

Planetary Prediction Engine: Autonomous Geospatial Prediction via Intelligent Data Selection and Foundation Model Embeddings

Evelyn Ma^{*,1}, Rama Kumar Pasumarthi^{*,†,1}, Kishwar Shafin¹, Mandar Sharma¹, Mimi Sun¹, Hamed Sadeghi¹, Dav M. Ebengo², Onesime Mbulayi², Rouslan Solomakhin¹, John Wamburu¹, William Ogallo¹, Aisha Walcott-Bryant¹, Sanxing Chen¹, Arbaaz Muslim¹, Yael Mayer¹, Ronald Ho¹, Roy Lee¹, Ruth Alcantara¹, Abdoulaye Diack¹, Monica Bharel¹, Lambert Rosique¹, Jeremy Amezcua-Droz¹, Christopher Haire¹, James Manyika¹, Yossi Matias¹, Niv Efron¹, Gautam Prasad^{**,1} and Shravya Shetty^{**,1}

^{*}Equal first authors (alphabetical order), ^{**}Co-last authors, ¹Google Research, ²Institut National de Recherche Biomédicale, Democratic Republic of Congo

Addressing critical global challenges, from food security and disaster risk to disease outbreaks and socio-economic vulnerability, demands high-fidelity geospatial modeling. However, building predictive planetary models remains bottlenecked by a fragmented data ecosystem, requiring manual data retrieval, multimodal data curation and fusion along with iterative model selection. As part of Google Earth AI, we present the Planetary Prediction Engine (PPE), an autonomous AI system that executes this end-to-end workflow directly from natural-language queries. PPE synthesizes multimodal datasets on the fly, retrieving spatiotemporally relevant covariates across open-web and Earth observation platforms (Data Commons, Google Earth Engine) and fusing them with geospatial foundation model embeddings (PDFM, AlphaEarth). Simultaneously, it searches over task-tailored model architecture families with automated overfitting guards. Across diverse tasks, geographies, and scientific domains, PPE consistently outperforms state-of-the-art or manually tuned expert baselines. For US spatial regression, PPE improves mean R^2 across 21 CDC health indicators (76.8% vs. 60.0%), FEMA national risk indices (64.9% vs. 60.0%), and the Social Vulnerability Index (66.2% vs. 58.6%). For spatial downscaling in data-scarce settings, PPE integrates localized proxies to double baseline accuracy in Nigerian food security indicators (R^2 of 66.1% vs. 31.5%). For epidemiological nowcasting of the 2026 DRC Bundibugyo Ebola outbreak, PPE achieves a Recall@10 of 83.3% (identifying 15 of 18 newly invaded health zones across five weekly forecasts), a +10.3 percentage-point improvement over the public state-of-the-art modeling (~73%). By combining autonomous multimodal planetary data discovery with targeted model optimization, PPE lowers the technical barrier to planetary-scale analytics, enabling rapid, customized, expert-level deployment.

1. Introduction

1.1. Motivation and Challenges in Planetary Analytics

Addressing pressing global crises such as mapping socioeconomic inequality, mitigating regional food insecurity, and tracking active disease outbreaks requires high-fidelity, real-time geospatial modeling. Accurate predictions allow humanitarian organizations, governments, and public health agencies to make critical, time-sensitive decisions. For instance, during an active viral outbreak, predicting spatial transmission corridors enables public health organizations to prioritize vaccine distribution and deploy mobile clinics to high-risk zones. Similarly, downscaling food consumption indicators to local administrative units allows agricultural agencies to target relief efforts to vulnerable communities [9, 24]. Historically, translating raw Earth observations and demographic data into these downstream interventions has relied on distinct, manual modeling paradigms. These

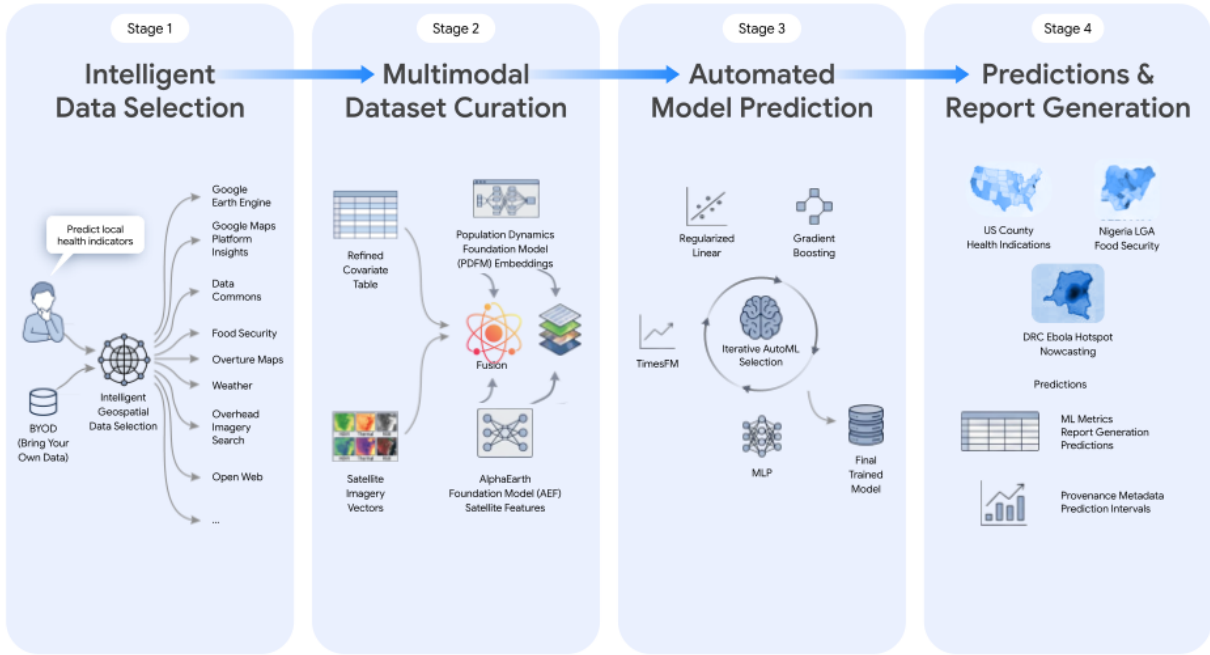


Figure 1 | The **Planetary Prediction Engine**’s end-to-end workflow. The system decomposes the predictive workflow into three modular stages: (1) Intelligent data selection, (2) dataset curation, and (3) AutoML & prediction to produce the predictions and report. Off-the-shelf LLMs serve as orchestrators within each stage.

include spatial regression and super-resolution downscaling to estimate socioeconomic indicators from satellite imagery [37, 80], as well as epidemiological nowcasting and spatial transmission modeling to correct for surveillance lags and simulate anisotropic transmission via Bayesian smoothing [49] and spatial flux equations [61, 71]. However, executing these predictive workflows is notoriously labor-intensive, creating a severe bottleneck when rapid deployment is critical. For example, a geospatial predictive modeling workflow for epidemiological nowcasting can have more than 700 steps across data selection, curation and model optimization. For any given socio-environmental task, specialized teams must manually navigate a highly fragmented geospatial data ecosystem. They must conduct rigorous domain research to identify relevant proxy signals, retrieve and clean data from decentralized public and private repositories, and fuse these covariates with domain-specific geospatial foundation models [1, 13, 36]. These models, such as the Population Dynamics Foundation Model (PDFM) [1] for demographics and AlphaEarth [8] for satellite-derived land-use semantics, yield expressive, high-dimensional embeddings that encode latent socio-environmental patterns. Finally, researchers must design model architectures that handle spatial dependencies while enforcing strict validation guardrails to prevent spatial target leakage [50, 59]. Because this pipeline requires deep domain expertise and significant manual engineering, developing these models can take weeks, hindering immediate humanitarian and policy response.

Existing automated machine learning (AutoML) frameworks [21, 66] are poorly suited to address these challenges. Recent AI-driven scientific discovery systems such as Empirical Research Assistance (ERA) [3] and AlphaEvolve [53] have achieved expert-level results across scientific domains by coupling LLMs with program search to optimize a pre-specified quality metric; however, they require a well-defined objective and curated task formulation. Similarly, while recent work has explored LLM-based autonomous agents for scientific discovery, data science, automating code generation, iterative debugging, and hypothesis formulation [28, 33, 45, 76], these frameworks are

typically restricted to executing code on clean, pre-curated tabular datasets or software engineering tasks. They lack the specialized capabilities needed to autonomously discover and curate large scale geospatial data on the fly, align and fuse heterogeneous geospatial embedding vectors (PDFM, AlphaEarth) via research-grounded workflows, or parameterize mechanistic epidemiological models. Consequently, planetary-scale analytics remains heavily restricted by the availability of specialized human engineering teams, leaving critical prediction tasks in data-scarce or crisis-prone regions under-addressed. We defer thorough discussion on related work to Appendix D.

1.2. The Planetary Prediction Engine Framework

To address these limitations, we propose the Planetary Prediction Engine (PPE), an autonomous AI architecture that translates a natural-language query into an executed geospatial model (Figure 1). A central goal of the Planetary Prediction Engine is rapid and accessible model building: by abstracting away data discovery, feature fusion, and hyperparameter tuning, the system enables non-experts to instantly build predictive models for their specific objectives. The Planetary Prediction Engine features three modular stages: Intelligent Data Selection, Multimodal Dataset Curation, and Automated Model Building and Prediction. Guided by LLM orchestrators, the system dynamically conducts signal discovery, retrieving raw covariates from both the open web and selecting appropriate Earth AI tool calls within each stage’s predefined tool set.

Importantly, task-type inference occurs early in Stage 1 during the initial query parsing: the system identifies the predictive paradigm (e.g., spatial regression, super-resolution, spatial transmission, or epidemiological nowcasting) from the user’s natural-language prompt and propagates these task-specific constraints to all downstream data retrieval sub-stages. This ensures that, for example, spatial transmission tasks trigger specialized mobility matrix construction and point-of-interest distance computation, while standard spatial regression tasks focus on static demographic covariate retrieval.

We benchmark the system across a multidimensional matrix covering varying geospatial predictive modeling tasks (spatial regression, super-resolution, epidemiological nowcasting), geographies (Global North, Global South), and thematic domains (public health, environmental vulnerability, epidemiology).

The system uses frontier large language models as black-box orchestrators within each stage. The LLMs serve as natural-language interfaces for interpreting user queries and selecting appropriate tool calls within each stage’s predefined tool set. See Section 4 for more details.

1.3. Summary of key contributions

Our primary methodological contributions are:

- **Autonomous Modular Agent Architecture:** A fully automated end-to-end agent architecture for geospatial prediction and epidemiological nowcasting that, given a natural-language query, performs research-grounded signal discovery to dynamically fetch and prioritize multi-source data, performs feature engineering, and trains and evaluates models without manual intervention.
- **Task Auto-Identification & Objective Alignment:** Automatic identification of geospatial predictive tasks (spatial regression, super-resolution downscaling, spatial transmission modeling, epidemiological nowcasting) from high-level human specification, allowing for AutoML optimization over the agent-defined objective.

- **Intelligent Data Selection & Multimodal Curation:** An intelligent data selection protocol that combines heterogeneous geospatial embeddings (PDFM, AlphaEarth) with on-the-fly intelligently selected statistical covariates, systematically exploring feature combinations that a human practitioner might overlook, while enforcing strict Automated Target Leakage Mitigation to prevent target leakage.
- **Automated Model Building & Prediction:** An iterative model selection protocol that searches over diverse model families (regularized linear models, gradient-boosted trees, extreme gradient boosting, multi-layer perceptrons) and hyperparameters, incorporating a multi-layered Overfitting Guard and Self-Correction loop to select the optimal model to produce final test-set predictions.

Our empirical findings demonstrate expert-level performance across this matrix:

- **Epidemiological Transmission Prediction (2026 DRC Ebola Outbreak):** For real-time prediction of new disease transmission hotspots during the [May–July 2026 Bundibugyo ebolavirus outbreak](#), the Planetary Prediction Engine achieves a Recall@10 of 83.3%, correctly identifying 15 of 18 newly invaded health zones across five sequential weekly forecasts. This represents a +10.3 percentage point absolute improvement over the published state-of-the-art Bayesian modeling baseline [19] (73%), driven by fusing epidemiological signals with PDFM embeddings and intelligently selected geospatial covariates.
- **Super-Resolution Downscaling (Nigeria Food Security and US Socioeconomics):** For Food Security indicators, Coarse regional reporting often obscures local vulnerability. By integrating localized market shocks and microclimate anomalies, the Planetary Prediction Engine doubled baseline accuracy when downscaling food security from the provincial (ADM1) level to the Local Government Area (LGA/ADM2) level (R^2 66.1% vs. 31.5%). For Social Vulnerability Index (SVI), Planetary Prediction Engine presents a 26.6 percentage improvement over baseline (R^2 37.6% vs. 11.0%).
- **Spatial Regression (US Socioeconomics, Health, and Environmental Risks):** For Social Vulnerability Index (SVI), the Planetary Prediction Engine achieved a mean R^2 of 66.2%, delivering a 6.8 percentage point improvement over statistical covariates baselines (58.6%). For CDC Health variables, the system achieves a mean R^2 of 76.8%, significantly outperforming expert pipelines (60.0%). On FEMA Environment risk metrics, Planetary Prediction Engine outperforms expert baselines on the Socioeconomic & Composite indicators (R^2 66.9% vs. 61.1%) and performs on par with expert benchmarks across the broader suite of environmental targets.

Across tasks, the PPE achieves 12–94% relative R^2 improvement over standard baselines, from a 12% gain on SVI spatial regression (66.2% vs. 58.57%) to near-doubling on Nigeria food security downscaling (66.1% vs. 31.5%), and a 10.3 percentage point gain on epidemiological nowcasting (Recall@10).

2. Results

2.1. General Experimental Design and Problem Matrix

To rigorously validate the adaptability and robustness of the proposed framework, we evaluate it across three distinct predictive paradigms: (1) Mechanistic Nowcasting, (2) Super-Resolution Downscaling, and (3) Spatial Regression (Table 1). These paradigms cover a diverse spectrum of

socioeconomic, public health, ecological, and humanitarian variables tailored to reflect real-world geospatial applications. Table 1 outlines the multidimensional matrix used to benchmark the system across varying machine learning tasks, geographic contexts, and thematic domains. The technical specifications, spatial granularities, and evaluation partitions for these benchmarks manifest highly differentiated distributions, reflecting dynamic real-world constraints.

Table 1 | Summary of Evaluation Benchmarks and Dataset Specifications. Overview of target variables, geographic scopes, and spatial resolutions across the three evaluation paradigms.

ML Paradigm	Geographic Scope	Thematic Domain	Benchmark Dataset	Prediction Targets	Train / Test Granularity	Train / Test Size
Nowcasting	Global South (DRC)	Humanitarian Epidemics	DRC Ebola Outbreak Tracking	one-week confirmed caseloads increase	Admin 3 health zones	7 weeks $\times$ 519 zones / 5 weeks $\times$ 519 zones
Super-Resolution Downscaling	Global South (Nigeria)	Food Security	Nigeria FCG Insecurity	FCG score	Admin 1 $\rightarrow$ Admin 2 (LGA)	30 states $\times$ 40 months / 581 LGAs $\times$ 40 months
	Global North (US)	Social Vulnerability	SVI	5 index scores	County $\rightarrow$ ZCTA	$\sim$ 3k / $\sim$ 33k
Spatial Regression	Global North (US)	Public Health	CDC Health	21 health variables	Census Tract	$\sim$ 67k / $\sim$ 17k
	Global North (US)	Environmental Risk	FEMA NRI	21 risk scores	Census Tract	$\sim$ 67k / $\sim$ 17k
	Global North (US)	Social Vulnerability	SVI	5 index scores	County	$\sim$ 2.4k / $\sim$ 0.6k

The *Nowcasting* paradigm expanded beyond stationary datasets: this module simulates the spatiotemporal trajectory of the May–July 2026 DRC Ebola Outbreak (BDBV virus) across all 519 health zones in DRC. Driven by dynamic sparsity and humanitarian urgency, the framework fuses official registry logs from National Institute of Biomedical Research (INRB) [34], OpenStreetMap infrastructure algorithms [29, 46], and Earth Engine climatology forcings [25] (e.g., ERA5 [30], WorldPop [63]). This enables stress-testing the agentic pipeline’s capacity to maintain predictive fidelity during time-sensitive, anisotropic crisis events. We benchmark our model against epidemiological predictions published by INRB [48], and defer detailed benchmark specifications to Appendix A and full ablation results to Appendix F.1.

Super-Resolution Downscaling explicitly engineered to validate model efficacy under multiscale administrative pooling where models must capture coarse signals to predict fine-grained targets. On the US SVI benchmark, downstream models are trained on coarse county-level covariates ($\sim 3k$ samples) but tested against high-resolution census tracts ($N \approx 84k$). Broadening applicability to data-scarce Global South contexts, we predict Nigerian Food Security by estimating Food Consumption Group (FCG) metrics via dynamic pooling of World Bank price indices, rainfall/NDVI vegetation proxies, and news sentiment datasets. For the Nigeria benchmark, the system is trained at the state level (ADM1, $N=30$) and evaluated at the Local Government Area level (ADM2/LGA, $N=581$).

Spatial Regression applied to the Global North (United States) to address environmental hazards and chronic disease. This includes the CDC health dataset (21 health indicators, e.g., obesity, diabetes) [26] and the FEMA National Risk Index (21 environmental indices) [20], both aggregating approximately 84k fine-grained census tracts under a conventional 80:20 random train/test split, which aligns with settings of public SOTA [5]. Conversely, the Social Vulnerability Index (SVI) module predicts 5 vulnerability scores at a coarser, county-level granularity ($N \approx 3k$) [10, 22].

Detailed descriptions of datasets and benchmark formulations are provided in Appendix A and Appendix B, with full benchmark evaluations provided in Appendix F.

Baselines and Ablation Structure. To rigorously quantify the contribution of each component, we structure our empirical evaluations across four standardized ablation tiers:

- Baseline / SOTA: Traditional manual expert pipelines (hand-curated features and grid-searched models by domain experts) or standalone foundation embedding baselines.
- PPE (Covariates): Models trained exclusively on raw statistical covariates (e.g., Data Commons, World Bank, Census) without foundation model embeddings.
- PPE (Embeddings): Models trained on tabular covariates fused with geospatial foundation model embeddings (PDFM and/or AlphaEarth Foundation Model (AEF)). See appendix B for description on foundation model embeddings.
- PPE (Full Stack): The complete autonomous system, incorporating Covariates + Embeddings + Intelligent Data Selection (dynamic open-web discovery, learned covariates, OSRM physical mobility networks, and nowcasting corrections where applicable).

For the predictive modeling approaches, we characterize the data preprocessing (DP) as follows:

- (i) Baseline DP, which executes traditional spatial aggregation, epidemiological signal processing, human mobility matrix construction, and static vulnerability scoring (see Appendix A); to
- (ii) Planetary Prediction Engine Automatic Tabular DP, utilizing automated, end-to-end tabular data engineering (see Section 4.1.1 and 4.2); and ultimately to
- (iii) Planetary Prediction Engine Automatic Full DP, which synthesizes both automated tabular engineering and deep representation alignment for multi-modal embeddings (see Section 4.1.2 and 4.2).

2.2. Epidemiological Transmission and Nowcasting

Outbreak Hotspot Detection. We evaluate the performance of the Planetary Prediction Engine in predicting future transmission hotspots in an outbreak, which we define as previously uninfected health zones that subsequently experience a viral spillover event. Specifically, we report Recall@10, representing the proportion of newly infected zones successfully captured within our top 10 highest-risk predictions. On a one-week rolling forecast tracking the spatial invasion of Bundibugyo ebolavirus across the Democratic Republic of the Congo (DRC), Planetary Prediction Engine achieves a Recall@10 of 83.3%. We defer the detailed configurations of our spatial and temporal splits to Appendix F.1.

Table 2 summarizes the hotspot detection performance (Recall@10) across feature configurations on tracking the spatial invasion of Bundibugyo ebolavirus across the Democratic Republic of the Congo (DRC). Integrating geospatial covariates with baseline epidemiological covariates improves upon the SOTA baseline (77.8% vs. 73%), demonstrating the utility of auxiliary geographic indicators. Peak predictive performance is achieved by the complete Planetary Prediction Engine, where fusing epidemiological signals and geospatial covariates with pre-trained PDFM embeddings yields a Recall@10 of 83.3%. This represents a marked absolute gain of 10.3 percentage points over the state-of-the-art baseline. PPE predictions on total caseloads at frontline regions show high correlation with the ground-truth caseloads (Figure 2). These results validate the representational power of foundation model embeddings in capturing latent demographic structure and spatial connectivity during localized spillover events.

We defer the ablation results regarding model optimization, as well as comparison of mechanistic modeling vs spatial transmission regression to Appendix F.1.

2.3. High-Resolution Spatial Downscaling

In this section, we evaluate the performance of the geospatial prediction agent on Super Resolution tasks. We present performance on SVI and FCG benchmarks, covering domains of social economics

Table 2 | Model Performance (Recall@10) across Feature Configurations on tracking the spatial invasion of Bundibugyo ebolavirus across the Democratic Republic of the Congo (DRC).

System	Feature Configuration	Data Preprocessing (DP)	Recall@10 (%) [95% CI]
Baseline/ SOTA (Bayesian modeling)	Baseline Signals	Baseline DP	~ 73
PPE (Covariates)	Baseline Signals + Geospatial Covariates	Baseline DP + PPE automatic tabular DP	77.8 [54.8, 91.0]
PPE (Full Stack)	Baseline signals + PPE automatic full data selection	Baseline DP + PPE automatic full DP	83.3 [60.8, 94.2]

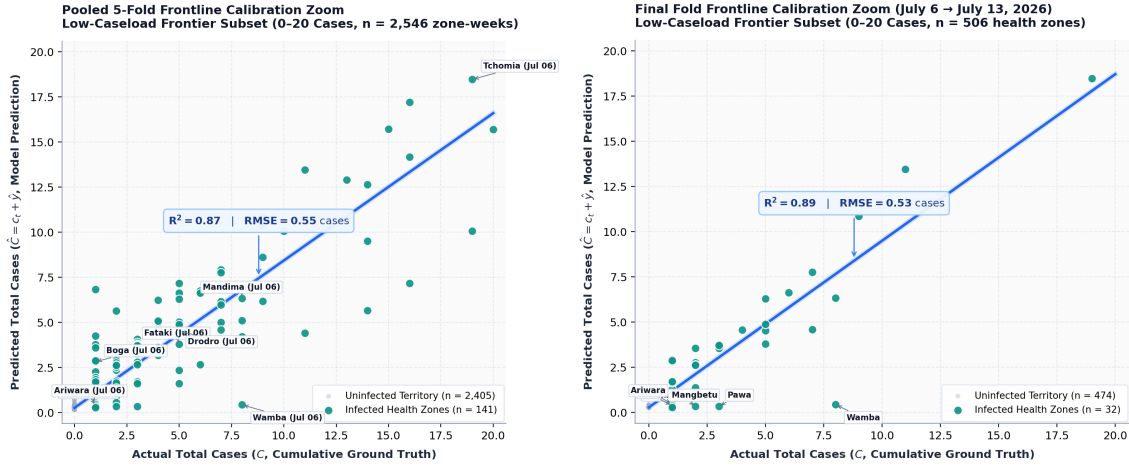


Figure 2 | Frontline Ebola Transmission prediction visualization: Comparison of predicted total caseloads against ground-truth caseloads.

and food security. We observe that the agent achieves significant outperformance over baselines.

2.3.1. Super-Resolution Downscaling in Nigeria for Food Security Indicators

We evaluate the Planetary Prediction Engine on downscaling food security indicators across Nigeria. In developing regions, food security assessments captured by metrics such as the Food Consumption Group (FCG) [72] and Integrated food security Phase Classification (IPC) [35]. Acute food insecurity phases are typically collected through representative surveys at the coarse State / Federal Capital (ADM1) level or Local Government Area (ADM2) level. (We will describe ADM1 with State level resolution for the sake of brevity.) However, humanitarian interventions require high-resolution, fine-grained estimates at the Local Government Area (LGA) or sub-LGA level to enable better targeted interventions [55]. In our benchmark, the system is trained on ADM1 state-level data (N=30) and evaluated at the ADM2 LGA level (N=581) on a monthly basis across 40 months, representing a substantial spatial downscaling challenge.

We implement the traditional spatial-downscaling approach, Macro-Covariates + Interpolation, as the baseline. The input features consist solely of calendar month timestamps encoded as first-order Fourier seasonal harmonics ($\sin(\frac{2\pi \cdot \text{month}}{12})$, $\cos(\frac{2\pi \cdot \text{month}}{12})$) and a linear year trend ($\text{Year} - 2022$). This formulation models the national agricultural harvest and lean cycle alongside multi-year macroeconomic drift. A gradient boosting model is trained on samples at a coarse administrative level. This baseline captures purely temporal and seasonal variance.

The Planetary Prediction Engine autonomously curates a rich multimodal feature space to drive

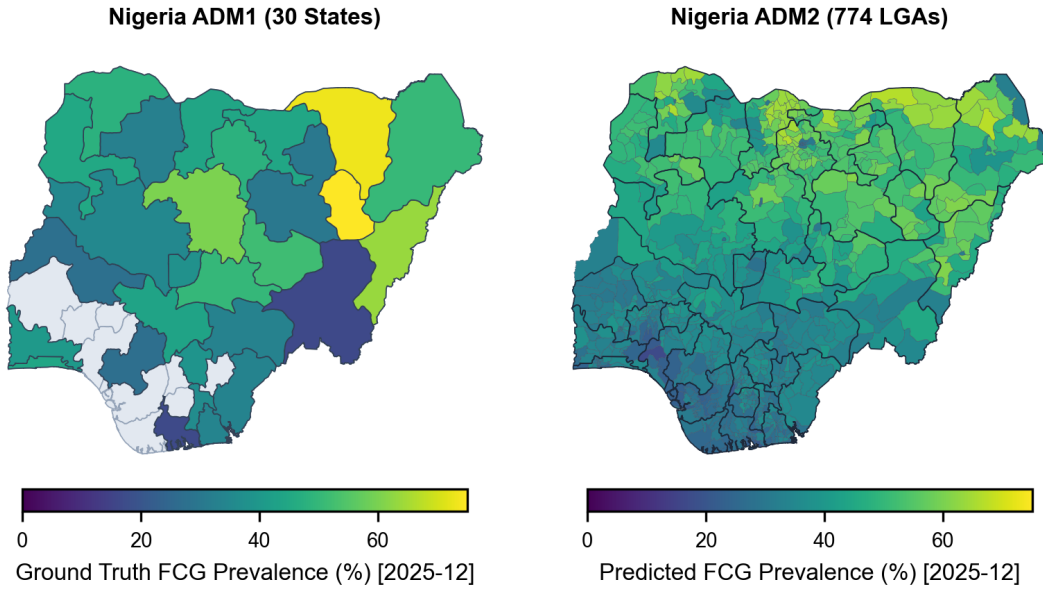


Figure 3 | Super-Resolution Food Security Downscaling in Nigeria (ADM1 State Level → ADM2 LGA Level)

the super-resolution downscaling model. It integrates several features including World Bank food price indices [74] (capturing localized food price index and food price inflation anomalies), World Food Programme (WFP) food insecurity metrics [39, 73], precipitation indicators [24], Normalized Difference Vegetation Index (NDVI) indicators [16], and VIIRS (Visible Infrared Imaging Radiometer Suite) nighttime lights (NTL) radiance [18]. These tabular data streams are fused with 330-dim PDFM [1] socioeconomic embeddings and 64-dim AlphaEarth satellite features [8]. Ground truth prevalence of insufficient food consumption was calculated as the proportion of households classified in the poor or borderline Food Consumption Group (FCG) from surveys conducted across Nigeria between August 2022 and December 2025 by the WFP, which are statistically representative at the ADM1 level. The household data provided by WFP is anonymized and does not contain any personal identifiable information. The coarse ADM1 FCG prevalence was used for model training and leave-one-state-out cross-validation, while localized ADM2 (LGA) ground truth was independently estimated using Multilevel Regression and Poststratification (MRP) [56] and kept blind to the model during training. The ADM2 MRP estimates were used for out-of-fold validation, where the Planetary Prediction Engine achieved a Mean Absolute Error (MAE) of 10.0% compared to 13.6% for the baseline model.

Table 3 demonstrates the improved downscaling accuracy achieved by the Planetary Prediction Engine, measured using R^2 measured against ADM1 level ground truth data with spatial cluster bootstrap 95% confidence interval (unit of resampling = state, $k=30$). While the baseline (relying on macro-covariates and basic interpolation) achieves an R^2 of 31.5%, and baseline + vegetation reach 60.1%, the complete Planetary Prediction Engine achieves an R^2 of 66.1%, doubling the baseline accuracy. The Planetary Prediction Engine’s intelligent data selection successfully captures agricultural seasonality, localized farming and drought shocks (via NDVI, vegetation index monthly, historical average, vegetation anomaly ratio), and urbanization and commercial infrastructure (via NTL), providing humanitarian organizations with actionable, high-fidelity vulnerability maps.

The Planetary Prediction Engine selected the following datasets:

- (i) **Macro-Covariates:** Temporal features of sine and cosine cyclical month encoding $[\sin(2\pi \cdot$

Table 3 | Model Performance (R^2) for Super-Resolution Food Security Downscaling (FCG) in Nigeria (ADM1 State Level $\rightarrow$ ADM2 LGA Level).

System	Feature Configuration	Overall R^2 [95%CI] (%)
Baseline	Macro-Covariates + Interpolation	31.5 [19.0, 39.3]
PPE (Covariates)	Macro-Covariates + nighttime lights (NTL)	49.8 [39.8, 57.9]
	Macro-Covariates + vegetation	60.1 [40.3, 72.8]
PPE (Full Stack)	Macro-Covariates + vegetation + NTL + Automatic Intelligent Selection	66.1 [55.9, 72.8]

$\text{month}/12), \cos(2\pi \cdot \text{month}/12)]$ and temporal trend $[\text{year} - 2022]$.

(ii) **Nighttime lights (NTL)**: The average nocturnal visible light radiance measured by the Day/Night Band (DNB) of the VIIRS instrument aboard the Suomi-NPP satellite.

(iii) **Vegetation**:

- (a) NDVI (Normalized Difference Vegetation Index): Derived from Sentinel-2 / Landsat surface reflectance using near-infrared (NIR) and red wavelengths.
- (b) Vegetation Index Monthly: The monthly observed vegetation index derived from the MODIS (Moderate Resolution Imaging Spectroradiometer) sensor aboard Terra and Aqua satellites.
- (c) Historical Climatological Baseline Vegetation Index: The long-term multi-year historical mean of the MODIS vegetation index for that specific calendar month.
- (d) Vegetation Index Quotient / Anomaly Ratio: The ratio of (b) / (c).

Planetary Prediction Engine performed a search across multiple machine learning architectures (Ridge Regression, Lasso Regression, ElasticNet, Random Forest, Gradient Boosting, XGBoost) and candidate datasets. It selected gradient boosting paired with temporal macro-covariates, NTL, and vegetation after discovering that this combination achieved the peak R^2 of 66.1%, whereas incorporating additional signals resulted in lower predictive accuracy.

In the side-by-side comparison (Figure 3), the left map displays the ground-truth food insecurity prevalence across 30 surveyed Nigerian states in December 2025, leaving unmonitored states blank. The right map presents our super-resolution model’s predictions across all 774 LGAs nationwide, which demonstrates how the framework can bridge gaps in data-sparse regions where ground surveys are absent, but satellite telemetry is ubiquitous.

2.3.2. Social Vulnerability Index

We employ the standard Macro-Covariates + Interpolation approach as the baseline. A regularized linear regression model is trained on county-level targets using geographic centroid coordinates (lat,lon) alongside one-hot administrative state fixed-effects.

To evaluate the cross-scale generalization of our framework under heterogeneous administrative boundaries, we formulate a super-resolution downscaling benchmark using the CDC’s Social Vulnerability Index (SVI). Standard socioeconomic vulnerability profiles are typically aggregated and published at the coarse county level ($\sim 3,000$ in the US) to maintain statistical representative validity and prevent privacy leakages. However, localized humanitarian, epidemiological, and policy interventions require high-resolution estimates at the ZIP-code level (approximated by ZIP Code Tabulation Areas, or ZCTAs). To rigorously simulate this scenario, our downstream models are trained exclusively on coarse, county-level statistical covariates but validated prospectively against

Table 4 | Model Performance (R^2) across Feature Configurations on SVI (Social Vulnerability Index) Benchmark Themes with Super-resolution tasks (from county-level to zipcode-level).

System	Feature Configuration	Mean R^2 [95% CI]
Baseline	Macro-Covariates + Interpolation	11.0 [10.0, 12.3]
PPE (Covariates)	Geospatial Covariates	25.6 [24.3, 26.9]
PPE (Embeddings)	PDFM	36.9 [36.0, 37.9]
PPE (Full Stack)	Covariates + PDFM	37.6 [36.7, 38.6]

fine-grained, ZIP-code-level targets.

Table 4 summarizes the super-resolution downscaling performance (R^2) across feature configurations on the US SVI benchmark. Resolving sub-county variation under administrative pooling presents a severe spatial super-resolution challenge due to extreme local socioeconomic heterogeneity. Standalone modalities, i.e., raw geospatial covariates (25.6% overall R^2) struggle to capture sub-county variance. A major performance breakthrough is achieved by fusing county-level covariates with our pre-trained 330-dimensional Population Dynamics Foundation Model (PDFM) embeddings, which drives the overall R^2 to 37.6%. This indicates that PDFM successfully encodes latent, cross-scale socioeconomic representations that remain robust to administrative boundary pooling.

2.4. Spatial Regression

In this section, we evaluate the performance of the geospatial prediction agent on spatial regression tasks. We present performance on CDC health [26], FEMA environment risk [20], and SVI benchmarks [10, 22], covering domains of public health, environment, and social economics. Detailed benchmark definitions are deferred to Appendix A. We observe that the agent demonstrates significant outperformance over SOTA baselines.

2.4.1. CDC Health Variables

Table 5 demonstrates the lift achieved by the Planetary Prediction Engine across the ablation tiers. While the previous state-of-the-art (SOTA) manual expert pipeline achieves a mean R^2 of 60%, the PPE’s intelligent data selection and multimodal fusion drive mean R^2 to 76.8%, a 23 percentage point improvement.

Table 5 | Model Performance (R^2) across Feature Configurations on 21 CDC Health Variables.

System	Feature Configuration	Mean R^2 [95% CI] (%)
Baseline / SOTA	Manual Expert Pipeline (PDFM + AEF)	60
PPE (Embeddings)	PDFM	59.7 [58.6, 60.9]
	PDFM+AEF	61.8 [60.6, 62.9]
PPE (Full Stack)	PDFM + AEF + Data Commons Covariates	76.8 [76.1, 77.6]

2.4.2. FEMA Environment Risk Variables

Table 6 evaluates the performance of Planetary Prediction Engine on 20 different county-level environmental and climate risk indices from FEMA against the performance previously reported by hand-curated models by experts [5].

We have split the 20 labels from FEMA into three different categories. The socioeconomic &

composite category contains targets that quantify human and institutional vulnerability, and overarching community risk. The atmospheric and climatological category contains hazards driven by meteorological, atmospheric and weather conditions. Finally, the geophysical and hydrological category covers hazards governed by geodynamic processes.

Given the FEMA target labels contain different modalities for risk assessment, from socioeconomic to hydrological hazard, we measured the impact of autonomous representation selection by conducting a systematic feature ablation study where the agent could choose the optimal feature configuration for each target. The agent could choose autonomously between the PDFM and AEF combined or in isolation, or it can combine them with Data Commons (DC) covariates where helpful. Overall, this autonomous feature ablation configuration allows Planetary Prediction Engine to achieve a higher mean R^2 (61.1%) compared to the hand-curated baseline’s mean R^2 (59.9%) over all 20 labels. Specifically in the socioeconomic and composite category, the feature ablation configuration achieves 66.9% mean R^2 compared to 61.1% mean R^2 of the hand-curated model.

We added the full Intelligent Data Selection (IDS) pipeline to be a part of the ablation suite to select the winning feature suite and we observed the performance further scales to a nationwide mean R^2 of 64.9% from 59.9% mean from the hand-curated model.

Beyond the category aggregate, noting standout single-target gains like Social Vulnerability ($R^2 = 67.6\%$ vs. 48.2% , a $+40.0\%$ gain) underscores the capability of autonomous multimodal feature selection and automated covariate discovery. This analysis overall demonstrates the generalizability of Planetary Prediction Engine to environmental and biophysical prediction targets.

Table 6 | Model Performance (R^2) across Feature Configurations on FEMA Environmental Risk Variables.

System	1. Socioeconomic & Composite	2. Atmospheric & Climatological	3. Geophysical & Hydrological	Full Nationwide Suite
Target Count	4	10	6	20
PDFM + AEF Manual expert Mean R^2 (%)	61.1	64.6	51.1	59.9
PDFM + AEF + DC (Feature ablation) Mean R^2 [95% CI] (%)	66.9 [66.2, 67.5]	64.3 [63.6, 65.0]	51.7 [49.3, 53.7]	61.1 [60.2, 61.7]
PDFM + AEF + DC + Intelligent Data Selection Mean R^2 [95% CI] (%)	69.4 [68.8, 70.0]	68.3 [67.5, 68.9]	56.2 [54.1, 57.8]	64.9 [64.1, 65.5]

Full results including which features were used with target-specific R^2 is reported in Appendix F.2. An example prompt on how the ablation was performed is also provided in Appendix E.3.2.

2.4.3. Social Vulnerability Index (SVI) Spatial Regression

To maintain methodological consistency with existing expert spatial regression benchmarks on CDC health and FEMA risk indicators[5], we establish the SVI baseline as the optimal performance attained across all unimodal embedding-only configurations.

Table 7 summarizes predictive R^2 across feature configurations, highlighting the benefits of multimodal fusion for social vulnerability modeling. Among standalone modalities, latent PDFM em-

beddings ($R^2 = 58.6\%$) outperform both explicit covariates (51.6%) and AEF signals (45.1%). Ultimately, the complete Planetary Prediction Engine pipeline (Covariates + PDFM + AEF + Intelligent Selection) achieves peak performance ($R^2 = 66.2\%$).

Table 7 | Model Performance (R^2) across Feature Configurations on SVI Benchmark Themes (Spatial Regression at County-Level).

System	Feature Configuration	Mean R^2 [95% CI] (%)
Baseline	Foundation Model Signals	60.3 [55.2, 64.9]
PPE (Covariates)	Geospatial Covariates	50.2 [44.2, 55.6]
PPE (Full Stack)	Covariates + PDFM	66.2 [61.6, 70.4]

This significant margin confirms the synergistic value of the full stack of synthesizing structured socio-demographic contexts, latent spatial semantics, and physical indicators to capture the multi-dimensional complexity of geospatial vulnerability.

3. Discussion

Based on analysis across spatial regression, super-resolution and epidemiological nowcasting, we look at key insights across these evaluations, along with limitations and future work.

3.1. Key Insights

Across the three tasks, three major insights emerge from our evaluation:

Multimodal fusion improves prediction quality. Across all benchmarks, the combination of explicit geospatial covariates with latent foundation model embeddings (PDFM, AlphaEarth) outperforms either modality in isolation. This confirms that pre-trained geospatial representations encode complementary information to traditional statistical covariates, and that their fusion, when mediated by appropriate feature engineering and leakage prevention, yields robust predictive gains.

Autonomous data curation closes the expertise gap. The Intelligent Data Selection pipeline, comprising grounded signal discovery, multi-repository retrieval, open-web search, and provenance-first prioritization, enables the system to assemble rich, task-specific covariate sets that rival or exceed those constructed by domain experts. This is particularly impactful in data-scarce settings (e.g., Nigeria food security, DRC Ebola), where the agent discovers localized proxy signals that would require substantial domain knowledge to identify manually.

Co-optimization of data and models is essential. Our ablation studies consistently show that neither intelligent data curation nor automated model selection alone achieves peak performance. The multiplicative interaction between what the model sees (curated, high-fidelity features) and how it learns (optimized architectures and hyperparameters) is the primary driver of the Planetary Prediction Engine’s performance advantage.

3.2. Multimodal Synergy in Spatial Regression

Our spatial regression evaluations in data-rich environments demonstrate a consistent performance lift across public health, environmental, and socioeconomic domains. On the 21 CDC health indicators, the Planetary Prediction Engine leverages multimodal fusion and autonomous model selection to achieve a mean R^2 of 76.8% , significantly outperforming the manual expert pipeline baseline of

60%. Similar improvements are observed for FEMA environmental risk variables, where the Planetary Prediction Engine reaches a mean R^2 of 64.9% compared to the baseline of 60%. For the Social Vulnerability Index (SVI), fusing explicit geospatial covariates with latent foundation model embeddings unlocks peak performance (mean R^2 of 66.2% across 5 themes), confirming that combining structured socio-demographic indicators with dense spatial representations yields a more robust understanding of regional vulnerability than standalone modalities.

3.3. Cross-Scale Generalization and Noise-Resolution Trade-offs

The super-resolution downscaling benchmarks validate the system’s ability to maintain predictive fidelity across heterogeneous administrative granularities. In data-scarce regions such as Nigeria, the Planetary Prediction Engine dynamically curates localized proxies, including food price anomalies and news sentiment, double the baseline accuracy when mapping food security at the LGA level (R^2 of 66.1% vs. 31.5%). However, cross-resolution projection [2] also uncovers an important noise-resolution trade-off. In the SVI super-resolution task from county to ZIP code level, adding high-resolution AlphaEarth Foundation (AEF) physical features to the covariate-PDFM stack causes a drop in overall performance (R^2 of 40.1% vs. 52.0%). This suggests that while satellite-derived terrain and land-use attributes provide rich localized context, they can introduce high-frequency noise or trigger spurious correlations at fine scales, potentially disrupting the generalization of broader socioeconomic proxies during administrative downscaling.

3.4. Trade-offs in Epidemiological Transmission and Nowcasting

Our experimental results demonstrate that combining intelligent data selection with automated model selection improves hotspot prediction accuracy. Relying solely on raw geospatial and dynamic epidemiological covariates yields sub-optimal performance. Integrating pre-trained PDFM embeddings with geospatial covariates provides an intermediate lift (Recall@10 of 77.8%), but performance remains bottlenecked by the passive ingestion of all features and non-optimized model configurations. The Planetary Prediction Engine agent resolves these limitations by jointly optimizing the feature input space (via Intelligent Data Selection) and the hypothesis space (via Model Search). This co-design drives Recall@10 to 83.3%, marking a 10.3 percentage point absolute gain over the published SOTA (73%). Also, see Appendix F.1 for comparison between spatial transmission modeling and epidemiological nowcasting in the context of the 2026 Ebola outbreak, which shows that spatial transmission modeling is better suited for predicting new hotspots compared to standard mechanistic SEIR models.

3.5. Limitations and Future Work

Firstly, the system currently relies on foundation model embeddings (PDFM, AlphaEarth) as frozen feature extractors; end-to-end fine-tuning of these representations for specific downstream tasks could yield further gains but would require careful regularization to avoid overfitting in small-sample regimes. Second, the noise-resolution trade-off observed in SVI super-resolution where high-frequency satellite features degrade cross-scale generalization highlights the need for adaptive feature selection mechanisms that account for the target spatial granularity. Third, while the Planetary Prediction Engine’s anti-leakage protocols (Feature Gate, Split-Isolated Imputation) provide strong safeguards, formal verification of causal direction filters remains an open challenge, particularly for targets with complex, bidirectional relationships to candidate covariates. Fourth, our epidemiological nowcasting evaluation is limited to a single outbreak; validation across diverse pathogens, geographies, and surveillance infrastructures would strengthen claims of generalization.

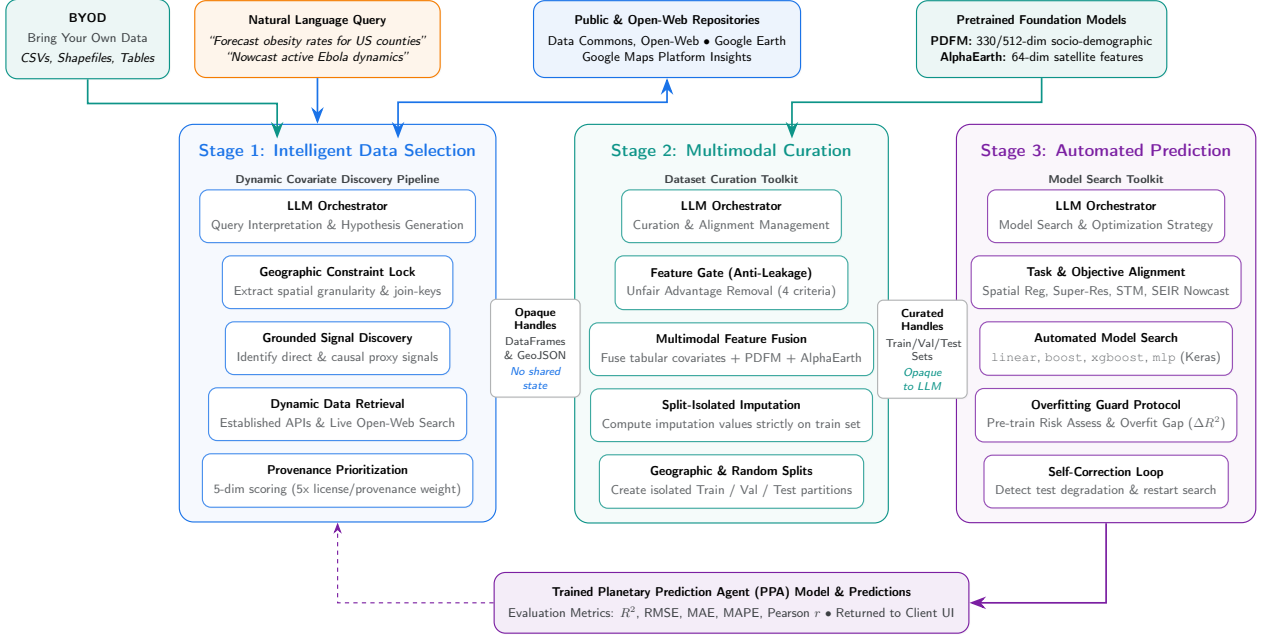


Figure 4 | The Planetary Prediction Engine’s multi-stage modular architecture.

Looking ahead, we identify several promising directions: (i) extending the framework to spatiotemporal forecasting with longer prediction horizons via temporal foundation models; (ii) scaling the Intelligent Data Selection pipeline to incorporate real-time streaming data sources (e.g., social media signals, mobility traces) for continuous model updating; and (iii) developing an ensemble architecture that combines spatial transmission models for frontier detection with mechanistic nowcasting models for resource allocation in operational outbreak response.

4. Methods

In this section, we introduce technical details of the three modular stages of the prediction workflow. The system follows rigorous modular software engineering principles: each stage operates on well-defined inputs and outputs, with no shared mutable state between stages. Data artifacts (DataFrames, GeoJSON geometries, mobility matrices) are passed between stages via opaque handles rather than serialized into LLM prompts, avoiding context-window limitations. All LLM calls use temperature = 0 for greater reproducibility.

4.1. Intelligent Data Selection

A distinguishing feature of the Planetary Prediction Engine is its Intelligent Data Selection stage: a pipeline that transforms a natural-language predictive query into a prioritized, join-ready multimodal DataFrame without manual data discovery, download, or curation. Based on the geographic and temporal constraints of the user’s labeled data, the system dynamically identifies, retrieves, scores, and assembles covariates at runtime for modeling.

4.1.1. Dynamic Covariate Discovery Pipeline

The pipeline is decomposed into six sequential sub-stages (See Appendix C for more details):

1. **Geographic Constraint Discovery:** Inspects the user-provided labeled training data to ex-

tract a geographic constraint (spatial granularity, join-key format, and temporal scope). 2. **Grounded Signal Discovery:** Formulates domain hypotheses to draft a structured Signal Guide of candidate covariates, distinguishing between *direct signals*, which are essential domain variables, and *proxy signals*, variables substituting for direct signals through well-established causal relationships. Every signal is validated against published literature and official reports.

3. **Established Repository Retrieval:** Systematically queries established, openly accessible geospatial repositories based on signal type: socio-demographic, economic, and pre-aggregated environmental statistics are retrieved from Data Commons [27]; raster environmental data requiring pixel-level aggregation are processed through Google Earth Engine [25]; and point-of-interest (POI) density metrics are obtained from the Google Maps Platform Insights.

4. **Ad-Hoc Open-Web Discovery:** For signals not resolved by established repositories, the system performs live open-web discovery for relevant datasets (CSV, GeoJSON, Parquet, etc.) across government portals (CDC, Census, WHO, UN OCHA HDX) and academic repositories (Zenodo, Harvard Dataverse), downloading and validating them programmatically.

5. **Provenance-First Prioritization:** Ranks datasets using a structured five-dimension scoring rubric (Provenance & License, Spatio-Temporal Fitness, Signal Alignment, Format Quality, Redundancy). Provenance and license openness carry the highest weight (5x), ensuring the assembled covariate matrix is composed primarily of openly licensed, institutionally authoritative data.

6. **DataFrame Assembly:** Merges all retained datasets into a single, join-ready covariate DataFrame, standardizing schemas (EPSG:4326 for geometries, ISO 8601 for dates) and generating a comprehensive data source audit and provenance metadata tracking report.

4.1.2. Multimodal Dataset Curation with Foundation Model Embeddings

To establish a robust, multimodal framework for geospatial modeling, the curation stage fuses the assembled tabular geospatial covariates (data sources deferred to Appendix B) with pre-trained geospatial foundation models. These features represent multi-dimensional geographical properties that collectively capture the socio-demographic, physical, industrial, and ecological dimensions of target regions. The agent is therefore able to leverage pre-trained deep neural representations to encode complex, non-linear geographical and demographic characteristics. These include Population Dynamics Foundation Models (PDFM) for socio-economic latent states (330/512-dim) and AlphaEarth Foundation models for satellite imagery semantics (64-dim).

4.2. Data preprocessing and Feature engineering

4.2.1. Automated Target Leakage Mitigation (Feature Gate)

To ensure the integrity of our empirical benchmarks and prevent unfair advantages [40, 41], the PPE implements a strict *Feature Gate* during dataset curation. Upon receiving the mathematical definition of the prediction target in the prompt, the agent evaluates every candidate covariate against four mandatory anti-leakage criteria:

Criteria 1 (No Mathematical Components): The covariate must not be a mathematical component, direct proxy, or sub-index used to calculate the ground truth target. (Example: Exclude "Median Rent" if predicting "Housing Affordability Index" which is income divided by rent).

Criteria 2 (No Synthetic/Shared Survey Leakage): The covariate and target must not rely on the exact same underlying survey data or imputation models. Furthermore, when predicting population-related targets, the system restricts covariates to non-enumerative, intensive socioeco-

nomic rates (e.g., Median Income) rather than enumerative counts (e.g., Count HousingUnit) to guarantee zero census enumeration leakage. (Example: Exclude Census age/income demographics if predicting a synthetic "Climate Vulnerability Score" derived from those same Census tables).

Criteria 3 (Causal Direction Filter): The covariate must be causally "upstream" (a driver, structural condition, or parallel confounder), not "downstream" (an effect, symptom, or response) of the target. (Example: Exclude "Number of Delayed Deliveries" when predicting "Traffic Congestion Levels").

Criteria 4 (Temporal Filter): The covariate feature must be from a time period before or during the target prediction window, preventing future information leakage.

4.2.2. Anti-Leakage Imputation

To ensure rigorous validation, the agent first quantifies the missingness rate for each candidate feature. Columns with insufficient coverage are discarded to mitigate overfitting to highly noisy or heavily imputed data. For the retained features, we implement a Split-Isolated Imputation protocol to prevent data leakage: imputation statistics (mean, median, or mode) are computed exclusively from the training partition and subsequently applied to the validation and test sets, thereby ensuring that downstream partitions represent strictly unseen data.

4.2.3. Feature-Specific Engineering

Feature engineering is applied conditionally based on the underlying data modality, which broadly comprises two categories: geospatial covariates and foundation model embeddings.

Geospatial Covariates: We apply standard tabular transformations to stabilize training:

- *Skewness Correction:* A $\log(1+x)$ transformation is applied to features exhibiting severe skewness ($|\text{skew}| > 1.0$).
- *Outlier & Redundancy Mitigation:* Features are clipped to their 1st and 99th percentiles to bound extreme values, and collinearity is addressed by dropping one feature from any pair with a high Pearson correlation coefficient ($|r| > 0.95$).
- *Scaling:* Variance is normalized using a standard Z-score transformation (StandardScaler).

Foundation Model Embeddings: We apply L2 normalization to project location-based model embeddings onto the unit hypersphere. Crucially, we eschew independent statistical transformations (such as Z-score or logarithmic scaling) on individual dimensions to preserve the semantic topology and angular relationships of the learned representation space.

4.3. Geospatial Prediction

The Automated Model Prediction stage is responsible for model training, hyperparameter optimization, and evaluation using pre-curated datasets. Operating within a strict isolation scope, the prediction agent is prohibited from fetching or mutating data, relying entirely on train and test DataFrame handles provided by the curation stage.

4.3.1. Modeling Objectives

The geospatial prediction stage automatically identifies the appropriate prediction task from the user's prompt and aligns the optimization objective across three core frameworks: Nowcasting, Super-Resolution, and Spatial Regression. Detailed problem formulations are explained below.

General Notations. Given a location i , let X_i^{cov} denote automatically selected geospatial covariates, and let X_i^{emb} denote location-based embedding signals generated by pretrained geospatial foundation models. The feature is constructed as $X_i = [X_i^{\text{cov}}, X_i^{\text{emb}}]$. Let f parameterized by θ denote a model to train, and y denote the prediction target. The general training strategy is to optimize parameter θ to θ^* which minimizes $\sum_i L(f(X_i; \theta), y_i)$, where $L(\cdot, \cdot)$ is the objective function. Using the trained model parameters θ^* , we infer the prediction target at location j with $\hat{y}_j = f(X_j; \theta^*)$. All three prediction tasks employ variants of this general formulation.

Epidemiological Nowcasting. Epidemiological Nowcasting refers to training on observed infection locations and predicting the infection spread frontier, i.e., the increase in infection caseloads, in a forecasting horizon (i.e., a 7-day window). Pandemic-specific signals such as infrastructure density, mobility corridors, etc., are of vital importance to prediction performance; therefore, when fetching geospatial covariates X_i^{cov} , the Planetary Prediction Engine automatically selects pandemic-focused signals beyond common signals (i.e., demographics). The prediction target y_i in this case is $y_i^{(t)} = C_i(t + \Delta t) - C_i(t)$, where $C_i(t)$ is the total confirmed caseloads at time t at location i , and Δt is the nowcasting horizon (i.e., 7 days). The model optimization is conducted spatially across the total infection-risky region and temporally across all historical expanding-window folds:

$$\theta^* = \operatorname{argmin}_{\theta} \sum_t \sum_i L(f(X_i^{(t)}; \theta), y_i^{(t)}).$$

Super-Resolution Downscaling. Super-Resolution Downscaling refers to training on observed data at coarse granularity (i.e., county-level) and predicting at fine-grained granularity (i.e., ZIP-code-level) by injecting zero-label ZIP rows with geospatial features. To overcome the ecological fallacy associated with coarse regional boundaries (where uninhabited forests receive identical operational priority to high-density transit nodes), we define a spatial downscaling framework. Let C represent a coarse geography (e.g., a county or health zone). Let Z denote fine-resolution target localities (e.g., ZIP codes or 1 km grid tiles). The model is trained on the coarse administrative boundaries:

$$\theta^* = \operatorname{argmin}_{\theta} \sum_{c \in C} L(f(X_c; \theta), y_c),$$

then we infer values across all fine-grained locations:

$$\hat{y}_z = f(X_z; \theta^*), \quad z \in Z.$$

Spatial Regression. Spatial Regression refers to training on observed locations and predicting missing values at the same geographic granularity. To address data sparsity and incomplete observations at a uniform geographic scale (where reporting constraints or missing surveys leave certain units without labels), we define a spatial regression framework for imputation. Let O and M represent observed and unobserved locations, respectively. O and M should be of the same granularity, i.e., both at the county-level. The model is trained on the observed locations:

$$\theta^* = \operatorname{argmin}_{\theta} \sum_{o \in O} L(f(X_o; \theta), y_o),$$

then we infer missing values across unobserved locations:

$$\hat{y}_m = f(X_m; \theta^*), \quad m \in M.$$

4.3.2. Model Search

Model Families and Hyperparameter Tuning. The Planetary Prediction Engine evaluates four core supervised model families (key hyperparameters listed in Table 8): Regularized Linear Models, Histogram-Based Gradient Boosting (Boost), Extreme Gradient Boosting (XGBoost), and Multi-Layer Perceptrons (MLP). To guard against resource exhaustion in automated search loops, all model families are bound by hard safety caps on model complexity.

Table 8 | Model types supported by the Planetary Prediction Engine.

Model	Class	Implementation	Key Hyperparams
linear	Linear	Scikit-Learn (Ridge, Lasso, ElasticNet) [57]	Regularization Type, Penalty α , L1 ratio ρ
boost	HistBoost, XGBoost [12]	Scikit-Learn HistGradientBoosting Regressor; xgboost.XGBRegressor	Learning rate, Max leaf nodes, Min samples per leaf, Loss Type, Max tree depth, # of estimators, Subsample, Regularization
mlp	MLPKeras	Keras	Hidden layer sizes, Dropout rate, Optimizer configs

Model Validation and Selection Strategies. Model selection can be executed via sequential searches or parallelized batch search. The agent evaluates configurations using any of the following validation strategies:

- **Random Split:** Divides the training data into an 80/20 train/validation split using a fixed random seed, making it ideal for standard i.i.d. tabular datasets and fast exploratory parameter sweeps.
- **Spatial Group Split:** Partitions data along geographic boundaries, which prevents spatial autocorrelation leakage and provides an accurate measure of true out-of-region generalization [50, 59].
- **K-Fold Cross-Validation:** Implements a standard 3-fold cross-validation scheme to ensure stable performance estimates on small datasets or high-variance tasks.

Model Overfitting Guard Protocol. To ensure generalization across unseen geographies, the agent implements a multi-layered Overfitting Guard Protocol:

1. **Pre-Training Risk Assessment:** Prior to training, the agent executes a heuristic check on dataset characteristics (sample size n , feature-to-sample ratio p/n , and spatial grouping) to classify overfitting risk as Low, Medium, or High. If classified as Medium, tree configurations are restricted to conservative depths. If High, strong regularization is enforced, biasing selection towards regularized linear models.
2. **Post-Training Self-Correction Protocol:** To mimic the diagnostic judgment of a human practitioner, the agent implements a self-correction protocol: if the validation-set evaluation reveals catastrophic generalization failure (e.g., negative validation metrics or a large train-validation gap), the agent discards the current model, increases regularization constraints, and restarts with conservative configurations. The self-correction loop is constrained to a single iteration to prevent infinite optimization cycles.

Evaluation Metrics. All models report R^2 , Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Pearson correlation (r).

4.4. Agent Execution and Runtime Profile

As an example, in case of Ebola nowcasting, we noticed that the agent executed 793 steps across 3 sessions above:

- Session 1 (15.8 min): Initialization and baseline reproduction (Steps 0–194)
- Session 2 (23.8 min): Feature vectorization and pipeline acceleration (Steps 195–542)
- Session 3 (15.6 min): Ablation evaluations and hyperparameter searches (Steps 543–792)

5. Conclusion

We present the Planetary Prediction Engine, an autonomous AI system that translates natural-language queries into executed geospatial predictive models, spanning spatial regression, super-resolution downscaling, spatial transmission modeling, and epidemiological nowcasting. The Planetary Prediction Engine addresses a fundamental bottleneck in planetary analytics: the labor-intensive, expertise-dependent process of discovering, curating, and fusing heterogeneous geospatial data with domain-appropriate modeling architectures.

Our empirical evaluation across a multidimensional benchmark matrix spanning the Global North and Global South, public health and environmental risk, and data-rich and data-scarce regimes demonstrates that the Planetary Prediction Engine achieves expert-level or superior performance across all evaluated paradigms. For spatial regression on 21 US CDC health indicators, the system achieves a mean R^2 of 76.8%, exceeding the manually curated expert baseline of 60%. For super-resolution food security downscaling in Nigeria, the Planetary Prediction Engine doubles baseline accuracy (R^2 of 66.1% vs. 31.5%). For real-time nowcasting of the 2026 DRC Ebola outbreak, the system achieves a Recall@10 of 83.3% in predicting newly invaded health zones, representing a 10.3 percentage point absolute improvement over the public state-of-the-art.

We believe this work represents a meaningful step toward democratizing geospatial prediction, making it accessible to researchers, humanitarian organizations, and policymakers who need expert-level models but may lack the specialized engineering teams traditionally required to build them. The Planetary Prediction Engine makes predictive modeling accessible by shifting the researcher’s role from manual data engineering to high-level hypothesis direction. Ultimately, it paves the way towards building geospatial prediction models useful for real-world applications in a fast, reliable and impactful manner.

Acknowledgements

We are grateful to the UN World Food Programme (WFP) and Vulnerability Analysis and Mapping (VAM) team for the data and research support. We also thank the Institut National de Recherche Biomédicale (INRB) for their collaboration on the DRC Ebola nowcasting, and the teams behind Data Commons, Google Earth Engine, Population Dynamics Foundation Models, and AlphaEarth for providing the foundational data and model infrastructure that powers the Planetary Prediction Engine.

We extend our sincere gratitude to Aviv Slobodkin, Hamsa Subramanian, Jeremy Amez-Droz, Joydeep Paul, Lambert Rosique, Lily Mihalkova, Milind Tambe, and Tim Thelin for their insightful discussions and valuable feedback on this work.

References

- [1] M. Agarwal, M. Sun, C. Kamath, A. Muslim, P. Sarker, J. Paul, H. Yee, M. Sieniek, K. Jablonski, S. Vispute, et al. General geospatial inference with a population dynamics foundation model. *arXiv preprint arXiv:2411.07207*, 2024.
- [2] P. M. Atkinson. Downscaling in remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 22:106–114, 2013.
- [3] E. Aygün, A. Belyaeva, G. Comanici, M. Coram, H. Cui, J. Garrison, R. Johnston, A. Kast, C. Y. McLean, P. Norgaard, et al. An ai system to help scientists write expert-level empirical software. *Nature*, pages 1–3, 2026.
- [4] L. S. Bastos, T. Economou, M. F. Gomes, D. A. Villela, F. C. Coelho, O. G. Cruz, M. S. Carvalho, and C. T. Codeço. A modelling framework for nowcasting disease incidence with applications to malaria and severe acute respiratory infection. *Statistics in Medicine*, 38(24):4853–4867, 2019.
- [5] A. Bell, A. Aides, A. Helmy, A. Muslim, A. Barzilai, A. Slobodkin, B. Jaber, D. Schottlander, G. Leifman, J. Paul, et al. Earth ai: unlocking geospatial insights with foundation models and cross-modal reasoning. *arXiv preprint arXiv:2510.18318*, 2025.
- [6] A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, and P. Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023.
- [7] B. Brown, J. Juravsky, R. Ehrlich, R. Clark, Q. V. Le, C. Ré, and A. Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.
- [8] C. F. Brown, M. R. Kazmierski, V. J. Pasquarella, et al. AlphaEarth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291*, 2025.
- [9] M. Burke, A. Driscoll, D. B. Lobell, and S. Ermon. Using satellite imagery to understand and promote sustainable development. *Science*, 371(6535):eabe8628, 2021.
- [10] Centers for Disease Control and Prevention. CDC/ATSDR social vulnerability index. <https://www.atsdr.cdc.gov/placeandhealth/svi/index.html>, 2020.
- [11] L. Chen, M. Zaharia, and J. Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023.
- [12] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.
- [13] Clay Foundation. Clay: An open source foundation model for earth observation. <https://huggingface.co/made-with-clay/Clay>, 2024. Accessed: 2026-08-01.
- [14] N. Cressie. *Statistics for Spatial Data*. Wiley, 1993.
- [15] A. Das, W. Kong, A. Leber, R. Mathews, and R. Sen. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

- [16] K. Didan. MOD13A3 MODIS/Terra vegetation indices monthly L3 global 1km SIN grid V006. *NASA EOSDIS Land Processes DAAC*, 2015. doi: 10.5067/MODIS/MOD13A3.006.
- [17] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch. Improving factuality and reasoning in language models through multiagent debate. *Proceedings of the 41st International Conference on Machine Learning*, 2024.
- [18] C. D. Elvidge, K. Baugh, M. Zhizhin, F. C. Hsu, and T. Ghosh. VIIRS night-time lights. *International Journal of Remote Sensing*, 38(21):5860–5879, 2017.
- [19] Epidemiological.org Consortium. Real-time spatiotemporal risk modelling of the Bundibugyo ebola virus outbreak 2026. <https://www.epidemiological.org/t/real-time-spatiotemporal-risk-modelling-of-the-bundibugyo-ebola-virus-outbreak-16>, 2026. Accessed: 2026-08-01.
- [20] Federal Emergency Management Agency. The national risk index: Technical documentation. Technical report, FEMA, US Department of Homeland Security, Washington, DC, 2023.
- [21] M. Feurer, A. Klein, K. Eggenberger, J. T. Springenberg, M. Blum, and F. Hutter. Efficient and robust automated machine learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.
- [22] B. E. Flanagan, E. W. Gregory, E. J. Hallisey, J. L. Heitgerd, and B. Lewis. A social vulnerability index for disaster management. *Journal of Homeland Security and Emergency Management*, 8(1):0000102202154151000271, 2011.
- [23] Flowminder Foundation. Population movements from bunia, mongbwalu and rwampara based on privacy secure analysis of mobile operator data from voda-com congo, may 2026. URL https://www.flowminder.org/media/eagbkohv/ebola-report-update-4-june_final.pdf. Accessed: 2026-06-03.
- [24] C. Funk, P. Peterson, M. Landsfeld, D. Pedreros, J. Verdin, S. Shukla, G. Husak, J. Rowland, L. Harrison, A. Hoell, et al. The climate hazards group infrared precipitation with station data (CHIRPS): A new dataset for monitoring extremes. *Scientific Data*, 2(1):150066, 2015.
- [25] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202: 18–27, 2017.
- [26] K. J. Greenlund, H. Lu, X. Zhang, J. Holt, K. Matthews, and J. Croft. PLACES: Local data for better health. *Journal of Public Health Management and Practice*, 28(S2):S123–S130, 2022.
- [27] R. Guha, N. Alon, B. Kaluza, and R. Ramakrishnan. Data Commons: Organising the world’s public data. In *Proceedings of the ACM Web Conference 2023*, pages 3900–3908, 2023.
- [28] S. Guo, C. Deng, Y. Wen, H. Chen, Y. Li, and C. Sun. DS-Agent: Automated data science by empowering large language models with case-based reasoning. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.
- [29] M. Haklay and P. Weber. OpenStreetMap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18, 2008.
- [30] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

- [31] M. Höhle. Epidemiological nowcasting: Assessing the progress of outbreaks in real time. *Epidemiology and Infection*, 145(15):3100–3110, 2017.
- [32] S. Hong, X. Zheng, J. Chen, Y. Cheng, C. Zhang, Z. Wang, S. K. S. Yau, D. Lin, L. Zhou, C. Ran, et al. MetaGPT: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations (ICLR)*, 2024.
- [33] S. Hong, Y. Zhuge, J. Chen, X. Zheng, Y. Cheng, J. Wang, J. Li, Z. Wang, D. Lin, J. Yu, et al. Data interpreter: An LLM agent for data science. *arXiv preprint arXiv:2402.18679*, 2024.
- [34] Institut National de Recherche Biomédicale (INRB). DRC Ebola Virus Disease (BDBV) 2026 Outbreak Caseload Registry. https://github.com/INRB-UMIE/Ebola_DRC_2026, 2026. Accessed: 2026-08-01.
- [35] IPC Global Partners. Integrated Food Security Phase Classification Technical Manual Version 3.1: Evidence and standards for better food security and nutrition decisions. Technical report, Food and Agriculture Organization of the United Nations (FAO), Rome, Italy, 2021.
- [36] J. Jakubik, M. Muszynski, C. Moffeit, R. Gao, S. Ramasubramanian, et al. Foundation models for generalist geospatial artificial intelligence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [37] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.
- [38] C. E. Jimenez, J. Yang, A. Wettig, S. Yang, S. Yao, K. Yang, K. Narasimhan, and S. Yao. SWE-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations (ICLR)*, 2024.
- [39] M. Kalkuhl, J. von Braun, and M. Torero. *Food Price Volatility and Its Implications for Food Security and Policy*. Springer Nature, 2016.
- [40] S. Kapoor and A. Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023.
- [41] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4): 1–21, 2012.
- [42] Y. Kim, K. Gu, C. Park, C. Park, S. Schmidgall, A. A. Heydari, Y. Yan, Z. Zhang, Y. Zhuang, M. Malhotra, P. P. Liang, H. W. Park, Y. Yang, X. Xu, Y. Du, S. Patel, T. Althoff, D. McDuff, and X. Liu. Towards a science of scaling agent systems. *arXiv preprint arXiv:2512.08296*, 2025.
- [43] B. Knyazev, H. de Vries, M. Cisse, A. Courville, and G. W. Taylor. SatCLIP: Global general-purpose location embeddings. In *International Conference on Learning Representations (ICLR)*, 2024.
- [44] V. Krechetova and D. Kochedykov. Geobenchx: Benchmarking llms in agent solving multi-step geospatial tasks. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Generative and Agentic AI for Multi-Modality Space-Time Intelligence*, pages 27–35, 2025.
- [45] C. Lu, C. Lu, R. T. Lange, J. Foerster, J. Clune, and D. Ha. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

- [46] D. Luxen and C. Vetter. Real-time routing with OpenStreetMap data. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 513–516, 2011.
- [47] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegrefe, U. Alon, N. Dziri, S. Prabhunoye, Y. Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 2023.
- [48] O. Mbulayi, P. Akilimali, C. Judge, B. Gutierrez, P. Mulu, C. M. Ibolobolo, J.-C. Sibo, R. Nkwele wa Nkwele, M. M. Hermann, L. Lawanga Ontshick, D. Mukadi, B. Kanku, E. Katanga, O. le Polain de Waroux, et al. Real-time epidemic intelligence in a public health emergency: The 2026 Bundibugyo virus outbreak. *The Lancet Infectious Diseases*, 2026. doi: 10.1016/S1473-3099(26)00330-0. URL [https://doi.org/10.1016/S1473-3099\(26\)00330-0](https://doi.org/10.1016/S1473-3099(26)00330-0).
- [49] S. F. McGough, M. A. Johansson, M. Lipsitch, and M. Santillana. Nowcasting by Bayesian smoothing: A framework for real-time outbreak tracking. *PLOS Computational Biology*, 16(4): e1007741, 2020.
- [50] H. Meyer, C. Reudenbach, T. Hengl, M. Katurji, and T. Nauss. Importance of spatial predictor variable selection in machine learning applications - moving from local to spatial cross-validation. *Ecological Modelling*, 411:108815, 2019.
- [51] S. Meyer, L. Held, and M. Höhle. Spatio-temporal analysis of epidemic phenomena using the R package surveillance. *Journal of Statistical Software*, 77(11):1–55, 2017. doi: 10.18637/jss.v077.i11.
- [52] G. Mialon, C. Fourrier, C. Swift, T. Wolf, Y. LeCun, and T. Scialom. GAIA: A benchmark for general AI assistants. *arXiv preprint arXiv:2311.12983*, 2023.
- [53] A. Novikov, N. Vü, M. Eisenberger, E. Dupont, P.-S. Huang, A. Z. Wagner, S. Shirobokov, B. Kozlovskii, F. J. R. Ruiz, A. Mehrabian, M. P. Kumar, A. See, S. Chaudhuri, G. Holland, A. Davies, S. Nowozin, P. Kohli, and M. Balog. AlphaEvolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.
- [54] I. Ong, A. Almahairi, V. Wu, W.-L. Chiang, T. Wu, J. E. Gonzalez, M. W. Kadous, and I. Stoica. RouteLLM: Learning to route LLMs with preference data. *arXiv preprint arXiv:2406.18665*, 2024.
- [55] S. Openshaw. *The Modifiable Areal Unit Problem*. Geo Books Norwich, 1984.
- [56] D. K. Park, A. Gelman, and J. Bafumi. Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis*, 12(4):375–385, 2004.
- [57] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [58] E. Real, C. Liang, D. So, and Q. Le. AutoML-Zero: Evolving machine learning algorithms from scratch. *International Conference on Machine Learning (ICML)*, 2020.
- [59] D. R. Roberts, V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, A. El-Gabbas, J. Räß, and C. F. Dormann. Cross-validation strategies for data with temporal, spatial, hierarchical or phylogenetic structure. *Ecography*, 40(8):913–929, 2017.

- [60] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 2023.
- [61] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–110, 2012.
- [62] C. Snell, J. Lee, K. Xu, and A. Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.
- [63] A. J. Tatem. WorldPop, open data for spatial demography. *Scientific Data*, 4(1):170004, 2017.
- [64] H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal. MuSiQue: Multihop questions via single hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.
- [65] G. Tseng, I. Zvonkov, H. R. Kerner, and D. Rolnick. Lightweight foundation model for earth observation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [66] C. Wang, Q. Wu, M. Weimer, and E. Zhu. FLAML: A fast and lightweight automated machine learning library. In *Proceedings of Machine Learning and Systems (MLSys)*, volume 3, pages 434–447, 2021.
- [67] J. Wang, J. Wang, B. Athiwaratkun, C. Zhang, and J. Zou. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024.
- [68] L. Wang, W. Xu, Y. Lan, Z. Hu, Y. Lan, R. K.-W. Lee, and E.-P. Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.
- [69] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. *International Conference on Learning Representations*, 2023.
- [70] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.
- [71] A. Wesolowski, N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow, and C. O. Buckee. Quantifying the impact of human mobility on malaria transmission in Kenya. *Science*, 338(6104):267–270, 2012.
- [72] D. Wiesmann, L. Bassett, T. Benson, and J. Hoddinott. Validation of the World Food Programme’s food consumption score and alternative indicators of household food security. IFPRI Discussion Paper 00870, International Food Policy Research Institute (IFPRI), 2009.
- [73] World Food Programme. HungerMap LIVE: Near real-time food security monitoring systems. Technical report, WFP Hunger Monitoring Unit, Rome, Italy, 2021.
- [74] World Food Programme, Vulnerability Analysis and Mapping (VAM). Global Food Prices Database. <https://data.humdata.org/dataset/global-wfp-food-prices>, 2024. Accessed via HDX. Accessed: 2026-08-01.
- [75] World Health Organization. Ebola disease caused by Bundibugyo virus – Democratic Republic of the Congo. <https://www.who.int/emergencies/disease-outbreak-news/item/2026-DON614>, 2026. Disease Outbreak News, 22 May 2026. Accessed: 2026-08-01.

- [76] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, et al. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.
- [77] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, and C. D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2018.
- [78] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2023.
- [79] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao. ReAct: Synergizing reasoning and acting in language models. *International Conference on Learning Representations*, 2023.
- [80] C. Yeh, A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications*, 11(1):2583, 2020.
- [81] M. Zaharia, O. Khattab, L. Chen, J. E. Gonzalez, I. Stoica, et al. The shift from models to compound AI systems. *Berkeley AI Research Blog*, 2024.

A. Benchmarks

- **SVI (Social Vulnerability Index):** County-level and ZIP-code level socioeconomic, household, minority, and housing/transportation vulnerability metrics across the United States [10, 22].
- **CDC Health Variables:** 21 chronic disease and health indicator variables at the census tract level (e.g., Obesity, Diabetes, Stroke, Asthma, COPD, High Blood Pressure) from the CDC PLACES dataset [26].
- **FEMA Environmental Risk:** Census-tract-level environmental and climate risk indices from the FEMA National Risk Index [20].
- **Nigeria Food Security Downscaling:** Predicting Food Consumption Group (FCG 72) scores (continuous food consumption index) and acute food insecurity across Nigeria. Features include World Bank food price indices (OHLC food price index, inflation food price index), HungerMap acute food insecurity metrics [73], HDX precipitation/rainfall time series, HDX NDVI vegetative indices, fatality datasets, and Gemini processed news data. Training at ADM1 state level (N=30), evaluation at ADM2 LGA level (N=581).
- **2026 DRC Ebola Outbreak:** Simulating the May–June 2026 outbreak of Ebola Bundibugyo virus (BDBV) across the Democratic Republic of Congo using official WHO surveillance reports [34, 75]. Datasets include the official INRB live caseload registry ([INRB GitHub](#)), HDX COD Admin 3 subnational boundaries ([HDX COD](#)), OpenStreetMap road network vectors and structural building footprints ([HDX OSM](#)), Data Commons demographics, and Earth Engine environmental layers (WorldPop, ERA5-Land temperature/precipitation, Copernicus built-up density).
- **Ebola Outbreak Hotspot Prediction:** Simulating the May–July 2026 outbreak of Ebola Bundibugyo virus (BDBV) across Ituri, Nord-Kivu, and Haut-Uele provinces in the Democratic Republic of Congo using official WHO surveillance reports [75] and INRB registry data [34].
 - **Signals and data sources.** The analysis drew on three categories of data, all harmonised to 519 Ministry of Health *zones de santé* (health zones) in the DRC. All non-confidential data sources are publicly available through an open-access GitHub repository.
 - * *Epidemiological:* Daily confirmed BDBV case counts by symptom onset date from the DHIS2 linelist (INRB/INSP), supplemented by daily case, death, and contact-tracing indicators manually transcribed from the INSP Situation Report (SitRep) MVE PDF series. Three health zones (Oicha, Makiso-Kisangani, Lubunga) absent from the linelist were manually added based on SitRep confirmation.
 - * *Demographic and socioeconomic:* Gridded population counts and density (WorldPop), socioeconomic deprivation and inequality indices (Climate-Conflict Vulnerability Index, CCVI), GDP per capita (Kummu et al.), and health facility counts and densities (GRID3 COD Health Facilities v8.0).
 - * *Flowminder Mobility:* Mobile-phone-based relocation estimates from the Flowminder Foundation [23], capturing subscriber proportions from the Ituri epicentre (Bunia, Mongbwalu, Rwampara) detected in other health zones over successive weekly windows; zone-to-zone road travel times via the OpenStreetMap OSRM API; and fitted gravity and radiation models for zone pairs not covered by Flowminder.
 - **Data preprocessing.** The raw data streams were processed through the following steps before entering the spatiotemporal invasion model:
 - * *Spatial aggregation:* All raster and point-source covariates were resampled and aggregated to the health zone level using the DART pipeline. Covariates were standardised (z-scored) prior to model fitting; population counts were additionally log-transformed.
 - * *Epidemiological signal processing:* Missing symptom onset dates were imputed from

sample collection dates using an estimated delay distribution. Case time series were then nowcasted to adjust for reporting delays. The nowcasted incidence in each source zone was convolved with a discretised Gamma generation-time distribution, evaluated at three parameterisations with means of 12.0, 15.3, and 18.0 days, and binned to weekly intervals to match the one- and two-week forecast horizons.

- * *Mobility matrix construction*: Flowminder empirical mobility estimates and parametric gravity/radiation models were combined into a composite mobility matrix W , which was multiplied by the generation-time-weighted incidence to compute the importation pressure into each at-risk zone.
- * *Vulnerability scoring*: A rank-based composite of health-site density, health-site count, travel time to the nearest facility, and CCVI deprivation was multiplied by the predicted invasion probability to yield a priority score rescaled to $[0, 1]$.

B. Geospatial Covariates and Foundation Models

Details of the inventories of Geospatial Covariates and Geospatial Foundation Models are listed in Table 9 and 10.

Table 9 | Inventory of Geospatial Covariates

Covariate Category	Cate- gory	Data Source / Provider	Key Variables & Dimensions Captured	Access Class
Socio-Demographic & Economic		Data Commons (Google Open Source)	Population density, household income trends, poverty rates, age/gender distributions, chronic disease rates.	Public (Open API & Free)
Ecological & Environmental		Google Earth Engine (EE)	Multi-spectral imagery (Sentinel-2, Landsat, MODIS), NDVI (vegetation), soil moisture, LST anomalies.	Public (Free for Research Use)
Infrastructure Mobility	&	Google Maps Platform Insights	Aggregated Statistics on POIs	Public (GCP Commercial Billing)
Infrastructure Mobility	&	OpenStreetMap (OSM) (via OCHA HDX)	Road network vectors, administrative boundary polygons, structural building footprints.	Public (Open Source / Free)
Meteorological		Open-Meteo / NOAA	Daily temperature, cumulative precipitation, relative humidity, wind velocity profiles.	Public (Open API / Free)

Table 10 | Inventory of Geospatial Foundation Models

Foundation Model	Population Dynamics Foundation Model (PDFM)	AlphaEarth Foundation Model
Modality / Coverage	Socio-economic & Demographic (State, County, ZCTA, City)	Ecological & Remote Sensing (High-resolution satellite raster)
Dimension	330 / 512	64
Key Semantic Characteristics	Encodes latent population density, mobility, economic growth, and regional demographics.	Encodes multi-scale topography, vegetation canopy, land-use semantics, and terrain profiles.
Access Class	Pre-General Availability + Research Use	Publicly available

C. Intelligent Data Selection Methodology & Rubrics

Table 11 details the five-dimension scoring rubric utilized by the Provenance-First Prioritization sub-stage to rank candidate datasets. Provenance and license compliance carry the highest weight (5x), ensuring the system preferentially assembles covariates from openly licensed, institutionally authoritative sources.

Table 11 | Five-Dimension Scoring Rubric

Dimension		Weight	Criteria & Point Allocations
Provenance & License		High (x5)	Institutional backing and license openness. Openly licensed data from established repositories (4 pts) → government/NGO open data (3 pts) → attribution-required academic data (2 pts) → restrictive or unknown license (0 pts).
Spatio-Temporal Fit-ness	Signal Alignment	Medium (x2)	Spatial coverage of the target area of interest and temporal overlap with the prediction window.
Signal Alignment		High (x3)	Degree to which the dataset directly addresses a signal from the Signal Guide: exact variable match (4 pts) → requires transformation (3 pts) → proxy coverage (2 pts) → tangential (0 pts).
Format & Quality		Low (x1)	Data readiness: structured and pre-parsed (4 pts) → standard interoperable format (2 pts) → legacy or malformed (0 pts).
Redundancy		Low (x1)	Whether the dataset provides unique coverage or duplicates a higher-scoring source for the same signal.

D. Related work

D.1. Autonomous AI Agents for Science

Recent work has explored using LLMs as autonomous systems and empirical research assistants for scientific discovery and data science tasks. These agent systems build on foundational LLM reasoning techniques including chain-of-thought prompting [70], ReAct-style reasoning and acting [79], self-consistency [69], plan-and-solve decomposition [68], tree-structured deliberation [78], iterative self-refinement [47], and tool-use paradigms [60]. Multi-agent architectures [17, 67] and test-time compute scaling [7, 62] further inform the design of compound AI systems [42, 81], with cost-efficient routing strategies [11, 54] enabling practical deployment. Systems such as The AI Scientist [45] and ChemCrow [6] demonstrate the potential of LLMs to autonomously formulate hypotheses, design experiments, and analyze results in complex scientific domains. In data science, frameworks like Data Interpreter [33], AutoGen [76], and MetaGPT [32] establish multi-agent architectures for code generation, tool execution, and iterative debugging, while benchmarks such as SWE-bench [38], GAIA [52], GeoBenchX [44], HotpotQA [77], and MuSiQue [64] evaluate real-world task automation and multi-hop reasoning. DS-Agent [28] automates machine learning competition workflows using case-based reasoning, while Empirical Research Assistance (ERA) [3] and AlphaEvolve [53], building on earlier automated algorithm evolution [58], explores automated algorithm discovery, but need a well defined objective, whereas our approach automatically identifies geospatial task and objective function to optimize, along with autonomously fetching relevant geospatial data sources. Our work extends these paradigms to the geospatial and epidemiological domains by incorporating specialized foundation model embeddings (PDFM, AlphaEarth) and a

research-grounded Intelligent Data Selection pipeline capable of dynamically fetching from heterogeneous data repositories, which are critical for high-quality planetary prediction.

D.2. Geospatial Foundation Models

The emergence of foundation models has transformed remote sensing and geospatial informatics. The Population Dynamics Foundation Model (PDFM) [1] produces 330-dimensional to 512-dimensional embeddings for administrative boundaries (states, counties, ZCTAs, cities) globally, capturing latent socioeconomic, demographic, mobility, and health-related patterns. These embeddings serve as compact, highly expressive spatial features that significantly improve downstream regression tasks. AlphaEarth [8] provides complementary 64-dimensional spatial features derived from high-resolution satellite imagery, encoding multi-scale topography, vegetation canopy, and land-use semantics at the administrative boundary level. Other notable efforts include Clay [13], Prithvi [36], SatCLIP [43], and Presto [65], which learn representations from multi-spectral earth observation archives and location coordinates, as well as temporal foundation models such as TimesFM [15] for time-series forecasting. Our system is, to our knowledge, the first to autonomously select, align, and fuse heterogeneous embedding sources in a unified, agentic ML pipeline.

D.3. Geospatial Prediction & Super-Resolution Downscaling

Spatial regression and geographic super-resolution downscaling have been studied extensively in geostatistics [14] and more recently with deep learning approaches. Pioneering work by Jean et al. [37] and Yeh et al. [80] demonstrated that combining satellite imagery with deep learning can accurately predict poverty and economic well-being across developing regions. Similarly, super-resolution techniques have been applied to downscale coarse climate and environmental variables to fine-grained grid tiles. Our work differs by automating the full pipeline from intelligent data selection and multimodal curation to model evaluation rather than focusing on a single manual modeling technique.

D.4. Epidemiological Nowcasting & Spatial Transmission Modeling

Real-time epidemic tracking faces two fundamental challenges: surveillance lag (reporting delays resulting in right-truncated caseload data) and anisotropic transmission (disease propagation along mobility corridors rather than uniform spatial diffusion). Foundational literature in epidemiological nowcasting [4, 31, 49] establishes statistical and Bayesian smoothing frameworks to correct for right-truncation in real-time reporting, while spatiotemporal surveillance analysis tools [51] enable systematic modeling of epidemic phenomena. Meanwhile, spatial transmission modeling traditionally relies on gravity models or radiation models to estimate inter-regional flux. We bridge these domains by introducing an automated SEIR nowcasting framework that couples discrete-time compartmental dynamics with empirical mobility flows and foundation model similarity kernels, while utilizing learned covariates to modulate transmission parameters.

E. Example Prompts

E.1. Epidemiological Transmission

```
Build a spatial invasion prediction model for the 2026 Bundibugyo
Ebola virus outbreak in DRC.
Data splits are as below:
- Train Cutoffs: ['2026-05-18', '2026-05-25', '2026-06-01']
- Eval Cutoffs: ['2026-06-08', '2026-06-15', '2026-06-22', '2026-06-29',
'2026-07-06']
- Test Set: Uninfected zones at each eval cutoff.
Predict which currently-uninfected health zones will experience new cases
within the next 1 week. Performance is evaluated by Recall@10 (how many
of the top 10 predicted zones actually get invaded).
```

E.2. Super Resolution

E.2.1. Food Security

```
Execute Super-Resolution downscaling of WFP HungerMap metrics (FCG and
rCSI) for Nigeria. Target Variables is Extract State-level FCG and rCSI
data. WFP survey data is constrained to coarse administrative levels.
Downscale this to the LGA level (1km2 hex grids) to map high-need zones
for targeted aid distribution.
```

E.2.2. SVI

```
RPL_THEME1(Socioeconomic Status) is mathematically defined as:
PercentileRank(PercentileRank(Poverty) +
PercentileRank(Unemployment) +
PercentileRank(PerCapitaIncome) +
PercentileRank(NoHighSchoolDiploma) +
PercentileRank(Uninsured)).
PercentileRank(x) = (Rank(x) - 1) / (N - 1).
Ties get lowest rank. Poverty, Unemployment, NoHighSchoolDiploma, and
Uninsured are ascending; PerCapitaIncome is descending.
Train a spatial regression model to predict the RPL_THEME1. Report the
R-squared score on the test set.
```

E.3. Spatial Regression

E.3.1. CDC Health

```
Train a spatial regression model for CDC Health indicators. The target
Variable: Percent_Person_WithHighCholesterol. Report the R-squared
score.
```

E.3.2. FEMA Environmental Risk Factors

Perform spatial regression and signal ablation for FEMA risk indicators. The Target Variable is {User_Target}. Report the R-squared score.

E.3.3. SVI

Train a spatial regression model for SVI Socioeconomic Status (RPL_THEME1). RPL_THEME1 is a percentile rank based on indicators like poverty, unemployment, income, and education. Report the R-squared score.

F. Expanded Benchmark & Ablation Results

F.1. Spatial Transmission

SEIR vs Spatial Transmission Regression

Table 12 | Spatial Transmission Regression vs SEIR Bayesian Modeling

System	Model Configuration	Val Top-10 Acc	Val RMSE	Test Top-10 Acc	Test RMSE
Baseline / SOTA	STM XGBoost: Baseline	0.4	4.5810	0.5	6.0810
(a) Covariates	STM XGBoost: + Earth AI Covariates	0.8	3.2510	0.8	4.2910
(b) Covariates + Embeddings	STM XGBoost: + PDFM Embeddings	0.6	3.9810	0.7000	5.2150
Baseline / SOTA	SEIR Nowcasting: Baseline	0.2	6.6571	0.3	7.2246
(b) Covariates + Embeddings	SEIR Nowcasting: + PDFM Embeddings	0.3	2.7229	0.4	3.4781
(c) Planetary Prediction Agent	SEIR Nowcasting: + Earth AI Learned Covariates	0.3	0.7019	0.4	0.9204

We compare Mechanistic SEIR Nowcasting and Spatial Transmission Regression (STR) for predicting outbreak escalation (case volume) and identifying vulnerable zones [48]. Table 12 reports the ablation results across regional health zones in Ituri and Nord-Kivu, evaluating validation and test metrics for static vulnerability ranking (Top-10 accuracy) versus dynamic caseload trajectory tracking (RMSE). While STR excels at static vulnerability ranking (achieving Test Top-10 accuracy of 0.8), SEIR Nowcasting is substantially stronger at predicting case volume (achieving a Test RMSE of 0.9204) when enhanced with empirical mobility flows and learned covariates.

F.2. Spatial Regression

FEMA predictions

We present the initial ablation comparison without intelligent selection via data commons in Table 13,14,15. We extend this with the inclusion of discovered intelligent data selection, which resulted in a modified feature suite and performance metrics (Table 16,17,18).

Table 13 | PPE performance of FEMA targets (Socioeconomic & Composite) with ablation on PDFM + AEF + DC features

FEMA Target Variable	PDFM + AEF Manual expert Mean R ²	PDFM + AEF + DC (Feature ablation) Mean R ² [95% CI]	Data Common variables used (Comma-Separated Variables)
Social Vulnerability	0.4824	0.6755 [0.6556, 0.6941]	Count_Person, BelowPovertyLevelInThePast12Months, Count_Household, HouseholderAge65OrMoreYears, SingleMotherFamilyHousehold, LimitedEnglishSpeakingHousehold, NoComputer, NoInternetAccess, With0AvailableVehicles, WithFoodStampsInThePast12Months, IncomeOfUpto10000USDollar, IncomeOf10000To14999USDollar, IncomeOf15000To19999USDollar, IncomeOf20000To24999USDollar, IncomeOf100000To124999USDollar, IncomeOf200000OrMoreUSDollar, Count_HousingUnit, Count_HousingUnit_Before1939DateBuilt
Resilience Score	0.7402	0.7358 [0.7287, 0.7427]	Count_Person, BelowPovertyLevelInThePast12Months, Count_Household, HouseholderAge65OrMoreYears, SingleMotherFamilyHousehold, LimitedEnglishSpeakingHousehold, NoComputer, NoInternetAccess, With0AvailableVehicles, WithFoodStampsInThePast12Months, IncomeOfUpto10000USDollar, IncomeOf10000To14999USDollar, IncomeOf15000To19999USDollar, IncomeOf20000To24999USDollar, IncomeOf100000To124999USDollar, IncomeOf200000OrMoreUSDollar, Count_HousingUnit, Count_HousingUnit_Before1939DateBuilt
Composite Risk Score	0.5983	0.6221 [0.6078, 0.6335]	Count_Person, BelowPovertyLevelInThePast12Months, Count_Household, HouseholderAge65OrMoreYears, SingleMotherFamilyHousehold, LimitedEnglishSpeakingHousehold, NoComputer, NoInternetAccess, With0AvailableVehicles, WithFoodStampsInThePast12Months, IncomeOfUpto10000USDollar, IncomeOf10000To14999USDollar, IncomeOf15000To19999USDollar, IncomeOf20000To24999USDollar, IncomeOf100000To124999USDollar, IncomeOf200000OrMoreUSDollar, Count_HousingUnit, Count_HousingUnit_Before1939DateBuilt
Expected Annual Loss	0.6227	0.6423 [0.6319, 0.6531]	Count_Person, BelowPovertyLevelInThePast12Months, Count_Household, HouseholderAge65OrMoreYears, SingleMotherFamilyHousehold, LimitedEnglishSpeakingHousehold, NoComputer, NoInternetAccess, With0AvailableVehicles, WithFoodStampsInThePast12Months, IncomeOfUpto10000USDollar, IncomeOf10000To14999USDollar, IncomeOf15000To19999USDollar, IncomeOf20000To24999USDollar, IncomeOf100000To124999USDollar, IncomeOf200000OrMoreUSDollar, Count_HousingUnit, Count_HousingUnit_Before1939DateBuilt

Table 14 | PPE performance of FEMA targets with ablation on PDFM + AEF + DC features

FEMA Target Variable	PDFM + AEF Manual expert Mean R ²	PDFM + AEF + DC (Feature ab- lation) Mean R ² [95% CI]	Data Common vari- ables used (Comma- Separated Variables)
Tornado Risk	0.7543	0.7542 [0.7468, 0.7617]	None
Hail Risk	0.5949	0.5976 [0.5807, 0.6138]	None
Strong Wind Risk	0.7339	0.7271 [0.7179, 0.7362]	None
Lightning Risk	0.4884	0.4918 [0.4758, 0.5090]	None
Drought Risk	0.6700	0.6866 [0.6332, 0.7323]	None
Heat Wave Risk	0.6336	0.6350 [0.6193, 0.6489]	None
Cold Wave Risk	0.6821	0.6724 [0.6489, 0.6928]	None
Wildfire Risk	0.7622	0.7541 [0.7361, 0.7730]	None
Winter Weather	0.6135	0.5933 [0.5793, 0.6063]	None
Ice Storm Risk	0.5320	0.5239 [0.5046, 0.5421]	None

Table 15 | PPE performance of FEMA targets (Geophysical & Hydrological) with ablation on PDFM + AEF + DC features

FEMA Target Vari- able	PDFM + AEF Manual expert Mean R ²	PDFM + AEF + DC (Feature ab- lation) Mean R ² [95% CI]	Data Common variables used (Comma- Separated Variables)
Coastal Flooding	0.3786	0.3840 [0.3394, 0.4240]	None
Riverine Flooding	0.2930	0.2935 [0.2695, 0.3179]	None
Hurricane Risk	0.8270	0.8124 [0.8011, 0.8232]	None
Earthquake Risk	0.8606	0.8608 [0.8540, 0.8672]	None
Landslide Risk	0.5614	0.5624 [0.5339, 0.5887]	None
Avalanche Risk	0.1474	0.1866 [0.0572, 0.2947]	Count_Person, BelowPovertyLevelInThePast12Months, Count_Household, HouseholderAge65OrMoreYears, SingleMotherFamilyHousehold, LimitedEnglishSpeakingHousehold, NoComputer, NoInternetAccess, With0AvailableVehicles, WithFoodStampsInThePast12Months, IncomeOfUpto10000USDollar, IncomeOf10000To14999USDollar, IncomeOf15000To19999USDollar, IncomeOf20000To24999USDollar, IncomeOf100000To124999USDollar, IncomeOf200000OrMoreUSDollar, Count_HousingUnit, Count_HousingUnit_Before1939DateBuilt

Table 16 | PPE performance on FEMA Socioeconomic & Composite risk targets with ablation on PDFM + AEF + DC and intelligent data selection.

FEMA Target Variable	PDFM + AEF + DC Feature Ab- lation Mean R ²	PDFM + AEF + Discovered DC Variables Intelligent Fea- ture Selection Mean R ² [95% CI]	Discovered / Selected External Variables (Sourced Non-Linearly)
Social Vulnerability	0.6755	0.6773 [0.6698, 0.6869]	Count_Person, HouseholderAge65OrMoreYears, SingleMotherFamilyHousehold
Resilience Score	0.7358	0.7258 [0.7062, 0.7348]	LimitedEnglishSpeakingHousehold, Median household income
Composite Risk Score	0.6221	0.6749 [0.6641, 0.6861]	WithFoodStampsInThePast12Months, Health insurance rates
Expected Annual Loss	0.6423	0.6974 [0.6888, 0.7055]	Count_HousingUnit_Before1939DateBuilt, Poverty Index metrics

Table 17 | PPE performance on FEMA Atmospheric & Climatological risk targets with ablation on PDFM + AEF + DC and intelligent data selection.

FEMA Target Variable	PDFM + AEF + DC Feature Ablation Mean R ²	PDFM + AEF + Discovered Variables Intelligent Feature Selection Mean R ² [95% CI]	Discovered / Selected External Variables (Sourced Non-Linearly)
Tornado Risk	0.7542	0.9168 [0.9123, 0.9208]	Agricultural vs Developed land ratios, Surface terrain elevation, Annual average precipitation, Distance to coast lines, Relative humidity
Hail Risk	0.5976	0.8656 [0.8540, 0.8807]	Cloud Cover fractions, Vegetation indices (NDVI) variance, Land surface temperature gradients, Distance to major water bodies
Strong Wind Risk	0.7339	0.6891 [0.6631, 0.7117]	NOAA Convective Storm Frequencies, Spherical coordinates (lat/lon 3D embeddings), Atmospheric pressure gradient proxies
Lightning Risk	0.4884	0.6607 [0.6051, 0.7055]	US Drought Monitor Palmer Indices, Agricultural land fraction, Water body area ratio, Mean annual precipitation
Drought Risk	0.6700	0.8711 [0.8648, 0.8775]	NOAA High Wind Event frequencies, Coastal proximity, Terrain elevation, Topographic wind exposure
Heat Wave Risk	0.6336	0.5964 [0.5790, 0.6139]	US Census Educational attainment, Health insurance coverage fraction, Median household income, Internet access
Cold Wave Risk	0.6821	0.7083 [0.6915, 0.7245]	US Census Gazetteer Water Area (AWATER_SQMI), Water-to-Land ratio, Coastal distance, 3D spatial spherical coordinates
Wildfire Risk	0.7622	0.8366 [0.8264, 0.8460]	AlphaEarth hydrography & surface water embeddings, PDFM Embeddings
Winter Weather	0.6135	0.6515 [0.6366, 0.6659]	NOAA Lightning Strike densities, Spatial coordinates, Elevation & topographic roughness
Ice Storm Risk	0.5320	0.6265 [0.5979, 0.6546]	CDC Social Deprivation Indices, Spatial area ratios, Housing structural density, Multi-hazard loss exposure counts

Table 18 | PPE performance on FEMA Geophysical & Hydrological risk targets with ablation on PDFM + AEF + DC and intelligent data selection.

FEMA Target Variable	PDFM + AEF + DC Feature Ablation Mean R ²	PDFM + AEF + Discovered Variables Intelligent Feature Selection Mean R ² [95% CI]	Discovered / Selected External Variables (Sourced Non-Linearly)
Coastal Flooding	0.3786	0.5047 [0.4882, 0.5202]	USGS Fault Line distance proxies + AlphaEarth seismic geodynamics embeddings
Riverine Flooding	0.2930	0.3144 [0.2924, 0.3377]	Coastal proximity buffer, Historical tropical cyclone track density, Elevation
Hurricane Risk	0.8270	0.7729 [0.7650, 0.7809]	AlphaEarth satellite foundation embeddings + discovered vegetative fuel covariates
Earthquake Risk	0.8606	0.7992 [0.7923, 0.8059]	Mean minimum temperature, Sub-zero degree day counts, Latitude spatial embeddings
Landslide Risk	0.5614	0.7547 [0.7385, 0.7705]	US Census Gazetteer spatial geometry, Census Tract Housing Age, Income distributions, Building Density, Population Density
Avalanche Risk	0.1474	0.6882 [0.6767, 0.6988]	NOAA Severe Hail probability metrics, Convective available potential energy (CAPE) proxies, Terrain elevation