Supplementary Information of Global hydropower infrastructure and its environmental risks mapped by multimodal AI

Jiahao Li[1,2], Jiancheng Pan[1], Ramit Debnath[2], Dabo Guan[1], Xiaomeng Huang[1]*

1 Department of Earth System Science, Ministry of Education Key Laboratory for Earth System Modelling, Institute for Global Change Studies, Tsinghua University, Beijing, China

2 Collective Intelligence and Design Group, Centre for Human-Inspired AI, University of Cambridge, Cambridge, UK

Contents

1. Multi-source Data

This study utilized multi-source datasets encompassing remote sensing, hydrological, and environmental disciplines. A detailed list of these data is available in **Supplementary Table S1**, which provides comprehensive information on their sources and download links.

Supplementary Table 1 Multi-source data

Data Name	Full Name	Data Source	Download Source
WRI GPPD	World Resources Institute Global Power Plant Database	WRI	https://datasets.wri.org
GHPT	Global Hydropower Tracker	GEM	https://globalenergymonitor.org/projects/global-hydropower-tracker/
GloHydroRes	GloHydroRes	Scientific Data	www.nature.com/scientificdata
HR Imagery	High-Resolution Imagery	Google Earth	https://earthengine.google.com/
Sentinel-2	Sentinel-2	ESA	https://earthengine.google.com/
HydroSHEDS	HydroSHEDS database	World Wildlife Fund US	https://www.hydrosheds.org/products/hydrosheds
HydroBASINS	HydroBASINS database	World Wildlife Fund US	https://www.hydrosheds.org/products/hydrobasins
WDPA	World Database on Protected Areas	IUCN	https://www.protectedplanet.net/en
GEDI	GEDI L4A Raster Aboveground Biomass Density	NASA GEDI mission	https://developers.google.com/earth-engine/datasets/catalog/LARSE_GEDI_GEDI04_A_002_MONTHLY
Flood Hub	Flood Hub	Google	https://developers.google.com/flood-forecasting
GRRR	Google Runoff Reanalysis & Reforecast	Google	https://sites.research.google/gr/floodforecasting/resources/

- **WRI GPPD** The World Resources Institute Global Power Plant Database is an extensive, open-access repository that consolidates global power plant data to facilitate comparative analysis and insights. The database comprises approximately 28,500 power plants across 164 countries, encompassing a wide range of energy sources, including coal, natural gas, wind, solar, and hydroelectric power, which collectively represent a significant portion of the world's electricity generation capacity. The global power plant dataset from the World Resources

Institute (WRI) includes information on **nearly 7,000** operational hydropower plants.

- **Global Hydropower Tracker** The Global Hydropower Tracker provides a comprehensive, worldwide dataset that catalogs hydropower facilities globally. This tracker focuses on hydroelectric power plants with a minimum capacity of 45 megawatts (MW), encompassing all facilities meeting this threshold—including those in operational status, as well as announced, pre-construction, and under-construction units. Additionally, it incorporates select data for plants in non-active states, such as shelved, mothballed, retired, or canceled projects. It encompasses data from 145 countries/areas globally, tracking a total of **5,617 hydropower projects**. Among these, there is an **operating capacity of 1,216 GW**, while the prospective capacity reaches 1,113 GW.

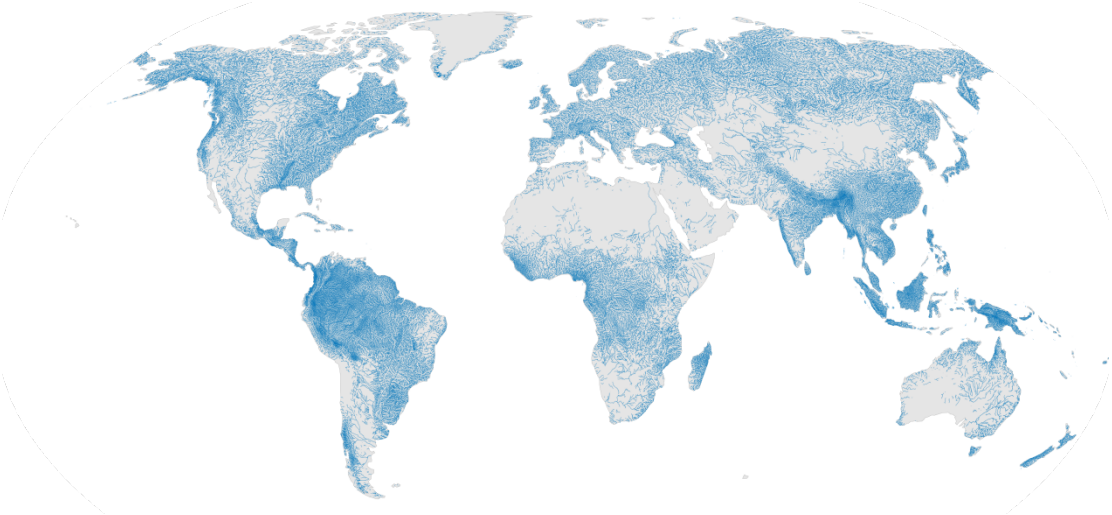
- **GloHydroRes** is a comprehensive, open-source global dataset developed to address the fragmentation of existing hydropower and reservoir data, where hydropower plant datasets lack reservoir attributes and reservoir datasets omit critical plant parameters, thereby hindering analyses of climate change impacts and water-energy nexus dynamics. Constructed by integrating multi-source open data, removing duplicates (prioritizing regional datasets), and linking plants to reservoirs via criteria of <10 km distance and lower plant elevation (using 15 arc-second DEM), followed by manual verification, it encompasses **7,775 hydropower plants** across 128 countries with a total installed capacity of **1,096.3 GW**, accounting for 79% (EIA, 2022) and 81% (IRENA, 2023) of global installed hydropower capacity.

- **Sentinel-2** The Sentinel-2 mission, developed under the European Space Agency's Copernicus program, comprises a dual-satellite system (Sentinel-2A and Sentinel-2B) designed for high-resolution multispectral imaging of Earth's surface. Operational since June 2015 for Sentinel-2A and March 2017 for Sentinel-2B, these satellites follow synchronized sun-synchronous orbits, enabling a combined five-day revisit interval at equatorial latitudes, with more frequent coverage at higher latitudes due to orbital convergence. Each satellite is equipped with a Multi-Spectral Instrument (MSI) capturing data across 13 spectral bands, spanning visible, near-infrared (NIR), and short-wave infrared (SWIR) wavelengths, with spatial resolutions ranging from 10 to 60 meters. This configuration supports detailed, frequent, and comprehensive monitoring of global land and coastal environments.

- **High-Resolution (HR) Imagery** High-resolution (HR) imagery aggregated by Google Earth (GE) is sourced from multiple satellite platforms, including WorldView-1, 2, 3, and 4, as well as GeoEye-1, resulting in variations in spatial and temporal resolution. The frequency of imagery updates is heterogeneous, driven by regional factors and data acquisition priorities. Areas of high population density or strategic importance, such as urban centers, typically experience more frequent updates, often on the order of several months or less in developed nations. Conversely, remote or less prioritized regions, particularly in developing countries, may experience update intervals that extend to a year or longer. Spatial resolution also varies significantly, with urban and high-interest areas benefiting from finer resolution imagery, while older or remote datasets often exhibit coarser resolution. Specifically, imagery at zoom level 19 corresponds to a spatial resolution of approximately 0.54 m, level 18 to 1.07 m, and level 17 to 2.15 m. These disparities underscore the influence of geographic and temporal factors on the availability and quality of GE HR imagery.

- **HydroSHEDS** database delivers a comprehensive suite of globally consistent hydrographic datasets, designed to underpin hydrological, ecological, and environmental research across multiple spatial scales—from local watershed studies to planetary water-resource assessments. Primarily derived from high-resolution Shuttle Radar Topography Mission (SRTM) digital elevation data, the database provides spatially contiguous information on critical hydrographic features, including river networks, watershed boundaries, drainage directions, and flow accumulation patterns, ensuring uniformity in data structure and quality across global biogeographic realms. HydroSHEDS systematically delineates all rivers draining catchments of 10 km^2 or more or exhibiting a mean annual discharge of greater than 0.1 m^3/s, yielding a detailed global river network that encompasses approximately 8.5 million individual river reaches with a cumulative total length of approximately 35.9 million km. A defining strength of the database lies in its standardized topological framework and explicit hierarchical connectivity across river basins—a feature that eliminates spatial discontinuities common in regional hydrographic datasets, enabling robust large-scale analyses. These analyses encompass assessments of freshwater ecosystem integrity and river connectivity, as well as evaluations of global water resource availability and vulnerability. By integrating rigorous data

processing protocols (including correction for topographic artifacts and validation against in situ hydrographic measurements) with open-access distribution, HydroSHEDS has established itself as a foundational resource for advancing interdisciplinary research on Earth's freshwater systems, supporting evidence-based conservation and water-management strategies at global and regional scales.

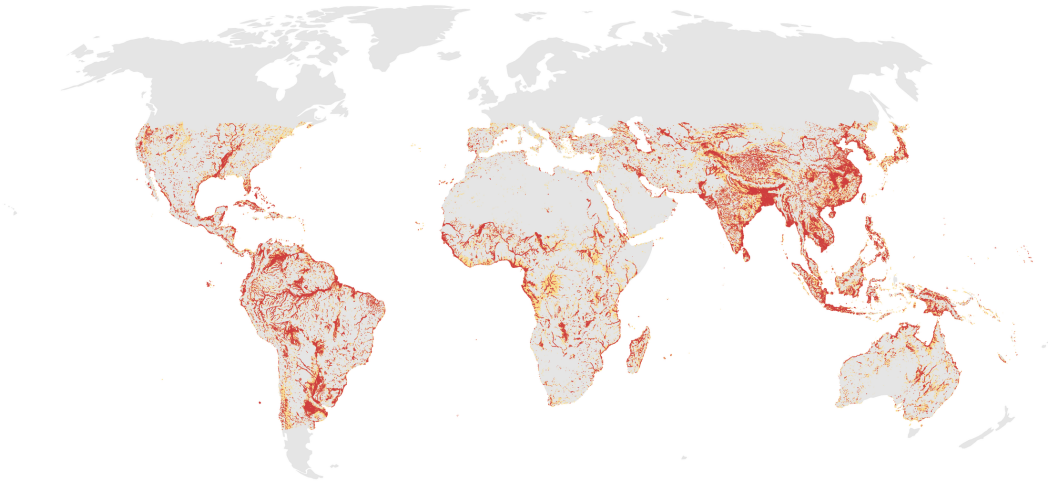


Supplementary Fig. 1 | Global hydrography map from HydroSHEDS.

- **HydroBASINS** is a global dataset comprising vectorized polygons that represent sub-basin boundaries, providing a detailed hierarchical subdivision of river basins across various spatial scales, ranging from tens to millions of square kilometers. Derived from the HydroSHEDS core layers at a 15-arc-second resolution, the product utilizes the 'Pfafstetter' coding system to facilitate the analysis of catchment topology, including upstream and downstream connectivity. At the finest scale of subdivision, it identifies sub-basins with a minimum upstream area of 100 km², ensuring comprehensive representation of the global hydrological network. With approximately 1 million polygons and an average size of 130.6 km², HydroBASINS covers 135 million km² of terrestrial surface, excluding Antarctica. The dataset primarily provides geometric attributes, including sub-basin area and distances from headwaters to ocean outlets.

- **GRRR** The Google Runoff Reanalysis & Reforecast (GRRR) dataset represents a pioneering advancement in global hydrological science, delivering comprehensive insights into streamflow derived from Google's state-of-the-art hydrological modeling framework—an

evolution of methodologies validated in prominent scientific literature. Spanning the entire globe with daily temporal resolution, GRRR encompasses three pivotal components: It offers streamflow reanalysis across more than 1 million hybas locations, chronicling hydrological dynamics from 1980 to 2023. It features streamflow reforecasts for the same extensive network of hybas locations between 2016 and 2022, with predictive lead times ranging from 0 to 7 days, enabling a robust evaluation of the efficacy of short-term hydrological forecasting. It integrates return period metrics derived from the reanalysis data, serving as critical benchmarks for assessing flood severity thresholds.



Supplementary Fig. 2 | Global map of flood-prone areas from Flood Hub (https://developers.google.com/flood-forecasting).

- **Flood Hub**, developed by Google Research, represents a transformative advancement in global riverine flood monitoring and early warning, leveraging state-of-the-art artificial intelligence (AI) and multi-source data fusion to address the critical gap in flood preparedness across data-scarce regions. This platform delivers actionable flood forecasts and hydrological insights for over 460 million people in more than 80 countries, with a particular focus on enhancing resilience in regions historically underserved by traditional hydrological monitoring networks. The platform provides multi-tiered flood intelligence: (1) 7-day ahead streamflow forecasts with daily resolution, achieving a 35% improvement in predictive skill (assessed via Nash–Sutcliffe efficiency) in 2025 compared to prior model iterations; (2) dynamic inundation maps

for select basins, visualizing spatially explicit flood extents; and (3) return period-based severity thresholds, derived from historical reanalysis datasets (e.g., GRRR), to contextualize flood magnitudes. These outputs are updated daily, ensuring timeliness for emergency response and long-term risk assessment.

- **WDPA** The World Database on Protected Areas represents the most comprehensive global repository of information on terrestrial and marine protected areas to date, compiled and maintained by the UN Environment Programme World Conservation Monitoring Centre (UNEP–WCMC) in collaboration with the World Commission on Protected Areas (WCPA) of the IUCN. As of its January 2025 update, the WDPA encompasses 305,195 total records, including **303,312 protected areas**, spanning 244 countries and territories (with spatial data composed of 293,259 polygons and 11,936 points) and provides standardized spatial and attribute data essential for tracking global progress in biodiversity conservation. Notably, the WDPA is updated monthly with submissions from governments, non-governmental organizations, landowners, and communities to ensure data timeliness; it also synergizes with complementary databases, such as the World Database on Other Effective Area-Based Conservation Measures (WDOECM) and the Global Database on Protected Area Management Effectiveness (GD-PAME), to deliver a holistic perspective on conservation efforts.

Supplementary Fig. 3 | The 303,312 protected areas defined by the IUCN.

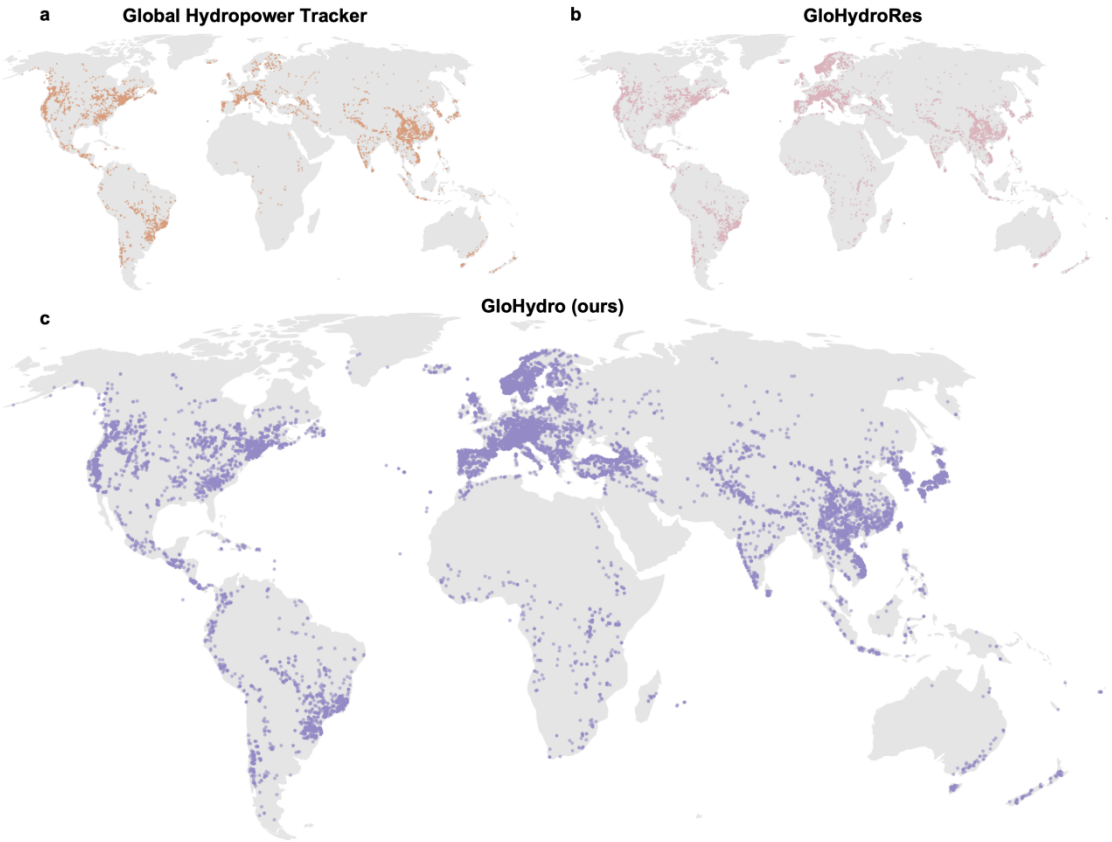
- **GEDI L4A Raster Aboveground Biomass Density** This dataset includes predictions of

aboveground biomass density (AGBD; in Mg/ha) and corresponding standard error estimates for each geolocated laser footprint, derived from Global Ecosystem Dynamics Investigation (GEDI) Level 4A (L4A) Version 2. Sub-orbits organize the data granules. Height metrics obtained from simulated waveforms, calibrated with field-based AGBD measurements across multiple regions and plant functional types (PFTs)—including deciduous broadleaf, evergreen broadleaf, evergreen needleleaf, deciduous needleleaf trees, and a composite category of grasslands, shrubs, and woodlands—were compiled to create the calibration dataset. Additionally, the GEDI02_A Version 2 algorithm settings for evergreen broadleaf trees in South America have been adjusted to minimize false positive errors caused by incorrectly selecting waveform modes above ground elevation as the lowest mode.

2. Comparison with the Existing Hydropower Plant Inventories

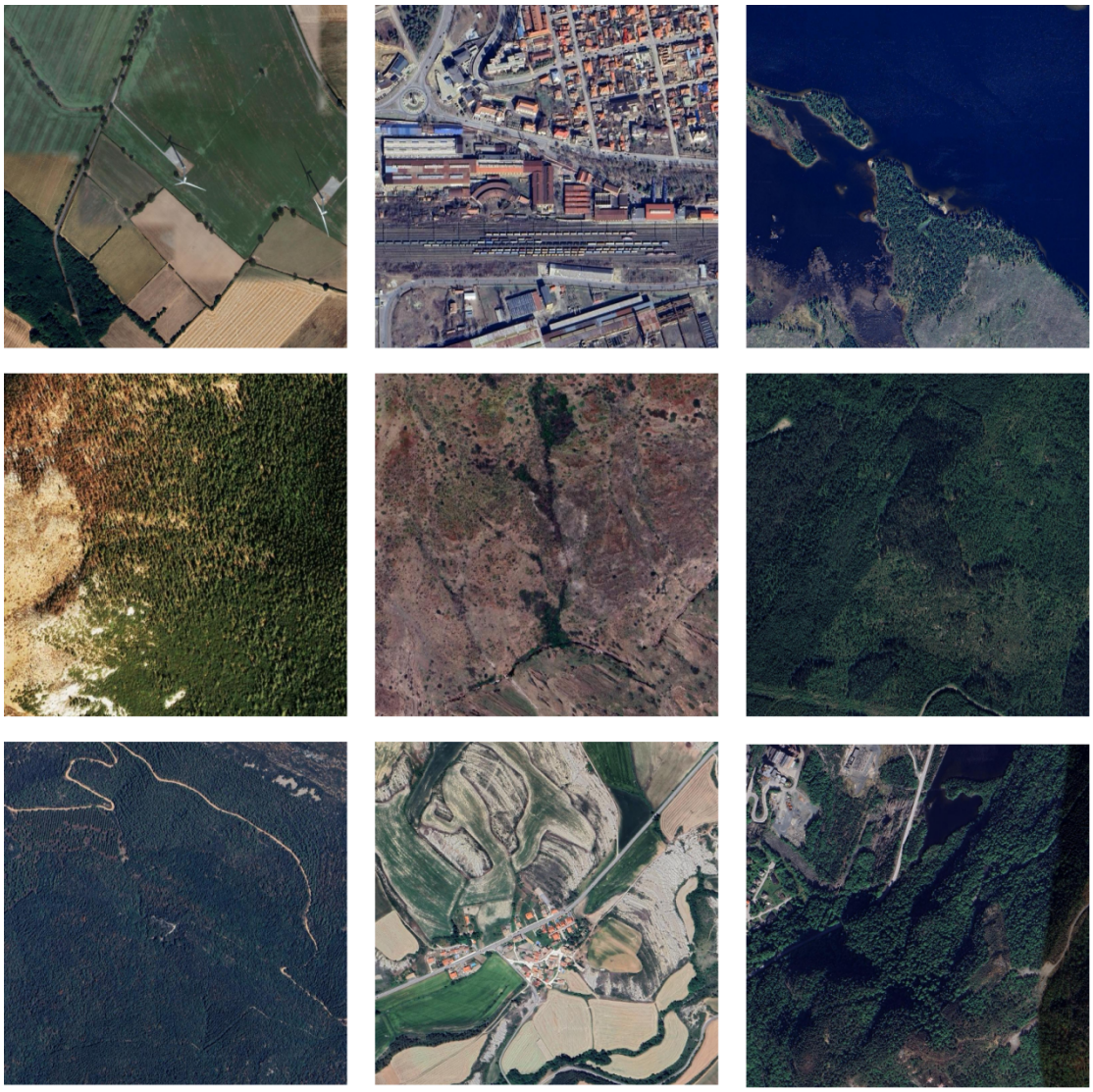
2.1 Comparison of GloHydro with existing hydropower plant inventories

The results from **GloHydro** were compared against existing public inventories. As presented in **Supplementary Fig. 4**, which compares the number and spatial distribution of records, 55.7% of the plants in GloHydro were not reported in the existing public hydropower inventories. **Supplementary Fig. 7** presents a comparison of the global hydropower installed capacity as reported by different stockpiles.

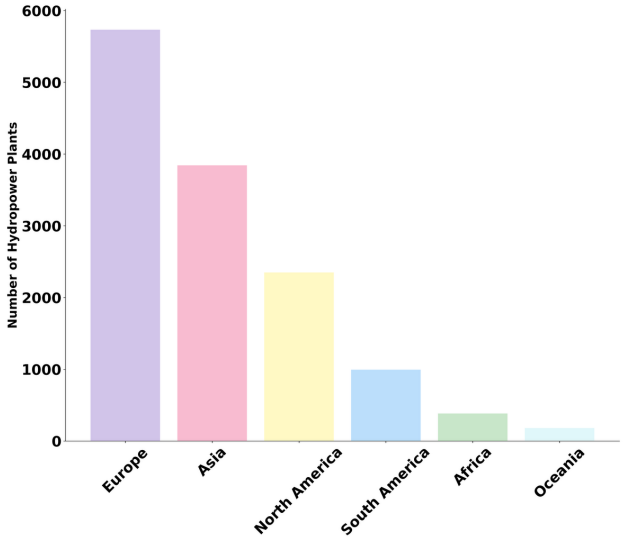


Supplementary Fig. 4 | Comparison of GloHydro with existing hydropower plant inventories.

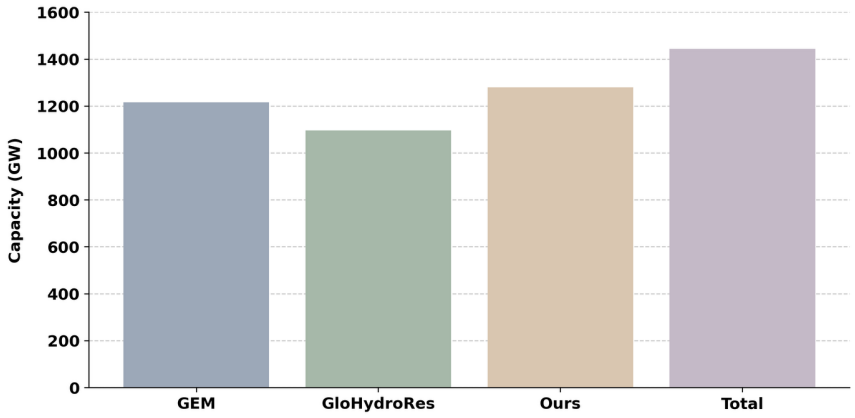
We manually verified all locations in the existing public inventories (WRI GPPD, GHPT, and GloHydroRes). We found that a significant proportion of sites were erroneous, likely due to the inclusion of planned projects or compilation errors, despite these inventories reportedly involving manual checks. Some examples of incorrect locations are shown in **Supplementary Fig. 5**. **Consequently, the validated hydropower locations plotted in Figure 1 in the main text represent a curated set of correct sites, integrated through a rigorous process of manual selection from three public inventories.**



Supplementary Fig. 5 | Examples of incorrect locations in the existing public inventories.



Supplementary Fig. 6 | The number of hydropower plants on each continent in GloHydro.



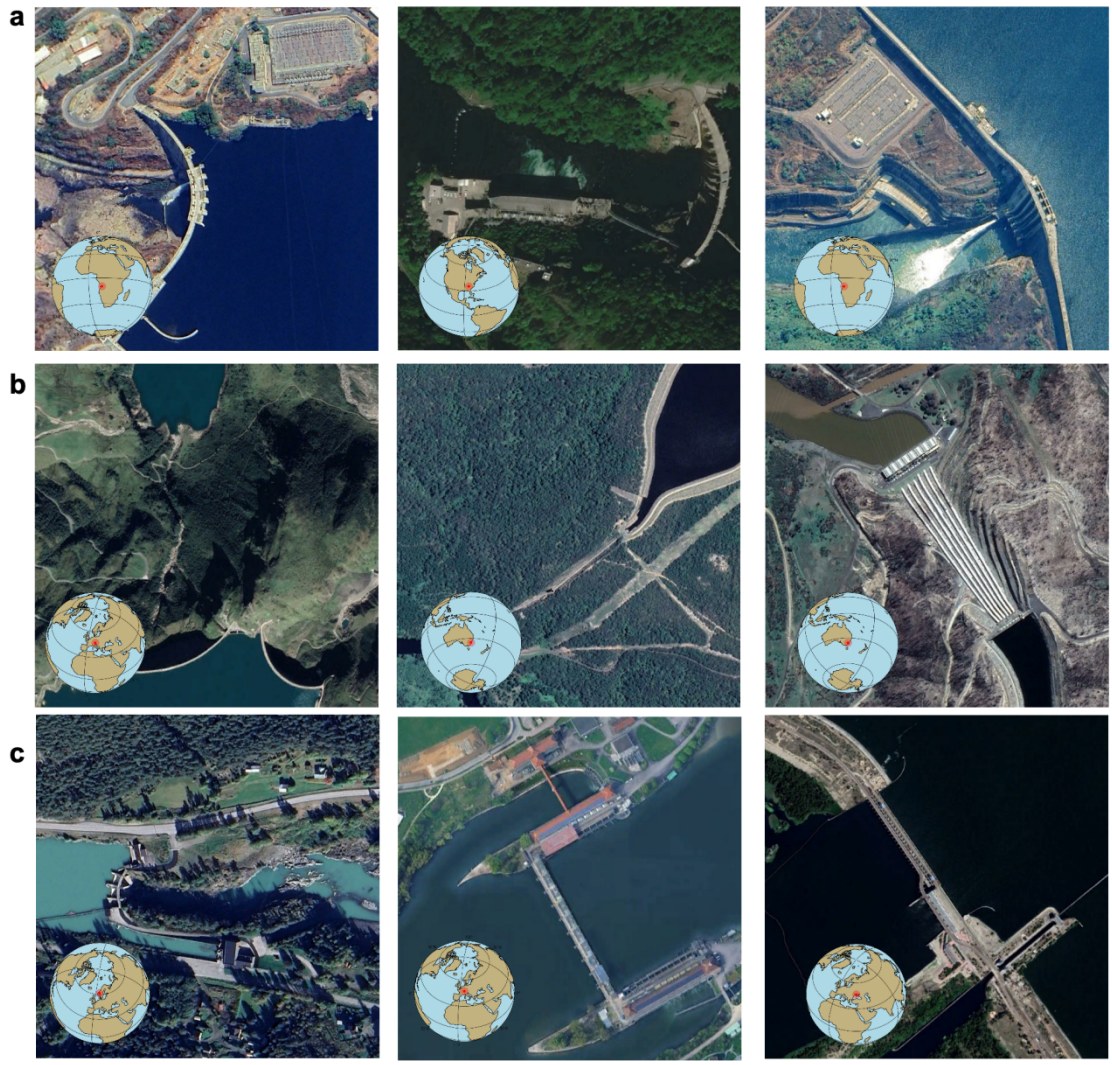
Supplementary Fig. 7 | Comparison of installed capacity across hydropower plant inventories.

2.2 Hydropower plant types and their remote sensing imagery

- **Storage Hydropower Plants (STO)** Storage hydropower plants are characterized by the use of large reservoirs to store water. These plants typically operate by releasing water from the reservoir through turbines to generate electricity when demand is high. The ability to store water during periods of low demand and release it during peak demand offers flexibility in grid management. Additionally, storage hydropower plants are capable of providing a stable and reliable source of power, making them a crucial component in many national energy systems.

- **Pumped Storage Hydropower Plants (PS)** Pumped storage hydropower plants (PS) function as a form of grid-scale energy storage, utilizing two reservoirs at different elevations. During periods of low energy demand, excess electricity is used to pump water from the lower reservoir to the upper one, storing potential energy. During periods of high energy demand, water is released from the upper reservoir to generate electricity as it flows back down through turbines. This cycle allows pumped storage plants to provide rapid-response power, contributing to grid stability by balancing fluctuations in demand and supply. Although they are considered a form of renewable energy, the environmental impacts of pumped storage, including habitat disruption and concerns about water quality, remain areas of focus for ecological assessments.

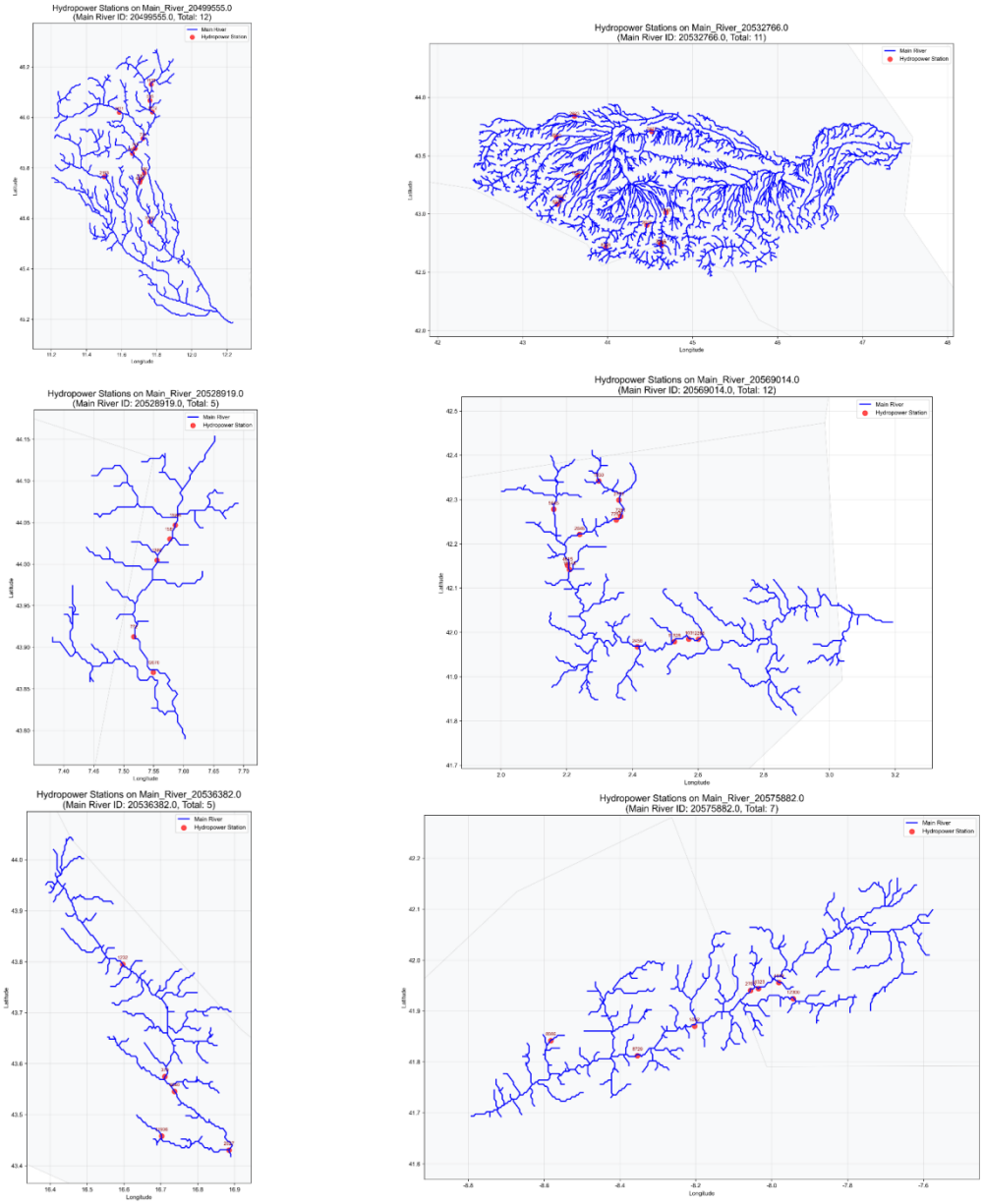
- **Run-of-River Hydropower Plants (ROR)** Run-of-river hydropower plants (ROR) operate without the need for large reservoirs, instead harnessing the natural flow of rivers or streams to generate electricity. Water is diverted from the river to flow through turbines, typically with minimal alteration to the river's natural course. This type of hydropower plant is often considered more environmentally friendly than storage or pumped storage plants, as it avoids the large-scale land inundation and ecosystem disruption associated with reservoirs. However, the efficiency and capacity of run-of-river plants are generally lower than those of storage plants due to the variable nature of river flow, which depends on seasonal and climatic factors. Despite this, ROR plants contribute to sustainable energy generation, particularly in regions with suitable river systems.



Supplementary Fig. 8 | Hydropower plant types and their remote sensing imagery. a, Storage Hydropower Plants. b, Pumped Storage Hydropower Plants. c, Run-of-River Hydropower Plants.

3. Hydropower Clusters and Transboundary Developments

By leveraging the precise locations of hydropower plants from **GloHydro** and global river network data from HydroSHEDS and HydroBASINS, we identified clustered hydropower developments worldwide. This analysis is crucial for identifying global trends in hydropower clustering and transboundary development, as illustrated in Supplementary Fig. 9, which outlines the methodology used to define and analyze these clusters.



Supplementary Fig. 9 | Clustering development of hydropower plants across river basins.

Supplementary Fig. 10 displays a cluster of hydropower plants in Europe, along with their corresponding remote sensing imagery. These facilities are distributed along a river and are situated near populated towns, serving as a vital energy source for these communities.

Supplementary Fig. 10 | A cluster of hydropower plants in Europe.

As listed in **Supplementary Table 2**, international organizations have played a crucial role in promoting transboundary hydropower development through mechanisms such as interstate coordination, equitable benefit-sharing, and the formulation of supportive policies.

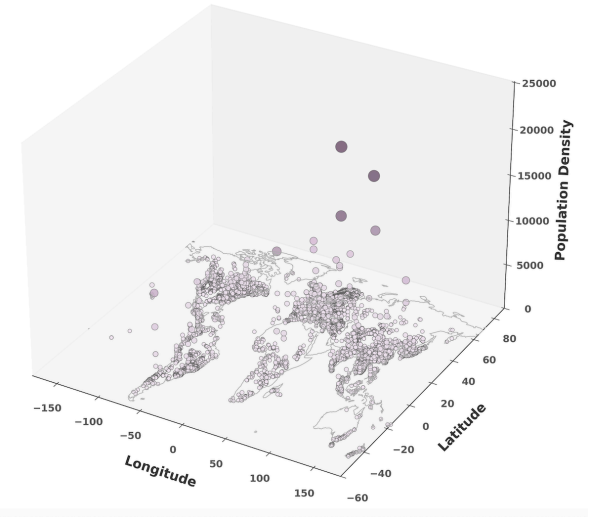
Supplementary Table 2: International organizations for transboundary hydropower

Name	Data Source	Download Source
HYDROPOWER EUROPE	European Union	https://hydropower-europe.eu/
UN-Water	United Nations	https://www.unwater.org/
Low Impact Hydropower Institute	LIHI	https://lowimpacthydro.org/
International Hydropower Association	IHA	https://www.hydropower.org/
Mekong River Commission	MRC	https://www.mrcmekong.org/

4. Global Coupling of Hydropower Plants and Protected Areas

4.1 Hydropower distribution and population density

The statistical analysis reveals a distinct relationship between global hydropower distribution and population density. As shown in **Supplementary Fig. 11**, 77.98% of global hydropower plants are located in regions with a population density below 100, whereas only 2.04% are situated in densely populated areas with a population density of more than 1,000 people per square kilometer.



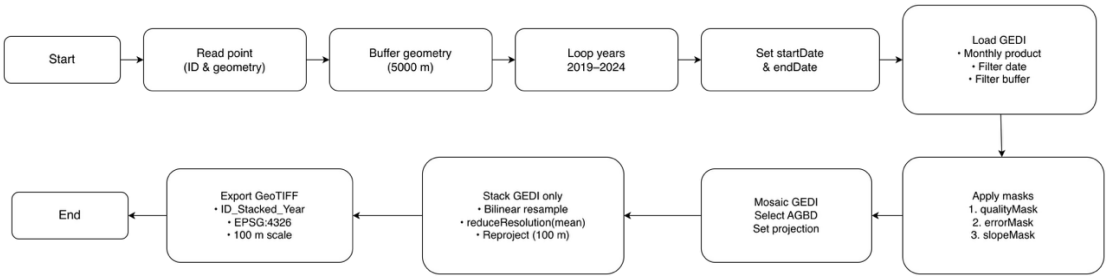
Supplementary Fig. 11 | Global hydropower plant distribution in relation to population density.

4.2 Aboveground biomass density processing

The workflow of **Supplementary Fig. 12** was designed to generate annual GEDI-based biomass products for the spatial neighborhood surrounding each hydropower plant. For every plant, its geographic coordinates and unique identifier were first extracted, and a circular buffer with a 5-km radius was delineated to define the analysis area. The procedure was repeated for each year from 2019 to 2024, with annual start and end dates initialized to constrain data filtering.

For each hydropower plant and year, the GEDI Level 4A monthly product was queried and restricted to both the corresponding temporal window and the buffered spatial extent. To ensure the quality and reliability of canopy biomass estimates, three sequential filtering steps were applied: (1) a quality mask retaining only observations flagged as valid, (2) an uncertainty mask excluding samples with relative biomass error (agbd_se/agbd) greater than 0.2, and (3) a terrain-based mask removing observations located on slopes exceeding 30°, with slope derived from the COPERNICUS GLO-30 digital elevation model. The filtered GEDI shots were subsequently mosaicked to produce a continuous annual AGBD layer while preserving the native projection.

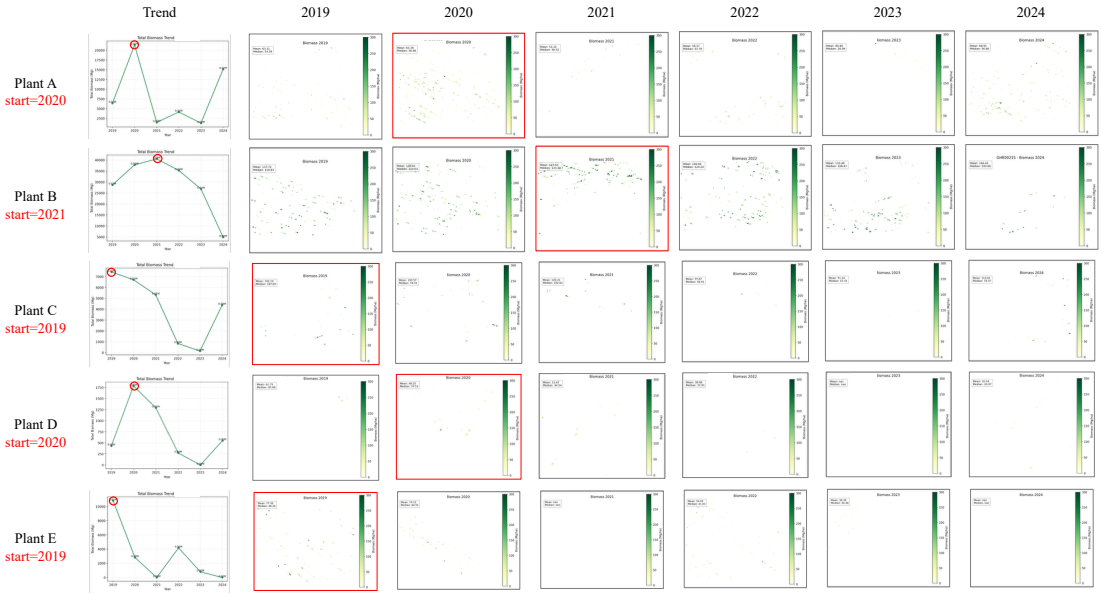
The resulting GEDI mosaic was resampled using bilinear interpolation, aggregated using a mean reducer with a pre-defined maximum-pixel limit, and reprojected to a uniform spatial resolution of 100 m. The final processed biomass layer for each hydropower plant and year was exported to Google Drive as a GeoTIFF in EPSG:4326 geographic coordinates. This complete workflow was executed iteratively across all hydropower plants and all years in the analysis period.



Supplementary Fig. 12 | Workflow for Annual GEDI-Based Biomass Processing Around Hydropower Plants.

For each hydropower plant, we performed a multi-year analysis of aboveground biomass (AGB) within its 5-km buffer zone using annual AGBD raster layers exported from Google Earth Engine. The workflow first scanned all files in the dataset directory. It automatically grouped them by hydropower plant identifier and year extracted from the filenames, thereby reconstructing a complete biomass time series for each plant from 2019 to 2024. For every annual raster, the biomass band was extracted, and invalid or missing pixels were removed. To convert pixel-level biomass density (Mg ha^{-1}) to total biomass (Mg), the geographic area of each pixel was calculated from the raster's affine transformation; when such metadata were missing or unreliable, a fallback pixel size of 100 m—consistent with the export resolution—was applied to ensure internal consistency across

years. The total biomass for each plant-year combination was then calculated by summing all valid pixel values, multiplied by the corresponding pixel area in hectares. To characterize interannual variations, the annual total biomass values were plotted as a time series, with each year labeled to facilitate visual assessment of increases, declines, or fluctuations. This automated and plant-centered workflow offers a reproducible and internally consistent approach for quantifying multi-year biomass dynamics in the landscapes surrounding hydropower plants.

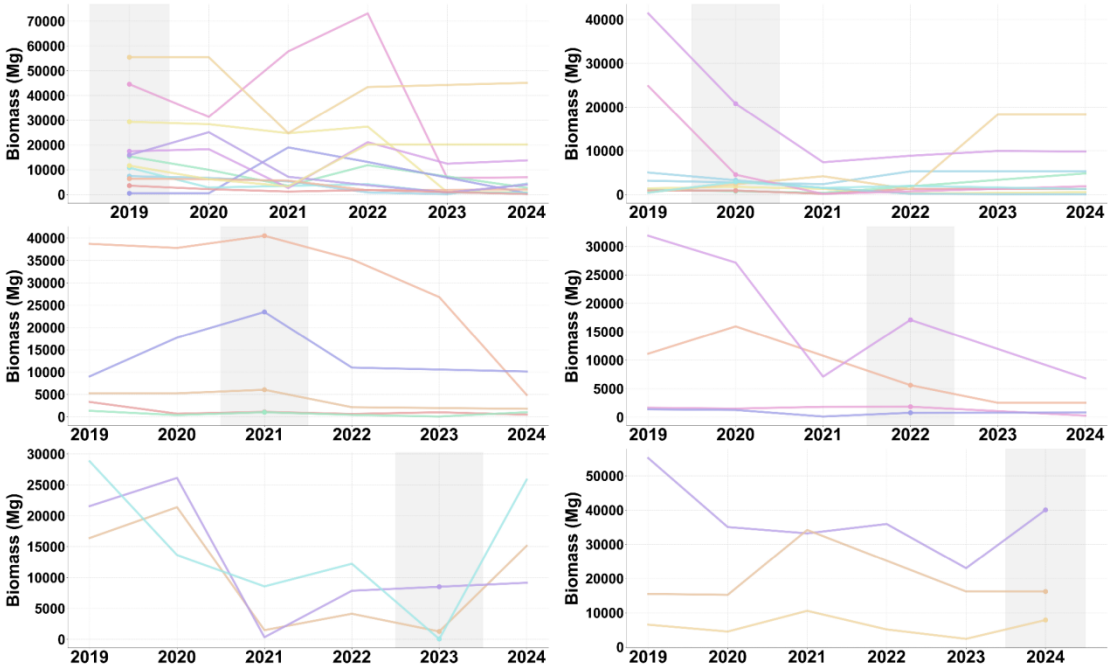


Supplementary Fig. 13 | Biomass trends within a 5-kilometre radius of hydropower plants. The red colour indicates the operational period of the hydroelectric power plant.

As shown in **Supplementary Fig. 13**, aboveground biomass within a 5-km radius of the hydropower plant exhibited a noticeable decline during the initial period following construction. This pattern is consistent with previous findings that infrastructure development can impose short-term disturbances on surrounding ecosystems, potentially through the removal of vegetation, land modification related to construction, or adjustments to local hydrological regimes. However, this decline does not appear to be persistent. In the case of GHR2022, biomass levels began to increase again in 2023, suggesting a degree of ecological recovery after the early-stage disturbance. Such recovery may reflect a combination of factors, including the regrowth of natural vegetation, the inherent resilience of the local ecosystem, and possible mitigation or management actions implemented during the operational phase—such as riparian restoration, reduced human disturbance, or improved soil and water conservation practices. It is essential to note that, although the timing of

biomass changes aligns with the construction and operation phases of the hydropower plant, the observed associations do not, in themselves, establish causal mechanisms; additional ecological indicators or management records would be required to disentangle the underlying drivers fully. Overall, **Supplementary Fig. 13** illustrates a dynamic process in which the surrounding ecosystem initially experiences disturbance but subsequently exhibits signs of recovery, highlighting potential lagged ecological responses and the role of management interventions in modulating the environmental impacts of hydropower development.

Supplementary Fig.14 presents a detailed, site-by-site tracking analysis, illustrating the temporal trajectory of aboveground biomass density relative to the commissioning year of each hydropower plant. While the comprehensive analysis is presented in the main text, this figure displays the results for individual cases. It reveals that hydropower development consistently leads to a marked reduction in biomass density; however, the subsequent recovery is highly heterogeneous across sites. This variability underscores the ecological diversity of hydropower impacts globally, which is influenced by local environmental conditions, construction practices, and regulatory policies.

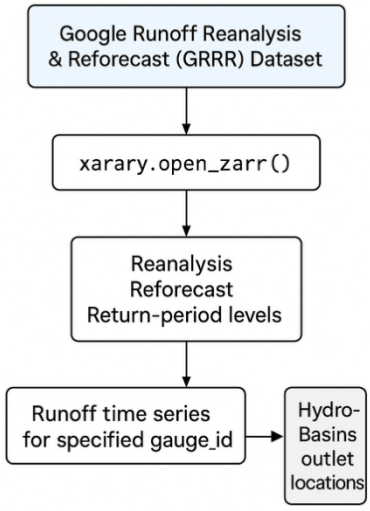


Supplementary Fig. 14 | Impacts of hydropower plant construction on the temporal dynamics of aboveground biomass density.

Notably, the impact of hydropower plants on aboveground biomass density is not necessarily confined to the year of their completion. Because the construction of a hydropower plant typically spans several years, substantial ecological disturbance can occur during this period. As shown in **Supplementary Fig. 14**, declines in aboveground biomass density were already evident in the years preceding the commissioning of hydropower plants. This observation motivated our analysis in the main text, where we specifically tracked and examined changes in biomass density over the three years preceding the completion of new hydropower projects.

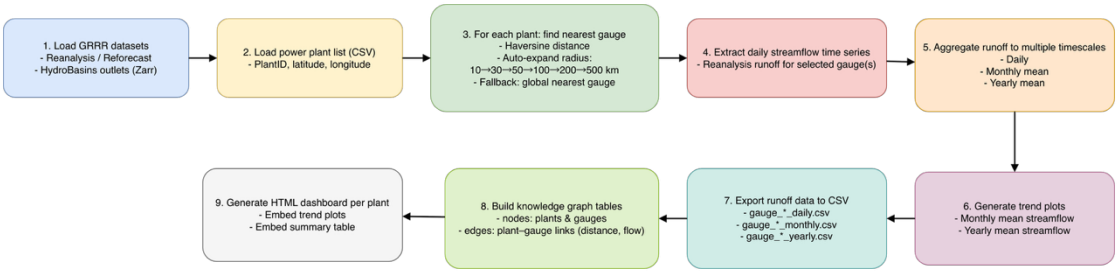
5. Runoff Alteration and Flood Risk

5.1 Runoff trend processing



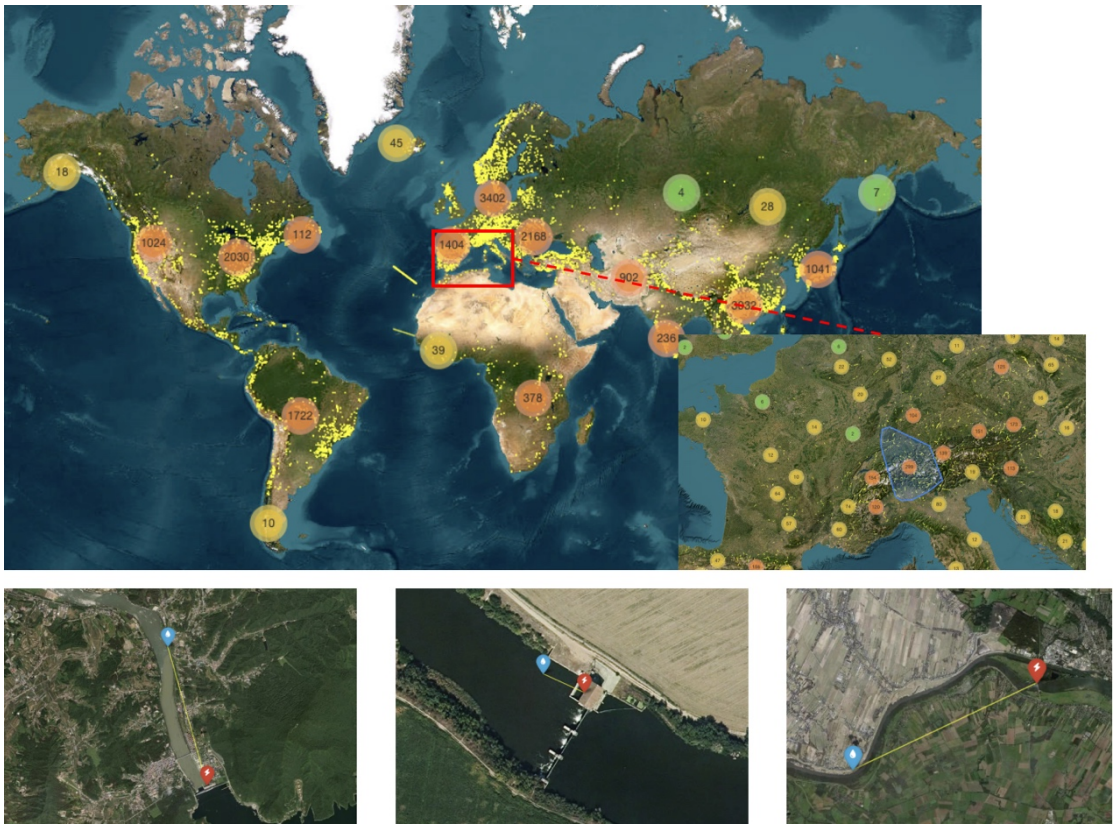
Supplementary Fig. 15 | Runoff extraction workflow.

The runoff data used in this study were obtained from Google's global Runoff Reanalysis & Reforecast (GRRR) dataset, which provides daily streamflow predictions for more than one million HydroBasins watershed outlets based on a state-of-the-art hydrologic modeling framework **(Supplementary Fig. 15)**. The dataset is distributed in cloud-optimized Zarr format and accessed directly via the xarray.open_zarr() interface, allowing anonymous, on-demand retrieval of its three primary components: (i) daily reanalysis streamflow (1980–2023), (ii) short-term reforecasts issued between 2016–2022 with lead times of 0–7 days, and (iii) return-period streamflow thresholds derived from the reanalysis. All datasets are indexed by a unique gauge_id corresponding to the outlet of each HydroBasins watershed, with reanalysis records organized along a daily time dimension and reforecast data structured by issue_time and lead_time. After loading the dataset, the notebook extracts the complete runoff time series for any specified gauge_id and, optionally, links it to approximate geographic coordinates using an auxiliary (unofficial) set of HydroBasins outlet locations included in the appendix. This indexing structure and data-access pipeline enable efficient retrieval of historical discharge, forecast trajectories, and severity thresholds for any watershed globally, providing a coherent and scalable basis for hydrological time-series analysis, flood-risk assessment, and broader environmental modeling.



Supplementary Fig. 16 | Global runoff extraction workflow.

After accessing the GRRR runoff dataset, we developed an automated global pipeline **(Supplementary Fig. 16)** to systematically extract hydrologic information for all hydropower plants worldwide by linking each facility to its nearest HydroBasins outlet. For every plant location (latitude, longitude), the workflow computes great-circle distances to all basin outlets using the Haversine formula. It employs a multi-scale adaptive search strategy, starting with a 10 km radius and progressively expanding the search radius (30, 50, 100, 200, 500 km) until at least one valid gauge is identified. If no outlet is found within the predefined radii, the algorithm falls back to the globally nearest gauge to ensure complete spatial coverage. Once the nearest gauge is identified, the corresponding daily streamflow record is retrieved directly from the GRRR reanalysis dataset and further aggregated into monthly and yearly mean discharge values, enabling multi-timescale hydrological characterization. For each plant, the pipeline automatically generates visual summaries—including monthly and annual streamflow trends—and exports structured tables that report key hydrological metrics, such as daily means, aggregated totals, and long-term averages. To facilitate knowledge-graph applications and large-scale network analysis, the system simultaneously constructs node–edge files representing plants and their associated river gauges as a relational graph. Collectively, this automated workflow enables high-throughput, globally consistent extraction, aggregation, and visualization of river runoff information for hydropower-relevant locations, providing a scalable and reproducible foundation for assessing hydrological variability associated with energy infrastructure.



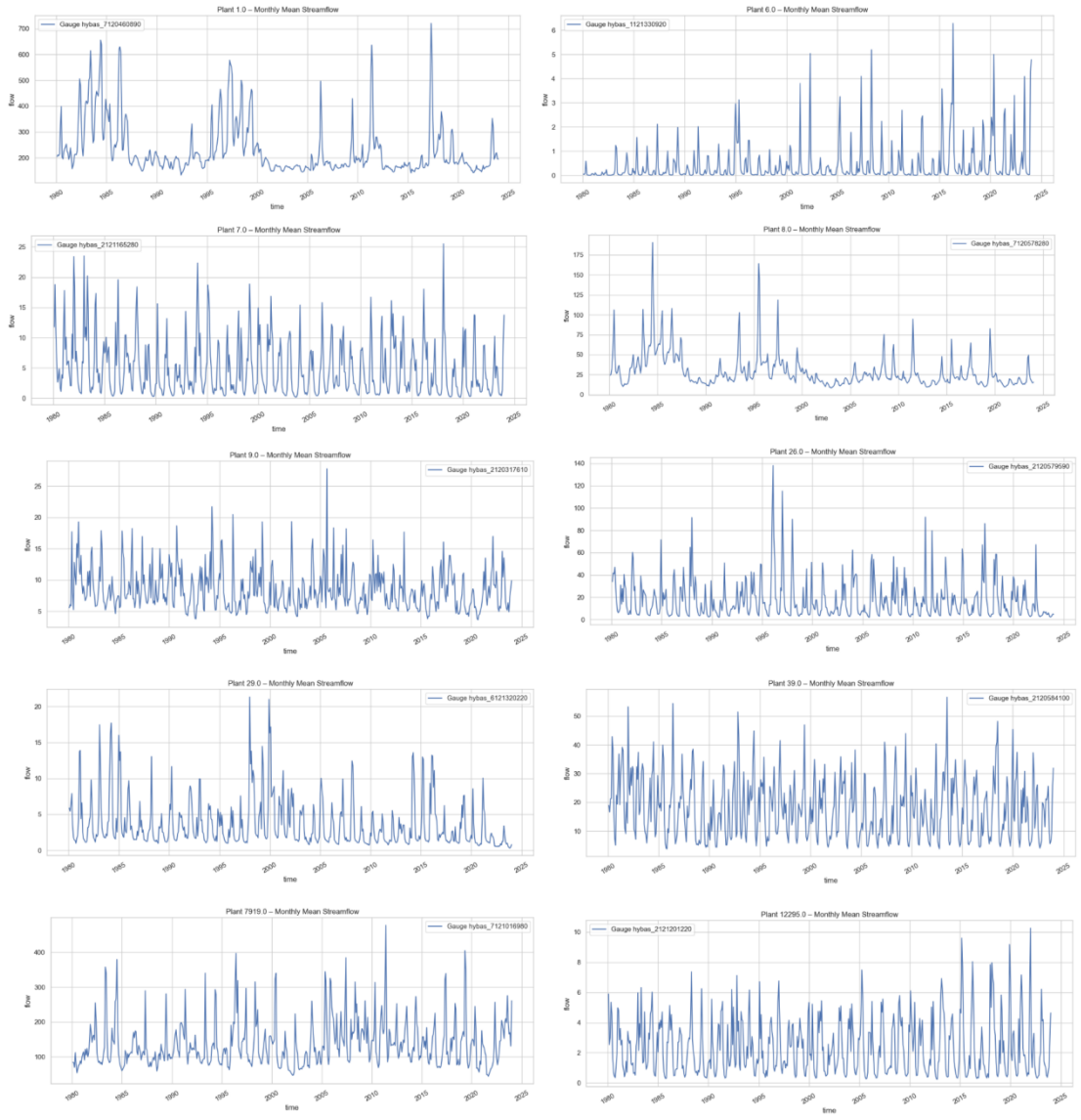
Supplementary Fig. 17 | Hydropower Topology Visualization.

We visualize the global hydropower topology **(Supplementary Fig. 17)**. The program loads hydropower plant nodes and water-flow relationships from a knowledge graph dataset, generating a global-scale topological map that illustrates both the spatial distribution and structural connectivity within the hydropower system. The visualization is rendered on top of the ESRI World Imagery satellite basemap, which provides high-resolution geographic context. All hydropower nodes are imported and displayed using a MarkerCluster mechanism to ensure efficient interaction even when visualizing a large number of spatial points worldwide. Each node is plotted according to its latitude and longitude, and different icon styles are applied to distinguish between node types (e.g., power plants versus reservoirs). For edge representation, the program identifies the nearest connection for each node from among all potential water-transfer paths and visualizes only this closest edge. This approach preserves the most representative hydrological linkage while avoiding excessive visual clutter that would result from rendering all available edges. Each edge includes information such as distance and flow magnitude, which is displayed through interactive tooltips. Finally, the program automatically adjusts the map's extent based on the geographic bounds of all nodes, ensuring that

the resulting visualization fully encompasses the global hydropower network. The final topological map is exported as an interactive HTML file, enabling dynamic exploration through a standard web browser. This visualization tool provides an intuitive and effective means of examining hydropower system structure, supporting analyses related to water-resource dispatching, flow pattern assessment, and global hydropower network modeling. This visualization program is accessible online via **https://glohydro.cn**.

5.2 Monthly runoff data of the global hydropower plant

Our study provides monthly runoff data (1980-2023) for rivers at all global hydropower plant locations. This data not only underpins our analysis of long-term streamflow trends but also enables the prediction of monthly power generation for each facility. The data are publicly accessible via **https://glohydro.cn**, and **Supplementary Fig. 18** displays some cases, clearly illustrating the monthly runoff dynamics of the hydropower plants.



Supplementary Fig. 18 | Cases of monthly runoff time series for hydropower plants (1980–2023).

6. Models and Training Details

To overcome the limitations of conventional vision-based identification models that rely primarily on single-modality visual cues, we present HydroVLM, an identification paradigm designed to overcome the limitations inherent in conventional approaches that rely predominantly on single visual features[1]. By harnessing the cross-modal comprehension and scene reasoning capabilities of Vision-Language Models (VLMs)[2,3], we establish a novel recognition paradigm founded on image-text associations. The training methodology employs Low-Rank Adaptation (LoRA), a technique that enhances training efficiency by utilizing low-rank matrices[4.] This approach involves freezing the majority of parameters in the pre-trained model while updating only a small set of low-rank adaptive parameters.

We fine-tuned the **Qwen2.5-VL-32B**[5,6] base model with LoRA on a meticulously curated, annotated dataset of 3,963 diverse hydropower samples, each comprising high-resolution remote-sensing imagery that captures the full spatial extent of the facility together with structured textual descriptions. **Rather than relying solely on static annotations, we implemented an expert-in-the-loop calibration workflow**: the initially trained model was used to automatically label additional imagery, and domain experts corrected and augmented these machine-generated annotations. Expert corrections and the newly validated labels were incrementally incorporated into subsequent LoRA fine-tuning rounds, with model performance tracked on a held-out validation set to guide stopping and sampling decisions. Training and iterative fine-tuning were performed on 8 NVIDIA A100 GPUs; throughout this cyclical process HydroVLM framed hydropower feature identification as an image–text association task, exploiting the cross-modal alignment and scene-reasoning strengths of vision–language models while progressively reducing manual labeling effort and improving predictive precision.

Given an input image I and a set of candidate textual descriptors $\mathcal{T} = \{T_1, T_2, \dots, T_K\}$, the visual encoder $E_v(\cdot)$ and text encoder $E_t(\cdot)$ of the underlying VLM project image and text into a shared semantic space:

$$\mathbf{v} = E_v(I), \mathbf{t}_k = E_t(T_k), k = 1, \dots, K. \quad (1)$$

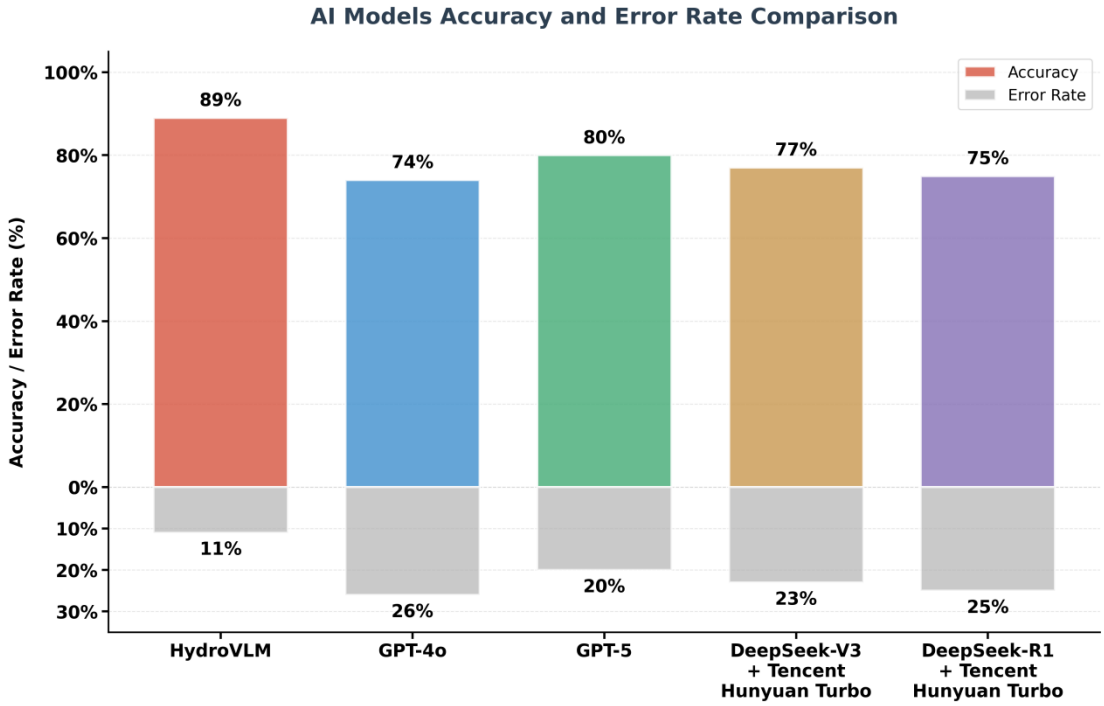
The association strength between the image and each textual descriptor is computed using a similarity function $S(\cdot,\cdot)$, typically cosine similarity:

$$s_k = S(\mathbf{v}, \mathbf{t}_k). \tag{2}$$

The predicted hydropower category $\hat{y}$ is determined by selecting the descriptor with the highest similarity score:

$$\hat{y} = \arg\max_k s_k. \tag{3}$$

This paradigm enables HydroVLM to utilize rich multimodal contextual cues rather than relying on isolated visual features, providing improved robustness and interpretability for complex remote sensing scenes.



Supplementary Fig. 19 | Comparison of HydroVLM and existing mainstream VLM models.

HydroVLM demonstrated robust performance metrics, achieving an accuracy of 89% and a recall rate of 93% on our primary task of identifying hydropower facilities. To rigorously benchmark HydroVLM's capability specifically for hydropower remote sensing image recognition, we constructed a dedicated test dataset comprising diverse samples of hydropower facilities. This

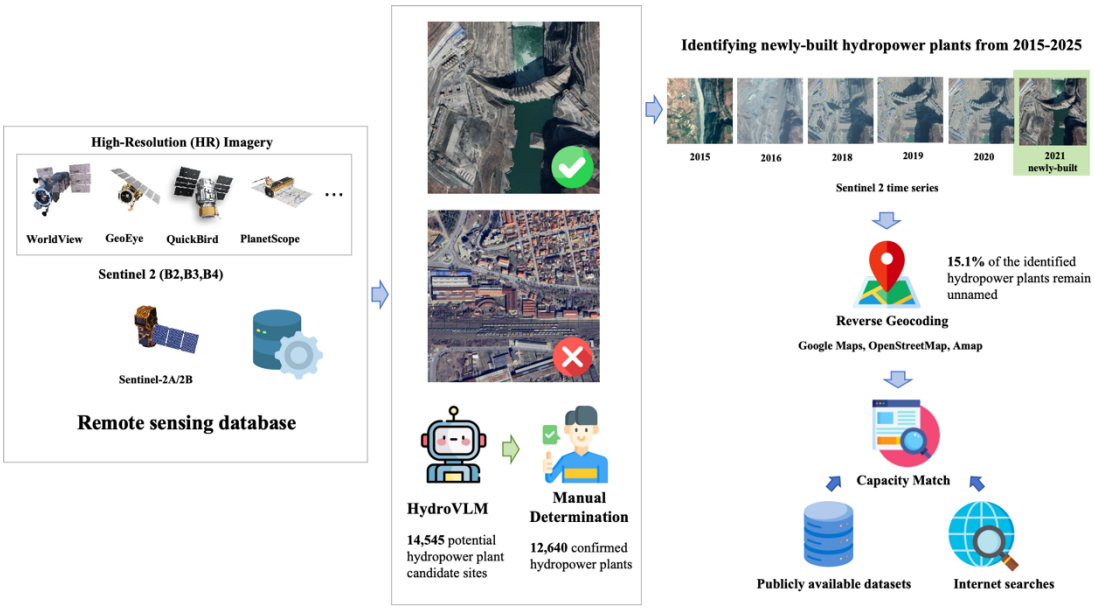
dedicated evaluation dataset shall consist of 254 meticulously curated samples, including 200 positive instances and 54 erroneous negative samples. Each sample features high-resolution remote sensing imagery that captures the complete spatial extent of the facility. We evaluated HydroVLM against several state-of-the-art generalist multimodal foundation models:

Supplementary Table 3: State-of-the-art generalist multimodal foundation models

Name	Source	Access
GPT-4o	OpenAI	https://chatgpt.com/
GPT-5	OpenAI	https://chatgpt.com/
DeepSeek-V3 + Tencent Hunyuan Turbo	DeepSeek AI Team	https://yuanbao.tencent.com/
DeepSeek-R1 + Tencent Hunyuan Turbo	DeepSeek AI Team	https://yuanbao.tencent.com/

Recognizing that DeepSeek-V3[7] and DeepSeek-R1[8] lack native image understanding[9], it was integrated with Tencent Hunyuan Turbo's visual capabilities via a joint inference approach (https://yuanbao.tencent.com/ for technical details).

Supplementary Fig. 19 presents the comparative performance of the VLM models[10]. Notably, the advanced reasoning capabilities of DeepSeek-R1[11] did not confer a discernible advantage in identifying hydropower targets. This finding implies that for specialized domains, enhancing generic reasoning prowess alone is insufficient; instead, VLM training must be strategically tailored, prioritizing the integration of domain-specific knowledge and architectural suitability over mere increases in reasoning complexity. To elucidate the underlying mechanisms of VLMs and to explore more advanced approaches, additional experiments would be required. In this study, however, our primary goal was to develop a practical, automated alternative to manual mapping, assisting us in the top-down identification of hydropower plants.



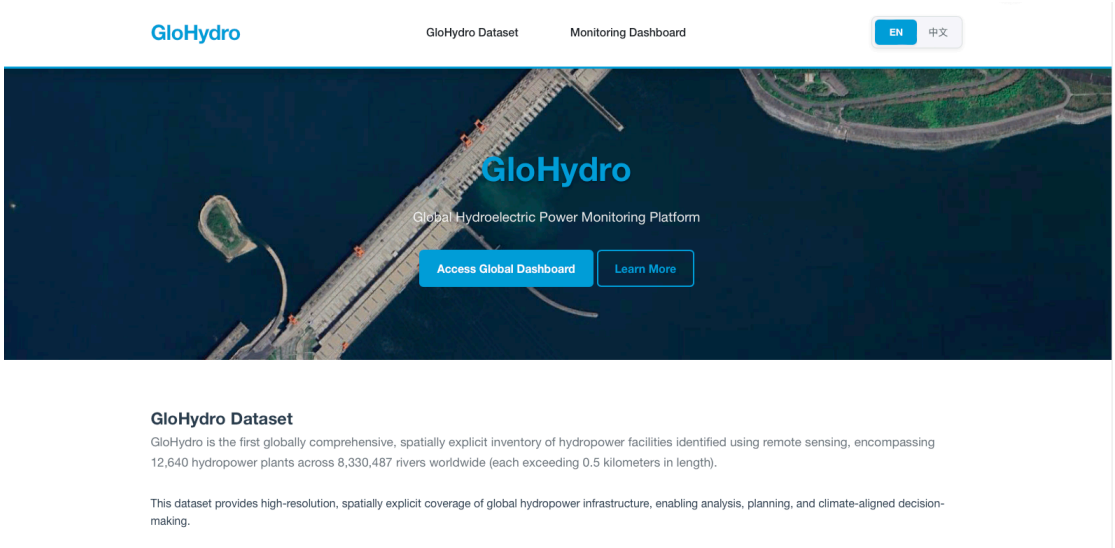
Supplementary Fig. 20 | Overall multimodal AI framework of HydroVLM.

As illustrated in **Supplementary Fig. 20**, the HydroVLM establishes an end-to-end workflow for hydropower station identification. The system consists of four main stages: (1) precise localization of hydropower plants from remote sensing imagery; (2) identification of their construction year; (3) name matching based on geospatial coordinates; and (4) retrieval of installed capacity information. This comprehensive pipeline enables automated and accurate recognition of hydropower plants at a global scale.

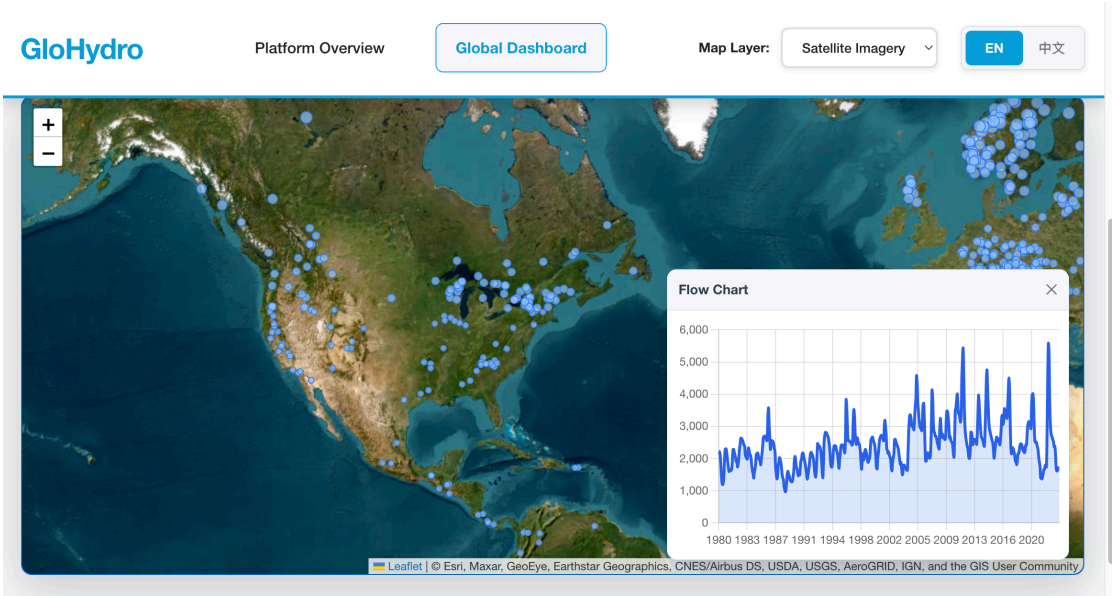
The HydroVLM analyzed high-resolution remote sensing imagery systematically retrieved along more than 8 million rivers worldwide, using pre-tiled 300 m × 300 m image patches. From these global river-aligned image collections, the model identified 14,545 candidate hydropower sites, of which 12,640 were validated as existing hydropower plants following expert review. This automated discovery pipeline marks a major advance over conventional manual mapping efforts that typically require months to years of intensive survey work, providing efficiency gains sufficient to support the global scale of our study. Furthermore, this framework establishes a scalable foundation for future extensions of VLM-based methodologies to identify and monitor other forms of energy infrastructure with comparable geographic breadth and precision.

7. Accessing the GloHydro Online System

Global-scale analyses of hydropower and the Water–Energy–Food (WEF) nexus, including policy and environmental assessments, constitute a vast and complex research domain. This study introduces a top-down remote sensing tracking approach for global hydropower plants and provides an inventory of these plants. We analyze the characteristics of hydropower plants in terms of their distribution, ecological impacts, and runoff patterns. Furthermore, we recognize that this inventory can play a crucial role in urban planning, sustainable development, and various other domains. To this end, we have developed an open-access platform for Glohydro, which can be accessed online at https://glohydro.cn. This platform displays a global map of hydropower plant distribution **(Supplementary Fig. 21)** and offers monthly runoff data for each plant from 1980 to 2023 **(Supplementary Fig. 22)**.



Supplementary Fig. 21 | GloHydro online system.



Supplementary Fig. 22 | Global Hydropower plants dashboard.

References

1. Alternative Energy Resources: The Way to a Sustainable Modern Society. vol. 99 (Springer International Publishing, Cham, 2021).
2. Phogat, R., Arora, D., Mehra, P. S., Sharma, J. & Chawla, D. A Comparative Study of Large Language Models: ChatGPT, DeepSeek, Claude and Qwen. in 2025 3rd International Conference on Device Intelligence, Computing and Communication Technologies (DICCT) 609–613 (IEEE, Dehradun, India, 2025). doi:10.1109/DICCT64131.2025.10986449.
3. Joshi, S. A Comprehensive Review of Qwen and DeepSeek LLMs: Architecture, Performance and Applications.
4. Li, Z. et al. A Survey of State of the Art Large Vision Language Models: Alignment, Benchmark, Evaluations and Challenges.
5. Ahmed, I. et al. Qwen 2.5: A Comprehensive Review of the Leading Resource-Efficient LLM with potentioal to Surpass All Competitors. Preprint at https://doi.org/10.36227/techrxiv.174060306.65738406/v1 (2025).
6. Bai, J. et al. Qwen Technical Report. Preprint at https://doi.org/10.48550/arXiv.2309.16609 (2023).
7. DeepSeek-AI et al. DeepSeek-V3 Technical Report. Preprint at https://doi.org/10.48550/arXiv.2412.19437 (2025).
8. DeepSeek-AI et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. Preprint at https://doi.org/10.48550/arXiv.2501.12948 (2025).
9. Puspitasari, F. D. et al. DeepSeek Models: A Comprehensive Survey of Methods and Applications. Preprint at https://doi.org/10.36227/techrxiv.174198511.15158242/v1 (2025).
10. Wu, W. et al. GPT4Vis: What Can GPT-4 Do for Zero-shot Visual Recognition? Preprint at https://doi.org/10.48550/arXiv.2311.15732 (2024).
11. Guo, D. et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature 645, 633–638 (2025).