in cases where no official shapefiles were available, survey region names were matched to administrative boundaries from the Global Administrative Areas database (GADM) version 3.6 using exact name matching (44). Survey regions that could not be linked to GADM region names by exact matching were submitted to fuzzy string matching using the Jaro–Winkler distance metric in the *fuzzyjoin* R package (v.0.1.8) (45), applied within country at each GADM level independently and manually reviewed. In cases where survey names could not be matched to existing GADM boundaries, custom boundaries were defined in Google Earth Engine (46) by merging relevant GADM sub-areas into a single geometry. For the MICS surveys in Nepal (2024) and Kazakhstan (2024), GADM boundaries were outdated and updated subnational administrative boundary files were downloaded from the Humanitarian Data Exchange (47, 48).

Geospatial covariate sampling

Static and dynamic geospatial covariates were sampled for each survey region using Google Earth Engine. Static covariates were sampled using long-term average conditions. Dynamic covariates were sampled using the respective survey year (either as defined in the DHS and MICS shapefiles or the year of survey data collection start). For each survey region, the spatial mean of each covariate was extracted within the survey boundary using zonal statistics. Population totals were additionally extracted as spatial sums for WorldPop and GHSL population layers. In total 87 raster datasets were sampled, spanning a range of environmental and human/socio-economic covariates (Table S1).

In addition, 13 country-level covariates including Gross Domestic Product (GDP) per capita, secondary education duration (49), Worldwide Governance Indicators (50, 51) and national basic sanitation coverage and open defecation prevalence from the JMP 2024 global dataset (7) were matched by country and survey year (Table S2). Wastewater collection, treatment, and reuse percentages were sourced from a global wastewater dataset and matched by country (35). To generate the 2024 predictions, the same static and dynamic covariates were sampled over the GADM (v3.6) admin-1 boundaries of all 130 LMICs for the 2024 prediction year using the identical zonal-statistics procedure, and the corresponding country-level covariates were matched by country.

Feature selection with hierarchical clustering

We applied hierarchical clustering to reduce redundancy amongst pre-selected predictors, also referred to as features, on the full set of 100 pre-selected covariates computed across all admin-1 units of 130 LMICs, using the ClustOfVar package in R (52). Dendrograms were cut at $k = 5, 10, 15, 20, 30, 45,$ and 90 clusters. For each number of clusters, the variable with the highest squared correlation with the first principal component of the variables within that cluster was automatically selected as the cluster representative. The seven subsets of 5 to 90 features, as well as the full set of 100 variables, were evaluated by stratified cross-validation model metrics (see *below*) and the feature set with the best predictive performance was selected to generate predictions of subnational sanitation coverage.

Modelling approach

We compared the performance of Distributed Random Forest models, a traditional machine learning approach previously applied to similar prediction tasks within the WASH domain (14, 20), to that of TabPFN-3 (53). Both models were applied to identical training datasets and evaluated using the same cross-validation framework, which consisted of two main steps: (i) a spatially stratified cross-validation for feature set selection and (ii) a leave-one-country-out cross-validation (LOCO-CV) for final model performance evaluation using the selected feature set.

For the spatially stratified cross-validation, the training data were split into five samples, also referred to as folds, each representing approximately 20% of the training data which are removed (i.e.: held out) during training and then used for validation. Folds were constructed using the proportion of admin-1 units per SDG region observed across the full prediction space of 130 LMICs as a sampling target (Fig. S12, S13). For each iteration, one-fold was held out, and countries were drawn from the remaining training data without replacement until the held-out fold matched this regional distribution across LMICs. The resulting income group distribution across folds was assessed as a diagnostic. The final feature sets were selected based on the R^2 and MAE computed for each fold and averaged across the five folds (equations 2 and 3).

Model performance was evaluated with PSU-weighted cross-validation metrics, namely the coefficient of determination (R^2) and mean absolute error (MAE) defined as:

$$R^2 = 1 - \frac{\sum_i w_i (A_i - \hat{P}_i)^2}{\sum_i w_i (A_i - \bar{A}_w)^2} \quad (\text{equation 2})$$

$$\text{MAE} = \frac{\sum_i w_i |A_i - \hat{P}_i|}{\sum_i w_i} \quad (\text{equation 3})$$

where w_i is the PSU weight for region i , A_i the observed value, $\hat{P}_i$ the predicted value and $\bar{A}_w$ is the PSU-weighted mean of observed values across the held-out set. R^2 describes the proportion of variance in the observed values that is explained by the model. It is upper-bounded by 1 which represents a perfect fit; negative values indicate a fit worse than the weighted mean of the held-out observations. The MAE describes the average absolute difference between predicted and observed values. It is bounded below by 0, which represents perfect prediction, with smaller values indicating closer agreement.

Final model performance was evaluated using LOCO-CV, in which all admin-1 units from one country were withheld and predicted from a model trained on all remaining countries. This was repeated for each country in the training set. Final model performance was summarised using PSU-weighted R^2 and MAE (as defined in equations 2 and 3) and calculated once over all out-of-fold predictions.

The main difference between the random forest and TabPFN implementations was the treatment of PSU-based observation weights: For the random forest, PSU weights were applied both during model training and when computing cross-validation metrics. For TabPFN, PSU weights were applied only at the evaluation stage as the model version used did not support the integration of observation weights.

TabPFN models were selected as the final model based on the best model performance and were used to predict basic sanitation and open defecation outcomes across 130 low- and middle-income countries at admin level 1.

Uncertainty and model diagnostics

Prediction uncertainty was quantified using TabPFN's probabilistic outputs and quantile levels 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, and 0.95 were reported. Based on these the 90% prediction interval (PI; q0.05–q0.95) and 80% PI (q0.10–q0.90) were calculated and used to plot and map prediction uncertainties. Model calibration was assessed by comparing predicted quantile levels to the observed proportion of admin-1 regions for which sanitation estimates fell below each predicted quantile. Priority countries for data collection were identified as those lacking both a recent (2015–2024) subnational survey and a 2024 JMP estimate, and falling above the median on both prediction uncertainty (median 90% PI width) and total population (WorldPop 2024) (Fig. S9).

To assess how the 2024 predictions related to the underlying survey data, we compared them with the observed subnational estimates from survey data gathered between 2015 and 2024 at overlapping areas. For each surveyed admin-1 region, the prediction was aggregated to the survey region's footprint as an area-weighted mean of the overlapping GADM prediction units, computed in an equal-area projection (Equal Earth, EPSG:8857). Boundary concordance between each survey region and its overlapping prediction unit(s) was quantified as the intersection-over-union (IoU) defined as:

$$IoU = \frac{area(S \cap G)}{area(S \cup G)} = \frac{area(S \cap G)}{area(S) + area(G) - area(S \cap G)} \text{ (equation 4)}$$

where S is the survey-region polygon and G the overlapping prediction-unit polygon, with areas computed in an equal-area projection (Equal Earth, EPSG:8857). IoU ranges from 0 (no overlap) to 1 (coincident boundaries). Agreement was summarised by the mean absolute error, overall and stratified by survey year (Fig. S6, Data S4). To evaluate how well predictions captured within-country inequality, we compared, for each country with ≥ 3 surveyed regions, the within-country range (maximum – minimum) of observed estimates with that of predictions over the same regions (Fig. S7).

Feature importance

We quantified predictor importance using SHapley Additive exPlanations (SHAP) values, which measure the extent to which each predictor moves an individual prediction above or below the average prediction (54). We used the model-agnostic permutation estimator (PermutationExplainer, shap v0.49) refitting each final model on all training observations with normalised PSU survey weights. SHAP values were estimated for 250 randomly sampled subnational units, three permutation rounds per unit and five background reference observations. We summarised each predictor's importance as the mean absolute SHAP value across units, expressed as a percentage of the total across all predictors, and aggregated these into thematic categories and by spatial level (country versus subnational) (Tables S1 and S2).

Use of AI tools

Claude Opus 4.8 (Anthropic), accessed via the Claude Code command-line interface, was used to assist with developing code for data cleaning, modelling and visualizing results. It was also used to assist authors in checking numbers and methods reported in the manuscript as well as improving flow and readability of the text. All AI-assisted code and outputs were reviewed and validated by the authors, who take full responsibility for the analyses and their interpretation.

Data and code availability

Estimates of basic sanitation and open defecation were calculated using Multiple Indicator Cluster Survey and Demographic Health Survey data are accessible via the programs' official websites (40, 41). Data sources of geospatial predictor variables are open access and documented and referenced in Tables S1 and S2. Geospatial boundaries are open access and can be accessed via the MICS and DHS websites as well the GADM website (44). Classifications on World Bank income groups, fragile contexts and SDG regions were accessed via the World Bank and Joint Monitoring Program websites (7, 18). All the code needed to replicate our work as well as the preprocessed geospatial data is publicly available on Github (55).

Acknowledgments

We acknowledge intellectual contributions from many colleagues, including Michael Berg, Joel Podgorski, Marco Beti Jesi, Katia Soland, Stuart Dennis, Diana Kriuchkova and Thomas Ost.

Author Contributions

Conceptualization by E.E.G., K.G., G.J., R.J.
Methodology by E.E.G., A.S., K.G., T.R.J., G.J., R.J.
Investigation by E.E.G
Formal analysis by E.E.G
Software by E.E.G., N.E. and J.v.d.H
Validation by E.E.G. and N.E.
Data curation by E.E.G., N.E., J.v.d.H , A.D., Y.R.H, Q.W
Supervision by K.G. and G.J.
Writing original draft by E.E.G. and K.G.
Writing – Review and editing by E.E.G., K.G., A.S., T.R.J., N.E., J.v.d.H., Y.R.H., Q.W., L.S., A.D., R.J.
Visualization by E.E.G., N.E., K.G., A.S., T.R.J
Funding acquisition by E.E.G. and K.G
Project administration by E.E.G and K.G.

Competing interest statement

The authors have the following conflicts of interest to declare: QW, YRH, and GJ are employees of The World Bank and were involved in the study conceptualization and design, results interpretation, and critical review of the manuscript.

Funding

This work was supported by the World Bank (consultancy contract ref. 7219818).

References

1. UNICEF; WHO, “Progress on household drinking water, sanitation and hygiene 2000-2024 Special focus on inequalities WHO/UNICEF Joint Monitoring Programme for Water Supply, Sanitation and Hygiene” (2024).
2. R. Bain, R. Johnston, F. Mitis, C. Chatterley, T. Slaymaker, Establishing Sustainable Development Goal Baselines for Household Drinking Water, Sanitation and Hygiene Services. *Water* 2018, Vol. 10, Page 1711 **10**, 1711 (2018).
3. G. Joseph, Y. R. Hoo, Q. Wang, A. Bahuguna, L. Andres, “FUNDING A WATER-SECURE FUTURE An Assessment of Global Public Spending;” www.worldbank.org/gwsp.
4. M. Gualavisi, D. Newhouse, Integrating Survey and Geospatial Data for Geographical Targeting of the Poor and Vulnerable: Evidence from Malawi. *World Bank Econ. Rev.* **39**, 377–409 (2025).
5. D. Bigman, H. Fofack, Geographical Targeting for Poverty Alleviation: An Introduction to the Special Issue. *World Bank Econ. Rev.* **14**, 129–145 (2000).
6. World Health Organization, United Nations Children’s Fund, JMP Methodology 2017 Update & SDG Baselines. (2018).
7. Downloads index | JMP. <https://washdata.org/data/downloads>.
8. UNICEF, Data Warehouse. [https://sdmx.data.unicef.org/databrowser/index.html?q=UNICEF:WASH_HOUSEHOLD_SUBNAT\(1.0\)](https://sdmx.data.unicef.org/databrowser/index.html?q=UNICEF:WASH_HOUSEHOLD_SUBNAT(1.0)).
9. T. Kilic, U. Serajuddin, H. Uematsu, N. Yoshida, Costing Household Surveys for Monitoring Progress Toward Ending Extreme Poverty and Boosting Shared Prosperity. (2017).
10. L. A. Andres Samir Bhatt Basab Dasgupta Juan A Echenique Peter W Gething Jonathan Grabinsky Zabludovsky George Joseph, Geo-Spatial Modeling of Access to Water and Sanitation in Nigeria. (2018).
11. R. L. Pullan, M. C. Freeman, P. W. Gething, S. J. Brooker, Geographical Inequalities in Use of Improved Drinking Water Supply and Sanitation across Sub-Saharan Africa: Mapping and Spatial Analysis of Cross-sectional Survey Data. *PLoS Med.* **11**, e1001626 (2014).
12. C. V. Nguyen, Coverage of safely managed water, sanitation and hygiene: An application of small area estimation. *Science of The Total Environment* **896**, 165277 (2023).

13. A. A. Ambel, R. Bain, T. B. Degefu, A. Donmez, R. Johnston, T. Slaymaker, Addressing gaps in data on drinking water quality through data integration and machine learning: evidence from Ethiopia. *npj Clean Water* 2023 6:1 **6**, 1–9 (2023).
14. E. E. Greenwood, T. Lauber, J. van den Hoogen, A. Donmez, R. E. S. Bain, R. Johnston, T. W. Crowther, T. R. Julian, Mapping safe drinking water use in low- and middle-income countries. *Science* (1979). **385**, 784–790 (2024).
15. A. Andries, S. Morse, R. J. Murphy, J. Lynch, E. R. Woolliams, Using Data from Earth Observation to Support Sustainable Development Indicators: An Analysis of the Literature and Challenges for the Future. *Sustainability (Switzerland)* **14**, 1191 (2022).
16. Ian Goodfellow, Aaron Courville, Yoshua Bengio, Deep Learning. *Request exam copy View preview Adaptive Computation and Machine Learning series*, 1–23 (2016).
17. N. Hollmann, S. Müller, L. Purucker, A. Krishnakumar, M. Körfer, S. Bin Hoo, R. T. Schirrmeister, F. Hutter, Accurate predictions on small data with a tabular foundation model. *Nature* 2025 637:8045 **637**, 319–326 (2025).
18. World Bank Country and Lending Groups – World Bank Data Help Desk. <https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups>.
19. Multidimensional fragility | OECD. <https://www.oecd.org/en/data/dashboards/multidimensional-fragility.html>.
20. Distributed Random Forest (DRF) — H2O 3.46.0.12 documentation. <https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/drfs.html>.
21. S. Vyas, N. Srivastav, D. Mary, N. Goel, S. Srinivasan, A. Tannirkulam, R. Ban, D. Spears, D. Coffey, Measuring open defecation in India using survey questions: evidence from a randomised survey experiment. *BMJ Open* **9**, e030152 (2019).
22. N. G. Exum, E. M. Gorin, G. Sadhu, A. Khanna, K. J. Schwab, Evaluating the declarations of open defecation free status under the Swachh Bharat ('Clean India') Mission: repeated cross-sectional surveys in Rajasthan, India. *BMJ Glob. Health* **5**, 2277 (2020).
23. G. Rosana, B. K. Kigen, C. Nyakundi, S. Okemwa, S. Toweet, J. O. Oguta, C. Akoth, Prevalence and factors associated with Open Defecation in Burkina Faso: A trend analysis of the 2003-2021 Burkina Faso Demographic and Health Surveys. doi: 10.21203/RS.3.RS-8920802/V1 (2026).
24. Local Burden of Disease Collaborators, Mapping geographical inequalities in access to drinking water and sanitation facilities in low-income and middle-income countries, 2000-17. *Lancet Glob. Health* **8**, e1162–e1185 (2020).
25. J. Wolf, R. B. Johnston, A. Ambelu, B. F. Arnold, R. Bain, M. Brauer, J. Brown, B. A. Caruso, T. Clasen, J. M. Colford, J. E. Mills, B. Evans, M. C. Freeman, B. Gordon, G. Kang, C. F. Lanata, K. O. Medlicott, A. Prüss-Ustün, C. Troeger, S. Boisson, O. Cumming, Burden of disease attributable to unsafe drinking water, sanitation, and hygiene in domestic settings: a global analysis for selected adverse health outcomes. *The Lancet* **401**, 2060–2071 (2023).
26. C. Milà, J. Mateu, E. Pebesma, H. Meyer, Nearest neighbour distance matching Leave-One-Out Cross-Validation for map validation. *Methods Ecol. Evol.* **13**, 1304–1316 (2022).
27. C. W. Ruktanonchai, J. J. Nieves, N. W. Ruktanonchai, K. Nilsen, J. E. Steele, Z. Matthews, A. J. Tatem, Estimating uncertainty in geospatial modelling at multiple spatial resolutions: the pattern of delivery via caesarean section in Tanzania. *BMJ Glob. Health* **4**, e002092 (2020).
28. N. Tzavidis, Small Area Estimation in the Era of Machine Learning and Alternative Data Sources: Opportunities, Challenges, and Outlook. *J. Off. Stat.* **41**, 921–929 (2025).

29. E. M. Delmelle, M. R. Desjardins, P. Jung, C. Owusu, Y. Lan, A. Hohl, C. Dony, Uncertainty in geospatial health: challenges and opportunities ahead. *Ann. Epidemiol.* **65**, 15–30 (2022).
30. C. Yeh, A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, M. Burke, Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications* **11** (2020).
31. Y. Ru, E. Tennant, D. Matteson, C. Barrett, S. B. J. Ashley, “Spatial Heterogeneity in Machine Learning-Based Poverty Mapping: Where Do Models Underperform?” (International, 2025); <https://doi.org/10.22617/WPS250340-2>.
32. Monitoring safely managed on-site sanitation (SMOSS) Synthesis of lessons from phase 1 pilots and recommendations for phase 2 pilots. (2021).
33. K. T. Swe, M. M. Rahman, M. S. Rahman, Y. Teng, S. K. Abe, M. Hashizume, K. Shibuya, Impact of poverty reduction on access to water and sanitation in low- and lower-middle-income countries: country-specific Bayesian projections to 2030. *Tropical Medicine & International Health* **26**, 760–774 (2021).
34. Domestic Wastewater Treatment Methodology 2024 | UN-Water. <https://www.unwater.org/publications/domestic-wastewater-treatment-methodology-2024>.
35. E. R. Jones, M. T. H. Van Vliet, M. Qadir, M. F. P. Bierkens, Country-level and gridded estimates of wastewater production, collection, treatment and reuse. *Earth Syst. Sci. Data* **13**, 237–254 (2021).
36. F. Mills, T. Foster, A. Kome, R. Munankami, G. Halcrow, A. Ndungu, B. Evans, J. Willetts, Indicators to complement global monitoring of safely managed on-site sanitation to understand health risks. *npj Clean Water* **2024 7:1 7**, 58- (2024).
37. S. Vedachalam, L. H. MacDonald, S. Shiferaw, A. Seme, K. J. Schwab, G. Guiella, L. M. Zan, P. Anglewicz, J. Bertrand, P. Kayembe, E. Otupiri, S. Kirlan, T. Y. Prihyugiarto, S. Wilopo, Ansariadi, N. L. Lubis, O. Habi, S. Oumarou, E. Omuluabi, F. OlaOlorun, F. Makumbi, S. P. S. Kibira, S. Radloff, A. Tsui, S. Ahmed, H. Olson, L. Zimmerman, Underreporting of high-risk water and sanitation practices undermines progress on global targets. *PLoS One* **12**, e0176272 (2017).
38. N. Tzavidis, Small Area Estimation in the Era of Machine Learning and Alternative Data Sources: Opportunities, Challenges, and Outlook. *J. Off. Stat.* **41**, 921–929 (2025).
39. OECD, Cuts in official development assistance: OECD projections for 2025 and the near term. *OECD Policy Briefs*, doi: 10.1787/8C530629-EN (2025).
40. Surveys - UNICEF MICS. <https://mics.unicef.org/surveys>.
41. The DHS Program - Data. <https://dhsprogram.com/data/>.
42. S. Khan, A. Hancioglu, Multiple Indicator Cluster Surveys: Delivering Robust Data on Children and Women across the Globe. *Stud. Fam. Plann.* **50**, 279–286 (2019).
43. Analyzing DHS Data. https://dhsprogram.com/data/Guide-to-DHS-Statistics/Analyzing_DHS_Data.htm.
44. GADM. https://gadm.org/download_country_v3.html.
45. D. Robinson, Join Tables Together on Inexact Matching [R package fuzzyjoin version 0.1.8]. *CRAN: Contributed Packages*, doi: 10.32614/CRAN.PACKAGE.FUZZYJOIN (2026).
46. N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, R. Moore, Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **202**, 18–27 (2017).
47. Kazakhstan - Subnational Administrative Boundaries | Humanitarian Dataset | HDX. <https://data.humdata.org/dataset/cod-ab-kaz>.
48. Nepal - Subnational Administrative Boundaries | Humanitarian Dataset | HDX. <https://data.humdata.org/dataset/cod-ab-npl>.

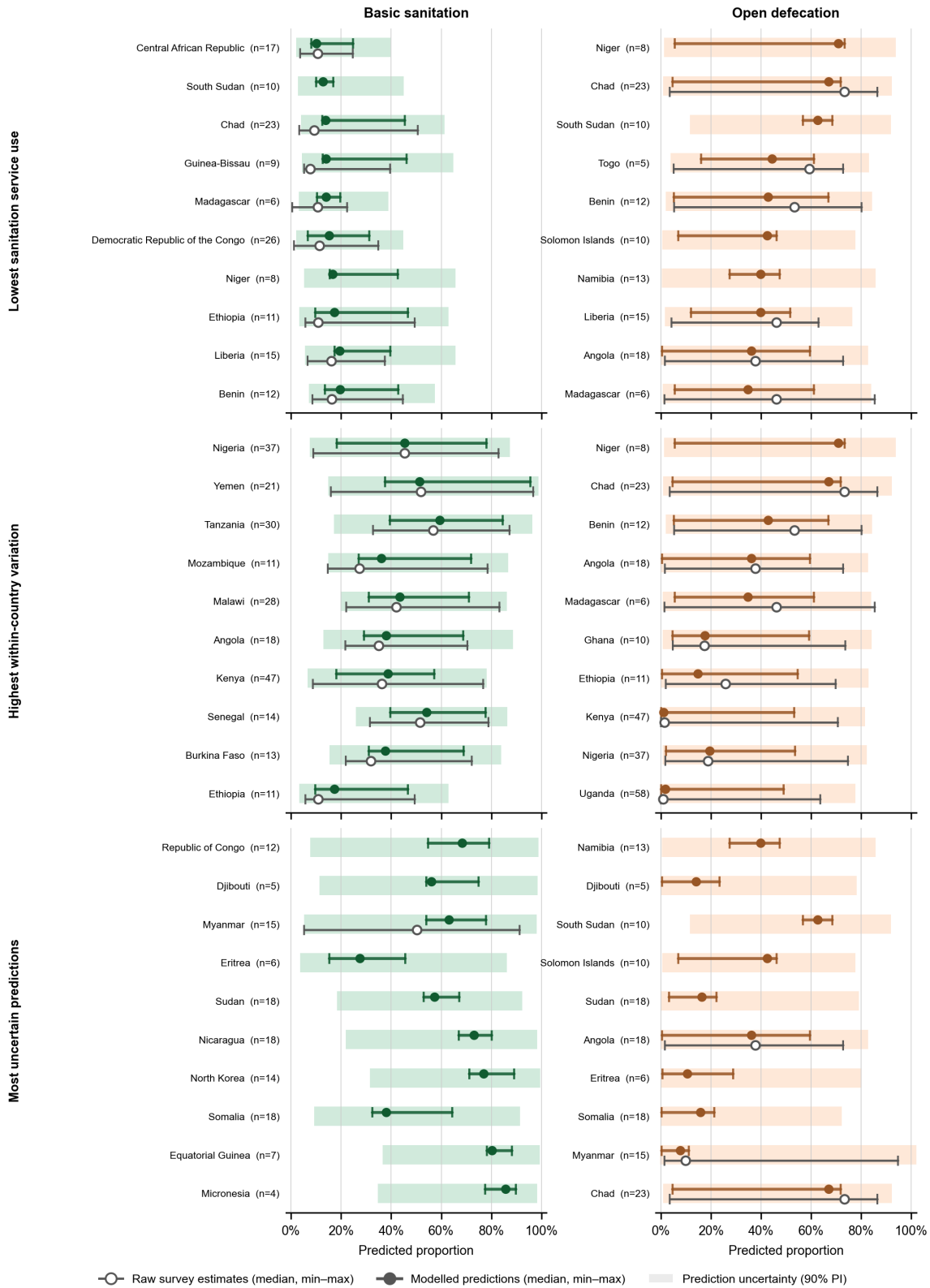
49. World Development Indicators | DataBank. <https://databank.worldbank.org/source/world-development-indicators/>.
50. World Bank, Worldwide Governance Indicators. <https://www.worldbank.org/en/publication/worldwide-governance-indicators>.
51. D. Kaufmann, A. Kraay, M. Mastruzzi, The Worldwide Governance Indicators: Methodology and Analytical Issues. [Preprint] (2010). <https://papers.ssrn.com/abstract=1682130>.
52. ClustOfVar package - RDocumentation. <https://www.rdocumentation.org/packages/ClustOfVar/versions/1.1>.
53. N. Hollmann, S. Müller, L. Purucker, A. Krishnakumar, M. Körfer, S. Bin Hoo, R. T. Schirrmeister, F. Hutter, Accurate predictions on small data with a tabular foundation model. *Nature* **637**, 319–326 (2025).
54. S. M. Lundberg, P. G. Allen, S.-I. Lee, A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* **30** (2017).
55. Code Availability in Github. https://github.com/PHIC-Eawag/subnational_water_sanitation_estimates.
56. Global Aridity Index and Potential Evapotranspiration (ET0) Climate Database v2. https://figshare.com/articles/dataset/Global_Aridity_Index_and_Potential_Evapotranspiration_ET0_Climate_Database_v2/7504448/3.
57. D. N. Karger, O. Conrad, J. Böhrer, T. Kawohl, H. Kreft, R. W. Soria-Auza, N. E. Zimmermann, H. P. Linder, M. Kessler, Climatologies at high resolution for the earth's land surface areas. *Scientific Data* **2017 4:1 4**, 1–20 (2017).
58. Envidat - Environmental Research Data at your Fingertips. <https://www.envidat.ch/#/metadata/chelsa-climatologies>.
59. G. T., R.-C. R.M., V. L.V., H. M., W. F., H. E., H. N., M. D., Tropical and Subtropical Wetlands Distribution version 2. *Tropical and Subtropical Wetlands Distribution version 2*, doi: 10.17528/CIFOR/DATA.00058.
60. C. M. Kennedy, J. R. Oakleaf, D. M. Theobald, S. Baruch-Mordo, J. Kiesecker, Managing the middle: A shift in conservation priorities based on the global human modification gradient. *Glob. Chang. Biol.* **25**, 811–826 (2019).
61. M. N. Tuanmu, W. Jetz, A global 1-km consensus land-cover product for biodiversity and ecosystem modelling. *Global Ecology and Biogeography* **23**, 1031–1045 (2014).
62. T. W. Crowther, H. B. Glick, K. R. Covey, C. Bettigole, D. S. Maynard, S. M. Thomas, J. R. Smith, G. Hintler, M. C. Duguid, G. Amatulli, M. N. Tuanmu, W. Jetz, C. Salas, C. Stam, D. Piotto, R. Tavani, S. Green, G. Bruce, S. J. Williams, S. K. Wiser, M. O. Huber, G. M. Hengeveld, G. J. Nabuurs, E. Tikhonova, P. Borchardt, C. F. Li, L. W. Powrie, M. Fischer, A. Hemp, J. Homeier, P. Cho, A. C. Vibrans, P. M. Umunay, S. L. Piao, C. W. Rowe, M. S. Ashton, P. R. Crane, M. A. Bradford, Mapping tree density at a global scale. *Nature* **2015 525:7568 525**, 201–205 (2015).
63. A. M. Wilson, W. Jetz, Remotely Sensed High-Resolution Global Cloud Dynamics for Predicting Ecosystem and Biodiversity Distributions. *PLoS Biol.* **14**, e1002415 (2016).
64. G. Amatulli, S. Domisch, M. N. Tuanmu, B. Parmentier, A. Ranipeta, J. Malczyk, W. Jetz, A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. *Scientific Data* **2018 5:1 5**, 1–15 (2018).
65. ESA Land Cover CCI project team; Defourny, P. (2016): ESA Land Cover Climate Change Initiative (Land_Cover_cci): Land Surface Seasonality Products. Centre for Environmental Data Analysis. <https://catalogue.ceda.ac.uk/uuid/7c114fc6e2884c1f9ca107e7a502fdbf>.
66. Y. Fan, H. Li, G. Miguez-Macho, Global patterns of groundwater table depth. *Science* (1979). **339**, 940–943 (2013).

67. M. Pesaresi, M. Schiavina, P. Politis, S. Freire, K. Krasnodębska, J. H. Uhl, A. Carioli, C. Corbane, L. Dijkstra, P. Florio, H. K. Friedrich, J. Gao, S. Leyk, L. Lu, L. Maffenini, I. Mari-Rivero, M. Melchiorri, V. Syrris, J. Van Den Hoek, T. Kemper, Advances on the global human settlement layer by joint assessment of earth observation and population survey data. *Int. J. Digit. Earth* **17** (2024).
68. M. Schiavina, S. Freire, Alessandra Carioli, K. MacManus, “JRC Data Catalogue - Dataset - European Commission” (2026); <http://data.europa.eu/89h/2ff68a52-5b5b-4a22-8f40-c41da8332cfe>.
69. M. Gilbert, G. Nicolas, G. Cinardi, T. P. Van Boeckel, S. O. Vanwambeke, G. R. W. Wint, T. P. Robinson, Global distribution data for cattle, buffaloes, horses, sheep, goats, pigs, chickens and ducks in 2010. *Scientific Data* **2018 5:1** **5**, 1–11 (2018).
70. UN-Adjusted Population Density, v4: Gridded Population of the World (GPW), v4 | SEDAC. <https://beta.sedac.ciesin.columbia.edu/data/set/gpw-v4-population-density-adjusted-to-2015-unwpp-country-totals>.
71. C. Giri, E. Ochieng, L. L. Tieszen, Z. Zhu, A. Singh, T. Loveland, J. Masek, N. Duke, Status and distribution of mangrove forests of the world using earth observation satellite data. *Global Ecology and Biogeography* **20**, 154–159 (2011).
72. K. Didan, MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V061 [Dataset]. NASA Land Processes Distributed Active Archive Center. (2021). <https://doi.org/10.5067/MODIS/MOD13Q1.061> Date Accessed: 2026-09-25.
73. S. Running, M. Zhao, MODIS/Terra Net Primary Production Gap-Filled Yearly L4 Global 500 m SIN Grid V006 [Dataset]. NASA Land Processes Distributed Active Archive Center. (2019). <https://doi.org/10.5067/MODIS/MOD17A3HGF.006>.
74. J. D. Pelletier, P. D. Broxton, P. Hazenberg, X. Zeng, P. A. Troch, G. Y. Niu, Z. Williams, M. A. Brunke, D. Gochis, A gridded global data set of soil, intact regolith, and sedimentary deposit thicknesses for regional and global land surface modeling. *J. Adv. Model. Earth Syst.* **8**, 41–65 (2016).
75. T. Hengl, J. M. De Jesus, G. B. M. Heuvelink, M. R. Gonzalez, M. Kilibarda, A. Blagotić, W. Shangquan, M. N. Wright, X. Geng, B. Bauer-Marschallinger, M. A. Guevara, R. Vargas, R. A. MacMillan, N. H. Batjes, J. G. B. Leenaars, E. Ribeiro, I. Wheeler, S. Mantel, B. Kempen, SoilGrids250m: Global gridded soil information based on machine learning. *PLoS One* **12**, e0169748 (2017).
76. Ardalan Tootchi, Anne Jost, Agnès Ducharne, Multi-source global wetland mapping: combining surface water imagery and groundwater constraints, *preprint* (2018). <https://hess.copernicus.org/preprints/hess-2018-48/>.
77. O. Venter, E. W. Sanderson, A. Magrath, J. R. Allan, J. Beher, K. R. Jones, H. P. Possingham, W. F. Laurance, P. Wood, B. M. Fekete, M. A. Levy, J. E. M. Watson, Global terrestrial Human Footprint maps for 1993 and 2009. *Sci. Data* **3** (2016).
78. C. Funk, P. Peterson, M. Landsfeld, D. Pedreros, J. Verdin, S. Shukla, G. Husak, J. Rowland, L. Harrison, A. Hoell, J. Michaelsen, The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific Data* **2015 2:1** **2**, 1–21 (2015).
79. J. Muñoz-Sabater, E. Dutra, A. Agustí-Panareda, C. Albergel, G. Arduini, G. Balsamo, S. Boussetta, M. Choulga, S. Harrigan, H. Hersbach, B. Martens, D. G. Miralles, M. Piles, N. J. Rodríguez-Fernández, E. Zsoter, C. Buontempo, J. N. Thépaut, ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth Syst. Sci. Data* **13**, 4349–4383 (2021).
80. M. Pesaresi, Politis Panagiotis, GHS-BUILT-S R2023A - GHS built-up surface grid, derived from Sentinel2 composite and Landsat, multitemporal (1975-2030). . *European Commission, Joint Research Centre [Dataset]* (2026).

81. D. J. Weiss, A. Nelson, C. A. Vargas-Ruiz, K. Gligorić, S. Bavadekar, E. Gabrilovich, A. Bertozzi-Villa, J. Rozier, H. S. Gibson, T. Shekel, C. Kamath, A. Lieber, K. Schulman, Y. Shao, V. Qarkaxhija, A. K. Nandi, S. H. Keddie, S. Rumisha, P. Amratia, R. Arambepola, E. G. Chestnutt, J. J. Millar, T. L. Symons, E. Cameron, K. E. Battle, S. Bhatt, P. W. Gething, Global maps of travel time to healthcare facilities. *Nature Medicine* 2020 26:12 26, 1835–1838 (2020).
82. C. D. Elvidge, M. Zhizhin, T. Ghosh, F. C. Hsu, J. Taneja, Annual time series of global viirs nighttime lights derived from monthly averages: 2012 to 2019. *Remote Sens. (Basel)*. 13, 1–14 (2021).
83. M. Bondarenko, R. Priyatikanto, N. Tejedor Garavito, W. Zhang, T. McKeen, A. Cunningham, T. Woods, J. Hilton, D. Cihan, B. Nosatiuk, T. Brinkhoff, A. Tatem, A. Sorichetta, Constrained estimates of 2015-2030 total number of people per grid square at a resolution of 3 arc (approximately 100m at the equator) R2025A version v1. University of Southampton. (2025).
84. B. Šavrič, T. Patterson, B. Jenny, The Equal Earth map projection. *International Journal of Geographical Information Science* 33, 454–465 (2019).

862

Supplemental Figures and Tables



863

Figure S1 Subnational raw survey estimates, predictions and their uncertainty, for basic sanitation (green, left) and open defecation (orange, right). For each country, the filled dot represents the country median of admin 1 median predictions; the empty dot represents the median of the raw survey estimates and is only present where data was available. The solid line with end-caps spans the range of district-level median predictions from minimum to maximum; and the shaded bar shows the full 90% prediction-interval, from the lowest district lower bound to the highest district upper bound. Countries are ordered by most-extreme case first within each block.

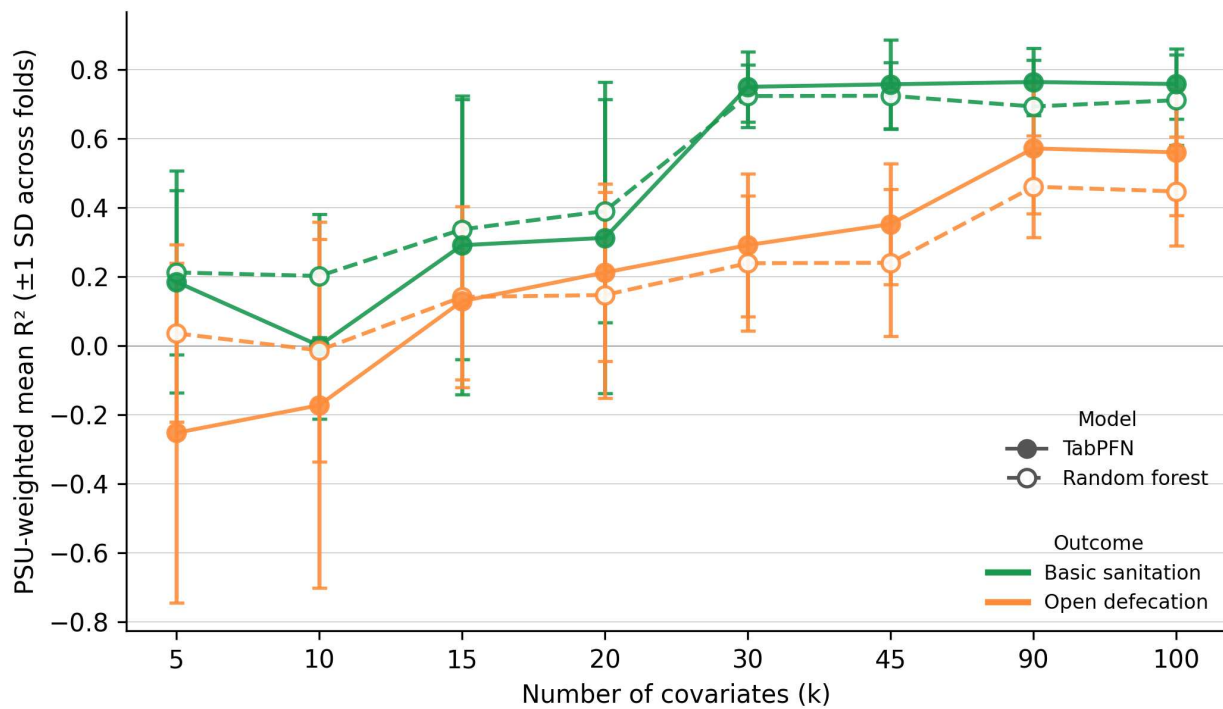


Figure S2 Stratified cross-validation performance of TabPFN and random forest models by feature set size for basic sanitation (green) and open defecation (orange). Mean PSU weighted R^2 (± 1 SD) across five repeated stratified folds is shown for candidate feature sets of $k = 5$ to $k = 100$ covariates, where each k represents the number of cluster representatives selected from the full set of 100 candidate covariates.

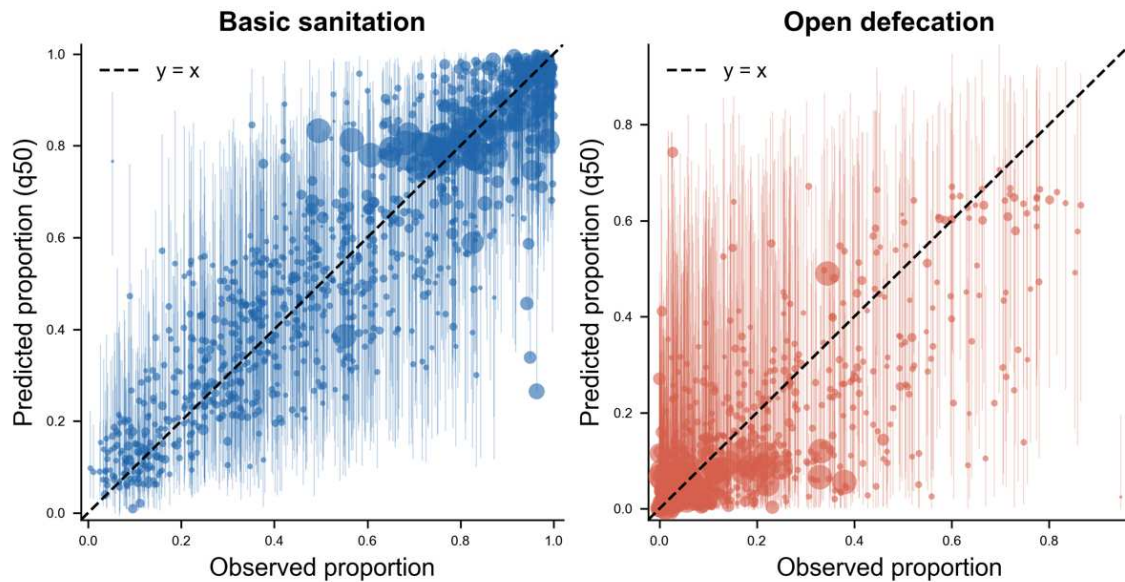


Figure S3 Observed versus predicted proportions (median prediction value) from TabPFN leave-one-country-out cross-validation for basic sanitation (left, blue) and open defecation (right, red). Each point represents one admin-1 region, sized proportionally to PSU weight. Vertical error bars represent the 90% prediction interval (q0.05–q0.95). The dashed line indicates perfect agreement between observed and predicted values.

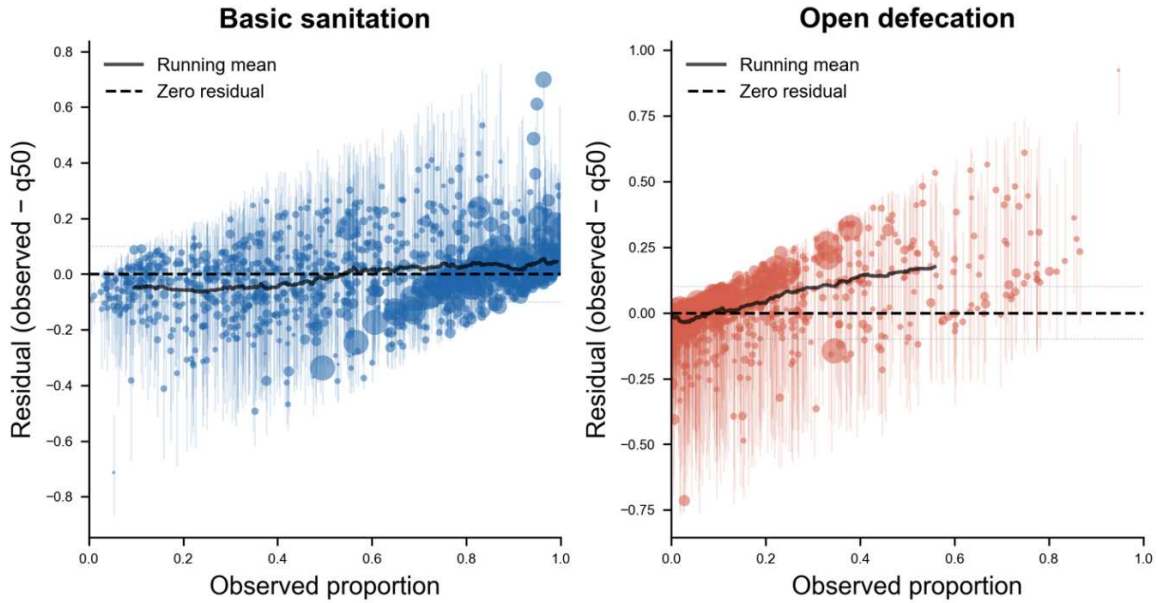


Figure S4 Residuals (observed – q50) plotted against observed proportions from TabPFN leave-one-country-out cross-validation for basic sanitation (left, blue) and open defecation (right, red). Each point represents one admin-1 region, sized proportionally to PSU weight. Vertical error bars represent the spread of the 90% prediction interval expressed as residuals (q0.05 – observed to q0.95 – observed). The solid line shows the running mean of residuals and the dashed line indicates zero residual (perfect prediction).

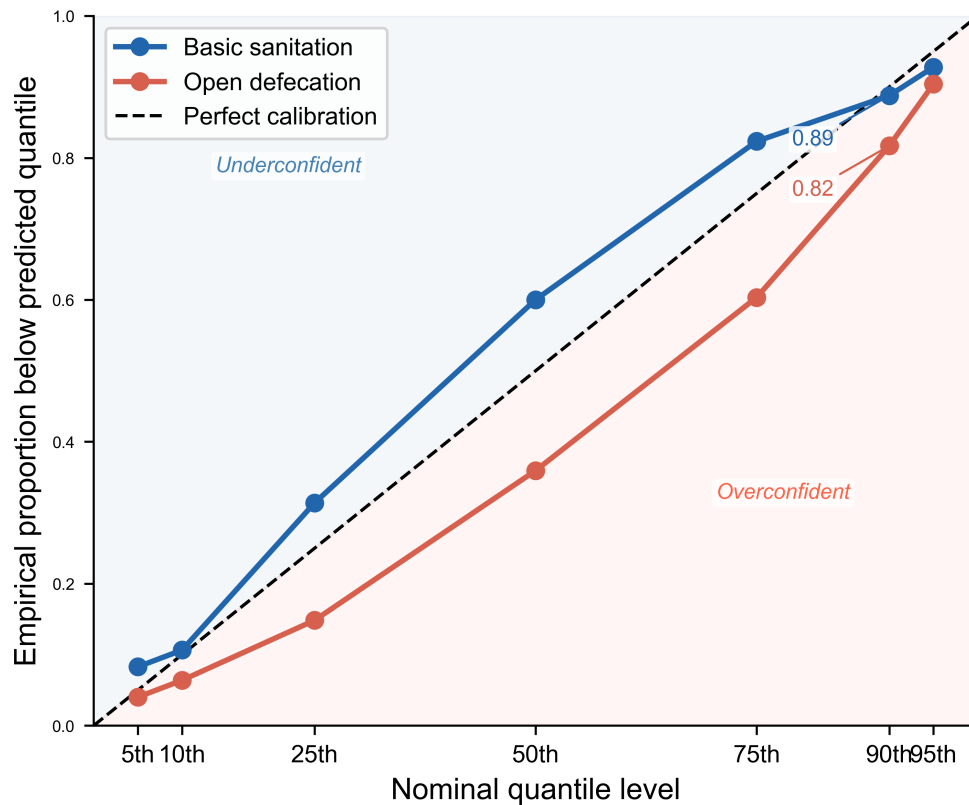


Figure S5 Calibration diagram for TabPFN quantile predictions from leave-one-country-out cross-validation for basic sanitation (blue) and open defecation (red). The x-axis shows the quantile level and the y-axis shows the PSU-weighted proportion of admin-1 regions whose observed value fell below the predicted quantile. The dashed line indicates perfect calibration, where empirical coverage matches the nominal quantile level exactly. Points above the diagonal indicate underconfident predictions (intervals wider than necessary) and points below indicate overconfident predictions (intervals too narrow).

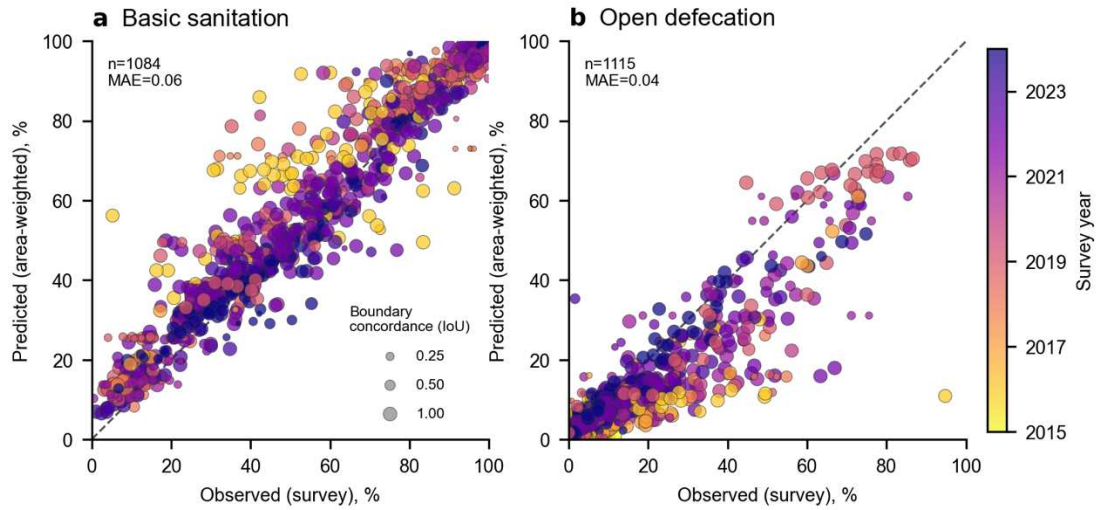


Figure S6 Agreement between observed subnational survey estimates (2015-2024) and model predictions (2024), for at least basic sanitation (left) and open defecation (right) (Data S4). Each point represents one surveyed admin-1 region and shows the observed survey estimate (x-axis) against the model prediction aggregated to the same survey footprint by area-weighted overlay (y-axis); the dashed line marks perfect agreement (1:1). Point colour indicates the survey year (2015–2024) and point size the boundary concordance between the survey region and the overlapping GADM prediction unit(s), measured as the intersection-over-union (IoU; larger points indicate closer boundary alignment). n, number of surveyed regions; MAE, mean absolute error (as a proportion, 0–1 scale).

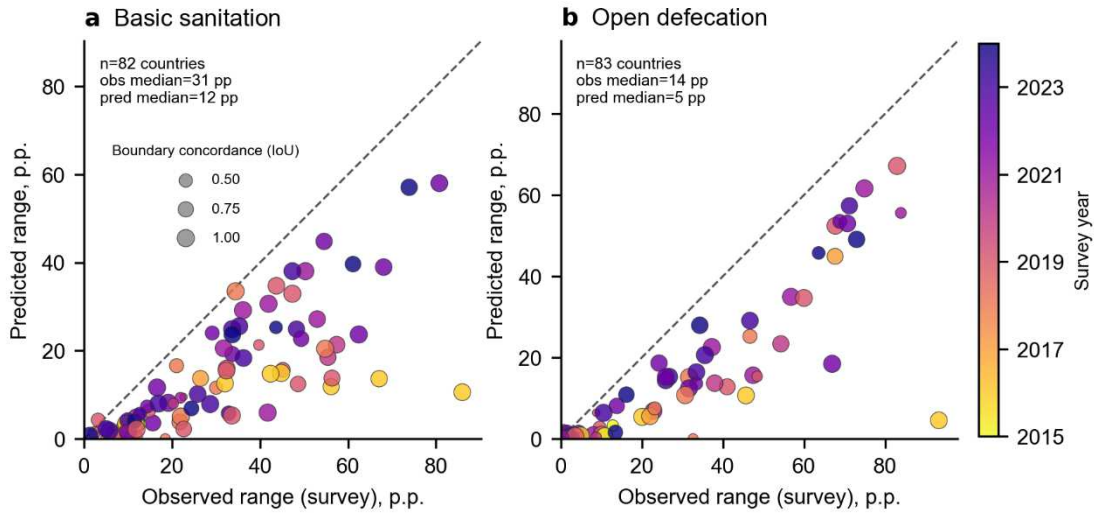


Figure S7 Within-country inequality: observed (2015-2024) versus predicted (2024). For each country with at least three surveyed admin-1 regions, the within-country range of the observed survey estimates (x-axis) versus the model predictions (y-axis), computed over the same regions, for (a) at least basic sanitation and (b) open defecation (Data S4). Predictions were aggregated to each survey region's footprint as an area-weighted mean of the overlapping prediction units. Each point is one country; point colour indicates survey year and point size the mean boundary concordance (IoU) across the country's regions. The dashed line marks perfect agreement (1:1).

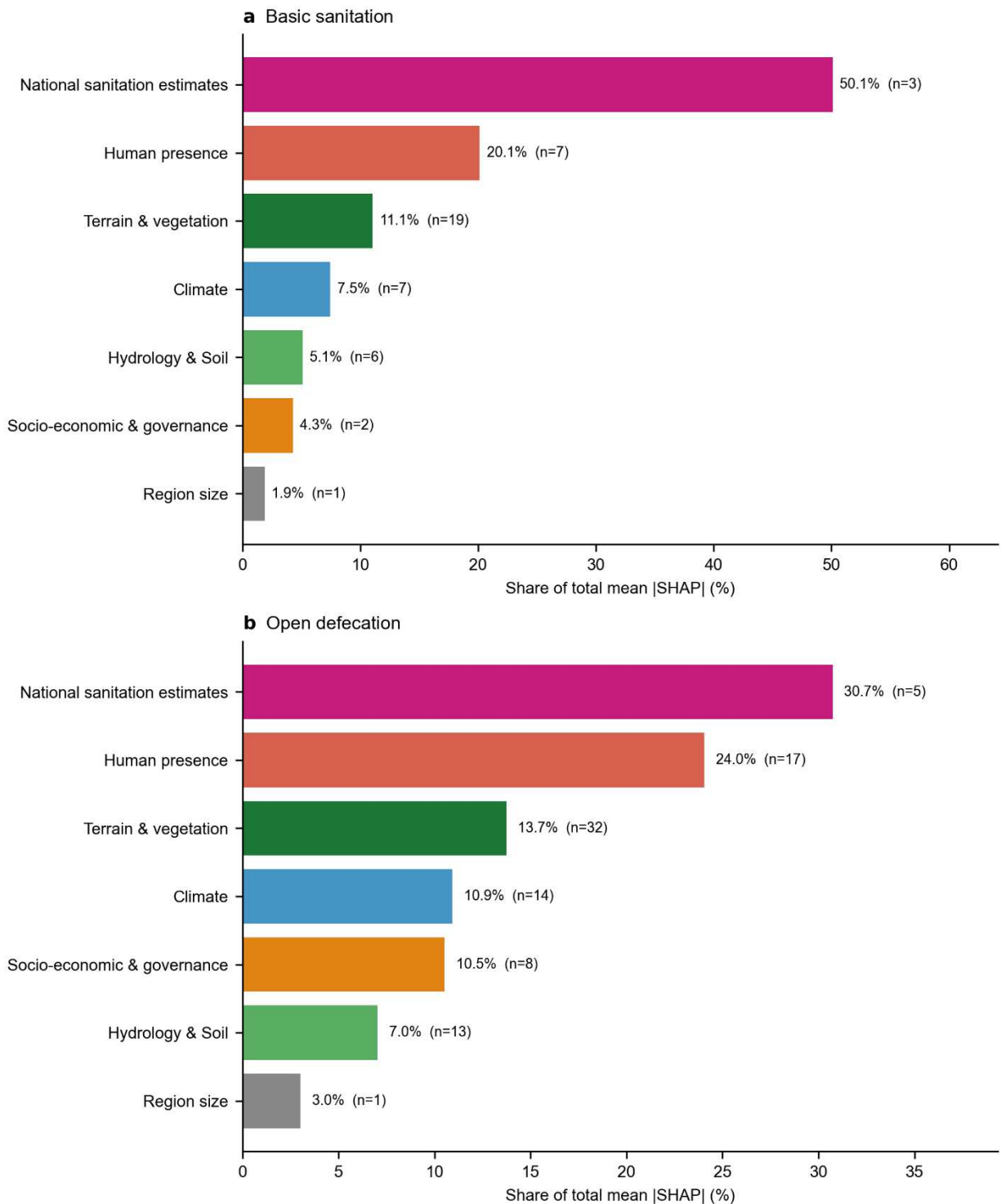
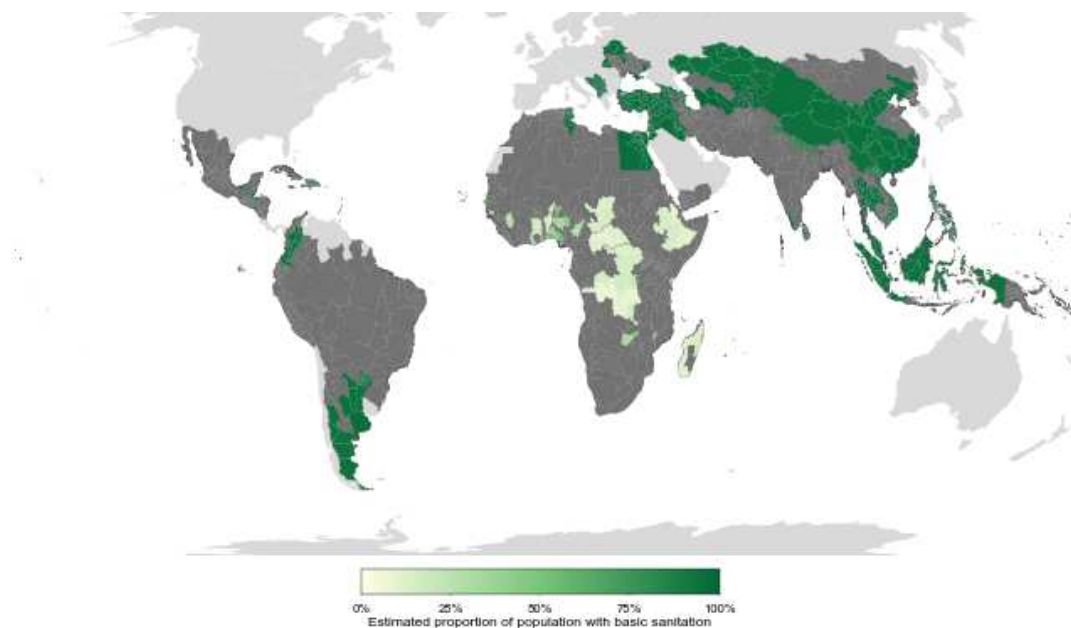


Figure S8 Relative importance of predictor categories for basic sanitation ($k = 45$) and open defecation ($k = 90$), as each category's share of total mean absolute Shapely explanation values ($|SHAP|$ (%)). n = predictors per category (Table S1/S2).

937



938

939 **Figure S9.** Subnational areas where basic sanitation predictions are the most certain. Dark grey
 940 areas represent areas where the 90% prediction interval width is greater than 25 percentage
 941 points. Light grey areas represent areas outside low-and middle-income countries.

942

943

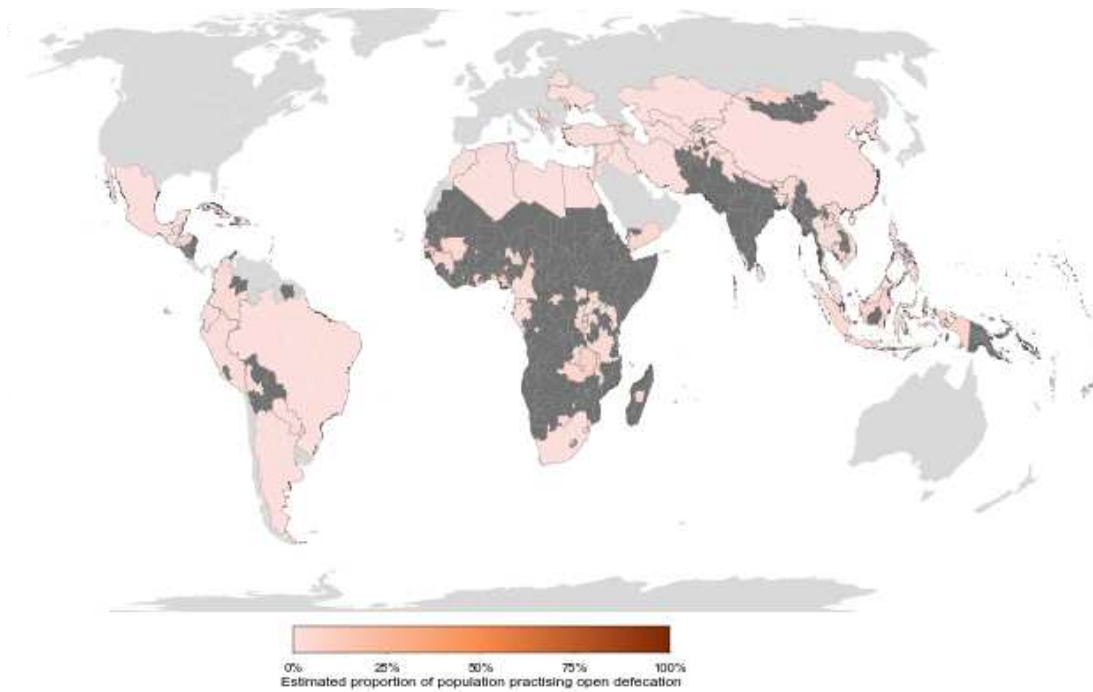


Figure S10. Subnational areas where open defecation predictions are the most certain. Dark grey areas represent areas where the 90% prediction interval width is greater than 25 percentage points. Light grey areas represent areas outside low- and middle-income countries

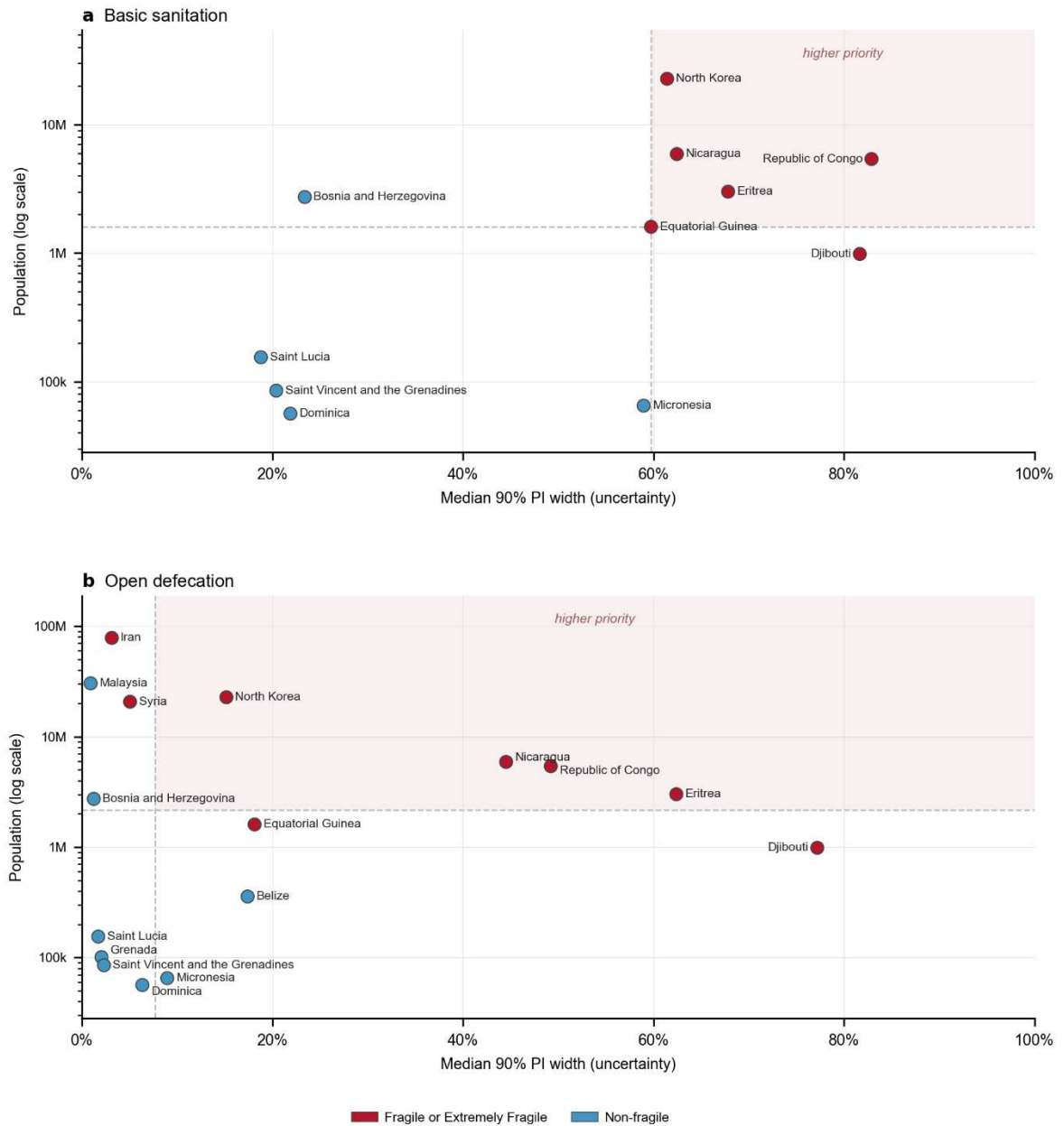


Figure S11 Priority countries for sanitation data collection. Countries lacking both a recent subnational household survey (2015–2024 training data) and a 2024 national JMP estimate, plotted by prediction uncertainty (median 90% prediction-interval width across districts) against total population (log scale WorldPop, 2024) for (a) basic sanitation and (b) open defecation. Point colour denotes fragility status (red = fragile or extremely fragile; blue = non-fragile using OECD classification). Dashed lines mark the median of each axis; countries in the shaded

upper-right quadrant, above-median uncertainty and above-median population, are the highest-priority candidates for new data collection.

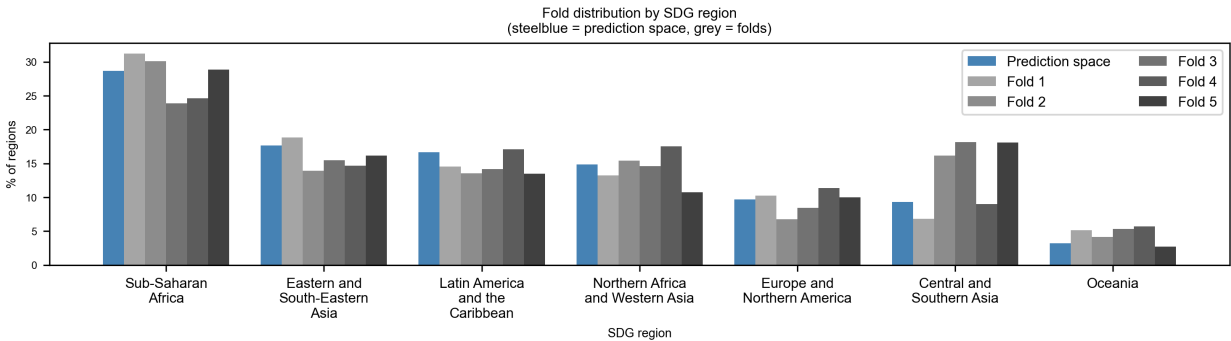


Figure S12 Distribution of admin-1 regions by SDG region across 130 low- and middle-income regions (blue) and five cross-validation folds (grey).

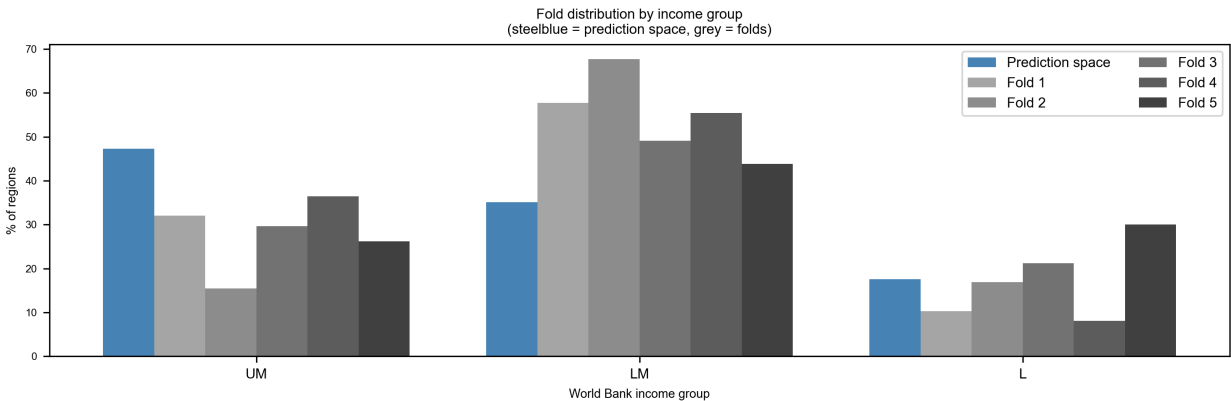


Figure S13 Distribution of admin-1 regions by income group across 130 low- and middle-income regions (blue) and five cross-validation folds (grey).

Table S1 Overview of geospatial covariates sampled per survey region categorized into: Climate; HP (human presence); S (hydrology and soil); TC (terrain and vegetation) and region size

Covariate name	Description	Category	Unit	Temporal resolution	Source
CGIAR_Aridity_Index	Global AI (Aridity Index) represents ratio between rainfall and vegetation water demand.	Climate	AI Value	Static	(56)
CGIAR_PET	Global PET (Potential Evapotranspiration)	Climate	PET Value (mm)	Static	(56)
CHELSA_BIO_Annual_Mean_Temperature	Bio1 = mean annual daily mean air temperatures averaged over 1 year	Climate	°C * 10	Static	(57)
CHELSA_BIO_Annual_Precipitation	Bio12 = Accumulated precipitation amount over 1 year	Climate	mm	Static	(57, 58)
CHELSA_BIO_Precipitation_Seasonality	Bio15 = The Coefficient of Variation is the standard deviation of the monthly precipitation estimates expressed as a percentage of the mean of those estimates (i.e. the annual mean)	Climate	Unitless	Static	(57, 58)
CHELSA_BIO_Precipitation_of_Coldest_Quarter	Average monthly precipitation during the coldest 3-month period of the year	Climate	kg m-2 month-1	Static	(57, 58)
CHELSA_BIO_Precipitation_of_Driest_Month	Minimum monthly precipitation total	Climate	kg m-2 month-1	Static	(57, 58)
CHELSA_BIO_Precipitation_of_Driest_Quarter	Average monthly precipitation during the driest 3-month period of the year	Climate	kg m-2 month-1	Static	(57, 58)

CHELSA_BIO_Precipitation_of_Warmest_Quarter	Average monthly precipitation during the warmest 3-month period of the year	Climate	kg m-2 month-1	Static	(57, 58)
CHELSA_BIO_Precipitation_of_Wettest_Month	Maximum monthly precipitation total	Climate	kg m-2 month-1	Static	(57, 58)
CHELSA_BIO_Precipitation_of_Wettest_Quarter	Average monthly precipitation during the wettest 3-month period of the year	Climate	kg m-2 month-1	Static	(57, 58)
CHELSA_BIO_Temperature_Annual_Range	Annual temperature range calculated as bio05 – bio06; measures amplitude between warmest and coldest months	Climate	°C	Static	(57, 58)
CHELSA_BIO_Temperature_Seasonality	Temperature seasonality given by the standard deviation of mean monthly temperatures	Climate	°C/100	Static	(57, 58)
CIFOR_Tropical_PeatlandExtent	Tropical and subtropical peatland extent	S	Unitless	Static	(59)
CSP_Global_Human_Modification	Cumulative measure of human modification of terrestrial lands.	HP	0-1 km2	Static	(60)
ConsensusLand_CoverClass_Barren	Percentage of the pixel area covered by barren ground	TC	percentage (0-100)	Static	(61)
ConsensusLand_CoverClass_Cultivated_and_Managed_Vegetation	Percentage of the pixel area covered by cultivated and managed vegetation (i.e., agriculture)	HP	percentage (0-100)	Static	(61)
ConsensusLand_CoverClass_Deciduous_Broadleaf_Trees	Percentage of the pixel area covered by deciduous broadleaf trees	TC	percentage (0-100)	Static	(61)
ConsensusLand_CoverClass_	Percentage of the pixel area covered by	TC	percentage (0-100)	Static	(61)

Evergreen_Broadleaf_Trees	evergreen broadleaf trees				
ConsensusLandCoverClass_Evergreen_Deciduous_Needleleaf_Trees	Percentage of the pixel area covered by deciduous needleleaf trees	TC	percentage (0-100)	Static	(61)
ConsensusLandCoverClass_Herbaceous_Vegetation	Percentage of the pixel area covered by herbaceous vegetation	TC	percentage (0-100)	Static	(61)
ConsensusLandCoverClass_Mixed_Other_Trees	Percentage of the pixel area covered by mixed other trees	TC	percentage (0-100)	Static	(61)
ConsensusLandCoverClass_Open_Water	Percentage of the pixel area covered by open water	S	percentage (0-100)	Static	(61)
ConsensusLandCoverClass_Regularly_Flooded_Vegetation	Percentage of the pixel area covered by regularly flooded vegetation	TC	percentage (0-100)	Static	(61)
ConsensusLandCoverClass_Shshrubs	Percentage of the pixel area covered by shrubs	TC	percentage (0-100)	Static	(61)
ConsensusLandCoverClass_Snow_Ice	Percentage of the pixel area covered by permanent snow and ice	Climate	percentage (0-100)	Static	(61)
ConsensusLandCoverClass_Urban_Builtup	Percentage of the pixel area covered by urban/built-up human areas	HP	percentage (0-100)	Static	(61)
ConsensusLandCover_Human_Development_Percentage		HP	percentage (0-100)	Static	(61)
CrowtherLab_Tree_Density	Tree_Density	TC	stems per pixel	Static	(62)
EarthEnvCloudCover_CloudForestPrediction	Relative occurrence rate of tropical montane cloud forests estimated using an	TC	Unitless	Static	(63)

	inhomogeneous point process model				
EarthEnvTexture_CoOfVar_EVI	Normalized dispersion of EVI	TC	Unitless (≥ 0)	Static	(63)
EarthEnvTexture_Contrast_EVI	Exponentially weighted difference in EVI between adjacent pixels	TC	Unitless (≥ 0)	Static	(63)
EarthEnvTexture_Correlation_EVI	Correlation (EVI Heterogeneity); Linear dependency of EVI on adjacent pixels	TC	Unitless (≥ -1 ; ≤ 1)	Static	(63)
EarthEnvTexture_Dissimilarity_EVI	Difference in EVI between adjacent pixels	TC	Unitless (≥ 0)	Static	(63)
EarthEnvTexture_Entropy_EVI	Disorderliness of EVI	TC	Unitless (≥ 0)	Static	(63)
EarthEnvTexture_Evenness_EVI	Evenness (EVI Heterogeneity); Evenness of EVI	TC	Unitless (≥ 0 ; ≤ 1)	Static	(63)
EarthEnvTexture_Homogeneity_EVI	Similarity of EVI between adjacent pixels	TC	Unitless (≥ 0 ; ≤ 1)	Static	(63)
EarthEnvTexture_Maximum_EVI	Dominance of EVI combinations between adjacent pixels	TC	Unitless (≥ 0 ; ≤ 1)	Static	(63)
EarthEnvTexture_Range_EVI	Range (EVI Heterogeneity); Range of EVI	TC	Unitless (≥ 0)	Static	(63)
EarthEnvTexture_Shannon_Index	Diversity of EVI	TC	Unitless (≥ 0 ; $\leq \ln(\text{max \# of different EVI})$)	Static	(63)
EarthEnvTexture_Simpson_Index	Diversity of EVI	TC	Unitless (≥ 0 ; $\leq 1 - 1/(\text{max \# of different EVI})$)	Static	(63)
EarthEnvTexture_Std_EVI	Dispersion of EVI	TC	Unitless (≥ 0)	Static	(63)
EarthEnvTexture_Uniformity_EVI	Orderliness of EVI	TC	Unitless (≥ 0 ; ≤ 1)	Static	(63)

EarthEnvTexture_Variance_EVI	Dispersion of EVI combinations between adjacent pixels	TC	Unitless (≥ 0)	Static	(63)
EarthEnvTopoMed_1stOrderPartialDerivEW	1st Order Partial Derivative of the Elevation (with reference to the East-West / X direction)	TC	m/m	Static	(64)
EarthEnvTopoMed_1stOrderPartialDerivNS	1st Order Partial Derivative of the Elevation (with reference to the North-South / Y direction)	TC	m/m	Static	(64)
EarthEnvTopoMed_Eastness	Eastness of a slope	TC	eastness index (-1 to 1)	Static	(64)
EarthEnvTopoMed_Elevation	Elevation (in meters)	TC	meters	Static	(64)
EarthEnvTopoMed_Roughness	Roughness is expressed as the largest inter-cell difference of a focal cell and its 8 surrounding cells	TC	m	Static	(64)
EarthEnvTopoMed_Slope	slope and terrain curvatures (defined as a measure of the concavity and convexity, or convergence and divergence)	TC	degrees ($^{\circ}$)	Static	(64)
EarthEnvTopoMed_TerrainRuggednessIndex	Terrain Ruggedness Index (TRI) is the mean of the absolute differences in elevation between a focal cell and its 8 surrounding cells	TC	m	Static	(64)
EarthEnvTopoMed_TopoPositionIndex	Topographic Position Index	TC	m	Static	(64)
EsaCci_BurntAreasProbability	Probability of Burnt Areas occurrence, detailed on a monthly frequency, based on	HP	0-100%	Static	(65)

	observations over the 2001-2019 period. Proportion of likely burned occurrence mapped.				
FanEtAl_Depth_to_Water_Table_AnnualMean	Mean annual depth of the water table on the terrestrial land surface (in m below land surface)	S	m below land surface	Static	(66)
FanEtAl_Depth_to_Water_Table_AnnualSD	Standard deviation of depth of the water table on the terrestrial land surface	S	m below land surface	Static	(66)
GHS_Population_Density	Distribution and density of population, expressed as the number of people per cell.	HP	Number of people per cell	Dynamic	(67, 68)
GLW3_Ruminants_Distribution_downsampled10km	Global distribution of ruminants (cattle, buffaloes, horses, sheep, goats)	HP	Number of animals per pixel	Static	(69)
GPWv4_Population_Density	Gridded Population of the World, Version 4 (GPWv4): UN-Adjusted Population Density for 2015 and 2020	HP	persons/km ²	Static	(70)
GiriEtAl_Mangroves_Extent	Status and distributions of global mangroves	TC	percentage (0-100)	Static	(71)
modis_evi, MODIS_EVI	Enhanced Vegetation Index (EVI) generated from the Near-IR, Red and Blue bands of each scene, and ranges in value from -1.0 to 1.0.	TC	Unitless	Dynamic & Static	(72)
modis_ndvi, MODIS_NDVI	Normalized Difference Vegetation Index generated from the Near-IR and Red bands of each scene as (NIR - Red) / (NIR + Red), and ranges in value from -1.0 to 1.0	TC	Unitless	Dynamic & Static	(72)

MODIS_NPP	Annual net primary productivity	TC	0.0001 * kg°C/m ²	Static	(73)
PelletierEtAl_Soil AndSedimentary DepositThicknesses	Average soil and sedimentary deposit thicknesses (unconsolidated material)	S	m	Static	(74)
SG_Absolute_depth _to_bedrock	Absolute depth to bedrock predicted using the global compilation of soil ground observations	S	cm	Static	(75)
SG_Bulk_density_ 015cm	Bulk density (fine earth) at 0.15m	S	kg / m-3	Static	(75)
SG_Depth_to_ bedrock	Depth to Bedrock (up to 200cm)	S	cm (up to 200)	Static	(75)
SG_H2O_Capacity_ 015cm	Derived available soil water capacity (volumetric fraction) until wilting point for depth 15 cm	S	percentage (0-100)	Static	(75)
SG_Saturated_H2O _Content_015cm	Volumetric water content at or near saturation (0 kPa suction) for the 0–15 cm soil depth	S	percentage (0-100)	Static	(75)
TootchiEtAl_ WetlandsRegularly Flooded	Regularly flooded wetlands, including regularly flooded land and lakes	S	binary or fraction of pixel	Static	(76)
WCS_Human_Foot print_2009	Quantifies cumulative impact of direct pressures on nature from human activities considering the extent of built environments, crop land, pasture land, human population density, night-time lights, railways, roads, and navigable waterways.	HP	Unitless	Static	(77)

CHIRPS_annual_precipitation	Annual precipitation	Climate	mm	Dynamic	(78)
ERA5_temperature_2m	Temperature of air at 2m above the surface of land, sea or inland waters.	Climate	Kelvin	Dynamic	(79)
ghsl_built_surface	Built-up surface per grid cell	HP	m2	Dynamic	(67, 80)
ghsl_population	Spatial distribution of residential population	HP	Population count by epoch	Dynamic	(67, 80)
ghsl_urban_frac	Rural-urban classification, applying the "Degree of Urbanisation"	HP	smod_code Class Table	Dynamic	(67, 80)
jrc_building_height	Average building height per grid cell	HP	m	Static	(67, 80)
map_friction	Land-based travel speed for all land pixels between 85 degrees north and 60 degrees south for a nominal year 2019	HP	min/m	Static	(81)
runoff_max_annual_max	Maximum runoff value each month	S	m	Dynamic	(79)
runoff_min_annual_min	Minimum runoff value each month	S	m	Dynamic	(79)
temperature_2m_max_annualmax	Maximum annual temperature at 2m	Climate	K	Dynamic	(79)
viirs_average	Average DNB radiance values.	HP	nanoWatts/sr/c m^2	Dynamic	(82)
worldpop	Estimated population density	HP	Number of inhabitants per cell	Dynamic	(83)
worldpop_sum	Estimated population sum	HP	Total number of people	Dynamic	(83)

ghsl_population_ sum	Estimated population sum	HP	Number of inhabitants per cell	Dynamic	(67, 80)
area_km2	Size of corresponding administrative region, calculated using the equal earth projection	Region size	km2	Static	(84)

Table S2 Country-level socio-economic and governance (SEG) and sanitation indicators and their sources.

Variable	Category	Temporal resolution
GDP per capita (constant 2015 USD) (49)	SEG	Annual
Secondary education duration (years) (49)	SEG	Annual
Wastewater collection, treatment and reuse (%) (35)	Sanitation	2015
Worldwide Governance Indicators (government effectiveness, control of corruption, political stability, regulatory quality, rule of law, voice and accountability)(50, 51)	SEG	Annual
National basic sanitation coverage and open defecation estimate (7)	Sanitation	Annual

Supplementary Files

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

- [DataS1.xlsx](#)
- [DataS2.xlsx](#)
- [DataS3.xlsx](#)
- [DataS4.xlsx](#)
- [DataS5.xlsx](#)