Global Assessment of Reservoir Water Storage Capacity and Sedimentation Rate Using Remote Sensing and Satellite Time Series Data

Josh Manto | STATS402 Final Project Presentation

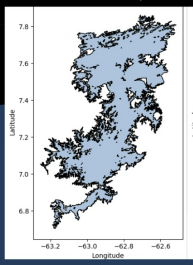
Google Earth Engine · PyTorch · TensorFlow · HydroSHEDS · USGS science for a changing world · Duke University · 昆山杜克大学 Duke Kunshan University

Abstract (1)

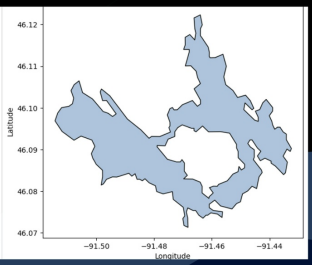
In this project, we address the critical need for scalable, global-scale monitoring of reservoir storage dynamics by integrating two premier geospatial inventories—HydroLAKES (HL) and the Global Dam Watch (GDW) database—to assemble precise reservoir geometries, morphometric attributes, and design capacities for 25,334 matched reservoirs. In this project, we developed a fully automated Python–Google Earth Engine (GEE) pipeline that ingests per-reservoir GeoJSONs, filters the COPERNICUS/S2_SR_HARMONIZED Sentinel-2 MSI collection (2018–2023, cloud cover ≤ 20%), computes spectral indices (e.g., NDWI) alongside all 13 original bands, and exports monthly composites as CSV time series. These dynamic features are merged with static hydromorphometric data such as surface area, storage capacity, mean depth, elevation, catchment area, and degree of regulation—extracted from a unified HL+GDW table of 35,646 records × 90 columns. To validate our area-to-volume proxy, we aligned NDWI-derived storage estimates for Fuqua Reservoir (GDW ID 3133) with USGS gauge measurements (2018–2025), observing RMSE ≈ 19.7 Mm^3, MAE ≈ 16.1 Mm^3, and MAPE ≈ 62.9% . For forecasting, we trained and compared LSTM and XGBoost regressors against a simple NDWI×capacity baseline using reservoir-wise 70/15/15 splits, quantifying performance via RMSE, MAE, MAPE, and R^2 and demonstrating that the NDWI-derived fill ratio remains the predominant predictor (> 60% explained variance), while additional Sentinel-2 bands (notably B8, B11) and static features yield consistent error reductions. This novel fusion of multi-temporal remote sensing, hydromorphic metadata, and advanced ML establishes a framework for near-real-time reservoir storage assessment. We look forward to integrating high-resolution bathymetry data and hydrometeorological inflows to forecast reservoir sedimentation buildup, and potentially packaging our workflow as an application for easy user towards resource management at a planetary scale.

Reservoir Querying Framework (2)

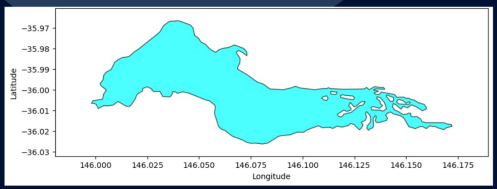
The first part of the pipeline involves building a database where we can easily query and extract the shapes of these reservoirs. We joined the Global Dam Watch (GDW) and HydroLAKES (HL) inventories via spatial intersection on pour-point centroids. HydroLAKES contributes precise reservoir geometries and morphometric metrics—surface area, mean depth, total volume, drainage area, elevation, and residence time—for 1.4 million lakes, while GDW supplies dam-specific attributes—design capacity (Cap_mcm), dam height/length, construction year, ownership, and quality flags—for 35 295 barriers. After co-registering 25 334 GDW entries to their matching HL polygons (71 % coverage), we produced a unified CSV of 35 646 records × 90 columns that merges both static and hydrologic metadata. This enriched table underpins our GEE queries (to extract 13-band monthly statistics) and feeds directly into downstream LSTM and XGBoost models. See below some reservoir outputs:



Guri Reservoir, located in Venezuela



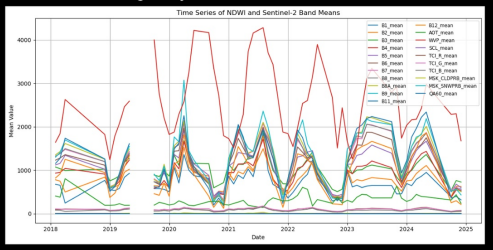
Nelson Reservoir, located in Sawyer County, Wisconsin



Mulwala Reservoir located in New South Wales, in eastern Australia

Communicating with Google Earth Engine (3)

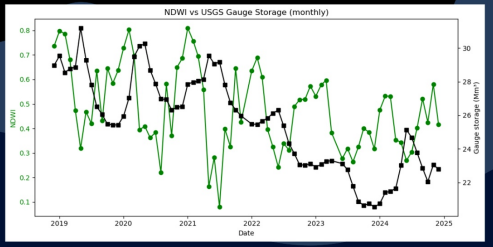
We harness Google Earth Engine (GEE) to automate monthly water-surface and spectral data extraction for each reservoir, turning thousands of shoreline polygons and their coordinates into ready-to-analyze tables. Using our prebuilt reservoir database, we ingested each reservoir's geometry into GEE as assets and then extracted spectral information from the Harmonized Sentinel-2 MSI collection spanning 2018–2023. We strictly filter for clear-sky images (≤20% cloud cover), employ the QA60 band to mask residual clouds, and sample all 13 spectral bands—ranging from visible blue, green, and red to near-infrared (NIR) and short-wave infrared (SWIR)—once per month. We also compute normalized indices such as the Normalized Difference Water Index (NDWI), defined as (Green − NIR) / (Green + NIR), which highlights surface water extent and its seasonal variations. This workflow produces an individual CSV time-series for each of the 35,646 reservoirs globally.



The image above is a randomly selected reservoir located in India. Among the spectral bands, NIR (around 800 nm) is highly sensitive to vegetation and water absorption, making it ideal for delineating water boundaries, while SWIR bands (around 1,600–2,200 nm) excel at detecting water quality indicators such as turbidity and moisture content. These monthly time series of raw reflectance values and indices become the predictor features for our machine-learning model. We basically use these bands to forecast reservoir storage volumes and sedimentation rates.

Prediction and Forecasting with LSTMs (4)

Our machine learning model is a two-layer long short-term memory (LSTM) network—a specialized form of recurrent neural network (RNN) with trainable input, forget, and output gates that explicitly learn which past signals to retain or discard. We train our model to translate multispectral time-series into accurate reservoir storage estimates per month. Our model input is a three-dimensional input tensor of shape (5 years × 12 months × 13 bands), where the 13 features include the full Harmonized Sentinel-2 MSI spectrum (Blue, Green, Red, four Red-Edge, NIR, two SWIR bands, plus QA60 cloud masks) and derived indices such as NDWI and the output would be the monthly water storage estimates. Our model consists of two stacked LSTM layers with 64 hidden units each (dropout = 0.2), followed by a 32-unit fully connected layer that outputs a scalar storage prediction.



The figure above shows our model's prediction on storage estimates of Fuqua Reservoir located in Stephen's County, Oklahoma. The green line is our prediction and the black line is groundtruth daily gauge readings taken from USGS (United States Geological Survey) sensors deployed in the reservoir. We trained for 100 epochs using an 80/20 train/test split. achieving a MAPE = 63%, overall pretty decent accuracy. Full training was held back due to time constraints, but we can bump this up further by refining our dataset further. Crucially, because tens of thousands of reservoirs worldwide lack in-situ gauges or even standardized names, our remote-sensing-driven LSTM approach offers a fully automated way to monitor water storage dynamics, detect emerging drought stress, and support water resource management at continental and global scales—all without any ground instrumentation.