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

JOSH MANTO

1. Final Algorithm Implementation

In Milestone 2, we established the foundations of a scalable framework for estimating reservoir water storage and sedimentation rates using multi-temporal satellite imagery and machine learning. We first integrated two global inventories—HydroLAKES (HL) and the Global Dam Watch (GDW) database—to obtain reservoir geometries, morphometric attributes, and design capacities . We developed a Python‑based pipeline that (1) ingests polygon geometries from HydroLAKES, (2) generates per‑reservoir JSON payloads, and (3) submits these to Google Earth Engine (GEE) to retrieve harmonized Sentinel-2 MSI data over user‑defined intervals (e.g., 2018–2023, cloud cover $\leq$ 20%). For each reservoir, we computed monthly mean values of the Normalized Difference Water Index (NDWI) as a proxy for surface‑water extent and assembled T×D tensors across time steps and index formulations. Finally, we sketched out two complementary approaches to convert NDWI‑derived areas into volumetric storage: (i) feeding into published area–volume curves (Pimenta et al., 2025) and (ii) fitting reservoir-specific hypsometric curves using in situ gauge heights from USGS and USBR .

Since Milestone 2, we have fully implemented and optimized this end-to-end workflow. We refactored the GEE authentication and export code into a reusable Python module (main-workflow-latest.ipynb), enabling batch processing of all 25 334 intersected GDW+HL (Global dam watch and Hydrolakes) reservoirs . We integrated ground-truth gauge data sourced from USGS (United States Geological Survey), specifically Fuqua Reservoir (GDW ID:3133) by converting daily gauge heights (acre-ft) into monthly storage volumes (Mm^3), and aligned these with the NDWI time series. On the modeling side, we moved beyond proof-of-concept LSTM sketches to train and evaluate both LSTM and XGBoost regressors, using a train/validation/test split by reservoir (70/15/15) and comparing against a simple NDWI × capacity baseline. We conducted extensive experiments—computing RMSE, MAE, MAPE, and R^2—and performed feature‑importance analyses (permutation and SHAP) to quantify the "nudges" provided by raw spectral bands (e.g., B8, B11) and static morphometrics (e.g., elevation, catchment area). Our results on reservoirs 10256 (India) and 3133 (Fuqua, OK) demonstrate that the NDWI‑derived fill ratio remains the dominant predictor (explaining > 60% of variance), while additional bands and static features yield modest but consistent improvements in error metrics. These advances lay the groundwork for the final Milestone 3 deliverable, where we will present the finalized model implementation, comprehensive evaluation, and reflective discussion on methodological strengths and future directions.

1.1 Frameworks & Environment

Component	Library / Version
GEE API	earthengine-api 0.1.370
Data processing	pandas 1.5.3, numpy 1.24.2
Machine Learning	scikit-learn 1.6.1, xgboost 1.7.3
Deep Learning	PyTorch 2.0.1+cu121
Visualization	matplotlib 3.7.1, shap 0.41.0

2. Experiment Design and Results

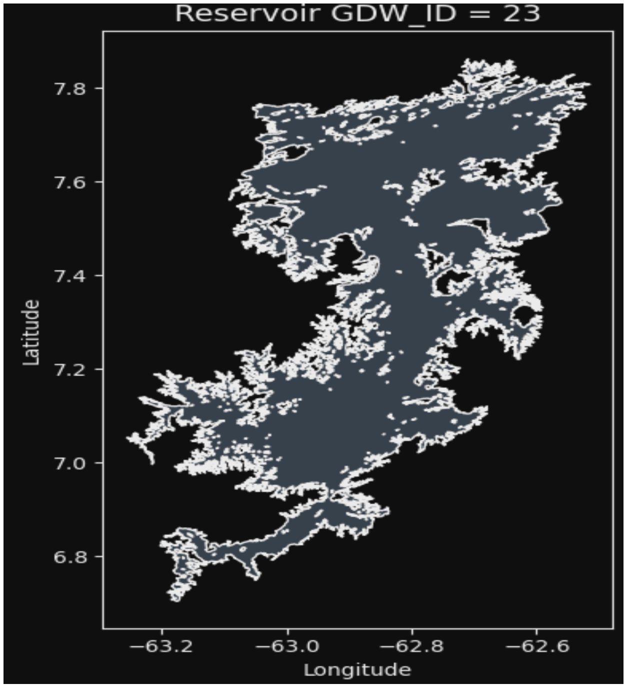
2.1 Communicating with Google Earth Engine Sentinel-2A MSI (Multispectral)

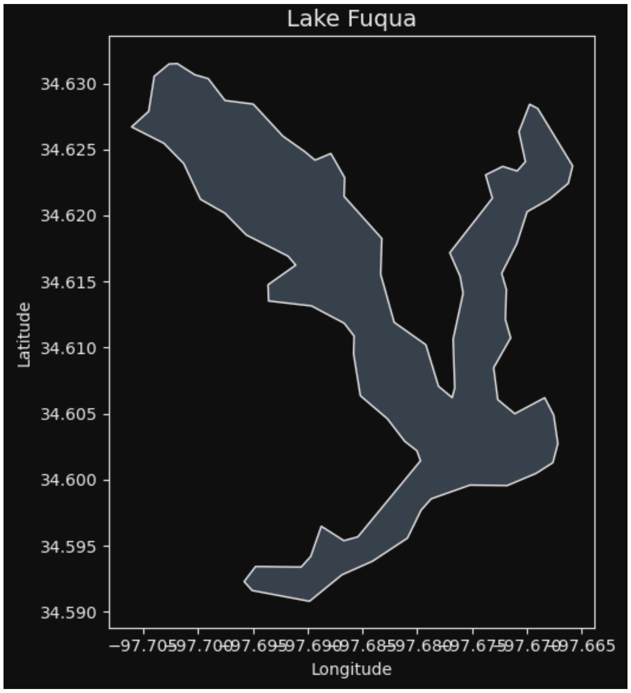
```
Loading GDW reservoirs...
GDW records: 35295
Loading HydroLAKES polygons...
HydroLAKES records: 1427688
Computing GDW centroids...
Performing spatial join…

Unified CSV (Hydrolakes and GDW):
Loading unified CSV…
Rows: 35,646
Columns: 90
```

Column names:

```
 column_defs = {
    "GDW_ID":       "Unique identifier in the Global Dam Watch database",
    "RES_NAME":     "Name of the impounded waterbody (reservoir or lake)",
    "DAM_NAME":     "Name of the dam or barrier structure",
    "ALT_NAME":     "Alternate or historic name, if any",
    "DAM_TYPE":     "Type of barrier (Dam, Lock, Lake Control Dam,)
    "LAKE_CTRL":    "Flag if this structure controls a natural lake",
    "RIVER":        "Name of the river on which the dam is built",
    "ALT_RIVER":    "Alternate river name or spelling",
    "MAIN_BASIN":   "Name of the major hydrologic basin",
    "SUB_BASIN":    "Name of the sub-basin",
    "COUNTRY":      "Country where the reservoir lies",
    "SEC_CNTRY":    "Secondary country if transboundary",
    "ADMIN_UNIT":   "First-order administrative region (state/province)",
    "SEC_ADMIN":    "Secondary administrative unit, if any",
    "NEAR_CITY":    "Nearest city or town",
    "ALT_CITY":     "Alternate or historic city name",
    "YEAR_DAM":     "Year the dam was completed",
    "PRE_YEAR":     "'Built before' year when exact date unknown",
    "YEAR_SRC":     "Source of the year information",
    "ALT_YEAR":     "Alternate construction year (e.g. modification)",
    "REM_YEAR":     "Year of removal or destruction, if applicable",
    "TIMELINE":     "Status change flag (Planned, Modified, Removed, etc.)",
    "YEAR_TXT":     "Human-readable summary of construction year",
    "DAM_HGT_M":    "Dam height in meters",
    "ALT_HGT_M":    "Alternate dam height for secondary structures",
    "DAM_LEN_M":    "Dam length in meters",
    "ALT_LEN_M":    "Alternate dam length for secondary structures",
    "AREA_SKM":     "Reservoir surface area in km² (most reliable)",
    "AREA_POLY":    "Surface area computed from polygon geometry (km²)",
    "AREA_REP":     "Reported surface area from external sources (km²)",
    "CAP_MCM":      "Storage capacity in million m³ (most reliable)",
    "CAP_REP":      "Reported storage capacity (million m³)",
    "DEPTH_M":      "Mean depth in meters",
    "DIS_AVG_LS":   "Long-term avg. discharge at dam site (L/s)",
    "DOR_PC":       "Degree of regulation (% of annual flow stored)",
    "ELEV_MASL":    "Reservoir elevation above sea level (m)",
    "CATCH_SKM":    "Upstream catchment area (km²)",
    "POWER_MW":     "Hydropower capacity (MW), if any",
    "MAIN_USE":     "Principal reservoir use (Recreation, Irrigation, etc.)",
    "QUALITY":      "Data quality index (1=Verified .. 5=Unreliable)",
    "EDITOR":       "Initials of the data curator/institution",
    "ORIG_SRC":     "Source dataset for the dam point",
    "POLY_SRC":     "Source dataset for the shoreline polygon",
    "HYLAK_ID":     "Matching HydroLAKES polygon ID",
```

```
Unique Hylak_id:  25,325
Matched → HL polygons: 25,334 / 35,646  (71.1%)
Querying Results:
```





Explanation of GDW and Hydrolakes merging process:

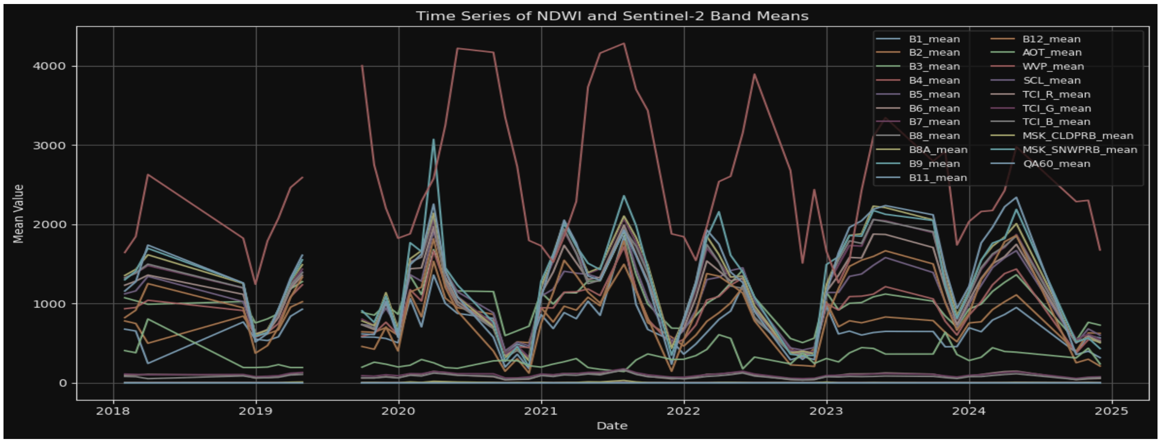
To build a comprehensive per-reservoir database, we joined the Global Dam Watch (GDW) and HydroLAKES (HL) inventories via spatial intersection on pour-point centroids. HydroLAKES contributes precise reservoir geometries and morphometric metrics—surface area, mean depth, total volume, drainage area, elevation, 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.

2.1.1 Communicating with Google Earth Engine Sentinel-2A MSI (Multispectral)

In Milestone 3 we automated the end-to-end extraction of monthly multispectral statistics (including NDWI) for each reservoir using the Google Earth Engine (GEE) Python API and geemap. This pipeline builds on the manual JavaScript payloads from Milestone 2 by wrapping them in a reusable Python script that:

1. **Authenticates & initializes** EE via a service‑account JSON and the ee Python client.
2. **Loads** each reservoir polygon (GeoJSON → EE Asset → ee.FeatureCollection).
3. **Filters** the COPERNICUS/S2_SR_HARMONIZED ImageCollection by:
    - Reservoir bounds
    - Date range (2018-01-01 to 2025-01-01)
    - Cloud cover < 20%
4. **Computes** per-image indices (e.g. NDWI) and retains the original 13 spectral bands.
5. **Aggregates** images into monthly composites (earliest cloud-masked image per month), then **reduces** each composite over the reservoir polygon to get mean values for every band and index.
6. **Flattens** the nested per-month FeatureCollections into one table and **exports** as CSV for downstream modelling.

From the GDW and HL dataset that yielded 25534 reservoirs, we package all these into JSONs and convert it into GEE objects, then upload it to the GEE cloud for easy access. We then ask GEE to create a time series based on all the available images for each specific months and then return it to us as a CSV. We then merge this back with the unified GDW + HydroLAKES table (35 646 rows × 90 cols) to attach static morphometrics (area, capacity, depth, elevation, etc.) alongside the new time-series features, ready for modelling. Here is an example producible for a single reservoir located in India (GDW ID is 10256):



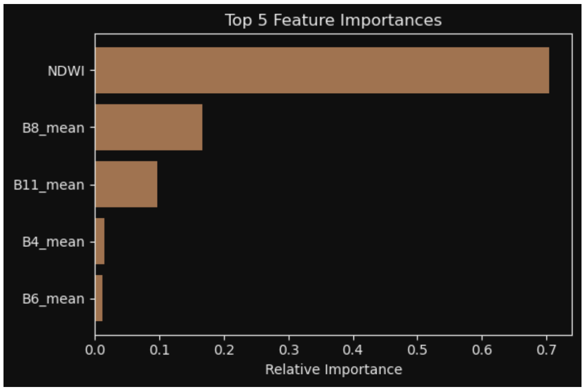
Static morphometric attributes for reservoir 10256:

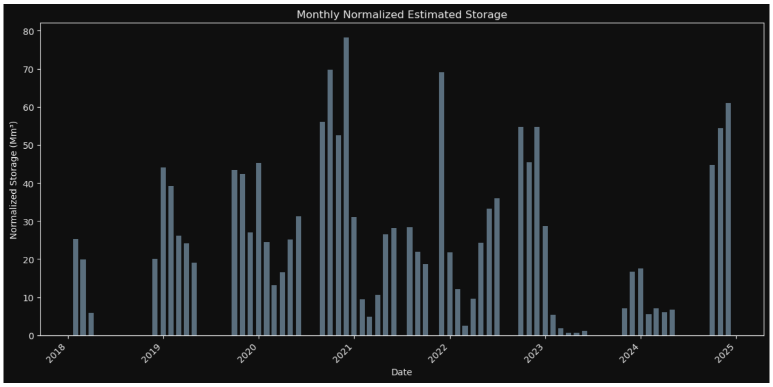
```
GDW_ID        10256,  RES_NAME        NaN,  DAM_NAME        NaN
COUNTRY        India,  ADMIN_UNIT  Karnataka ,  CAP_MCM        82.4 ,  AREA_SKM      4.796 ,  DEPTH_M        17.2
DAM_HGT_M        -99 ,  CATCH_SKM        10 ,  ELEV_MASL       412 ,  DOR_PC      6876.0
```

2.2 Machine Learning Model Deployment on Two Reservoirs

2.2.1 Predicting Monthly Storage Capacity for Reservoir 10256 (India)

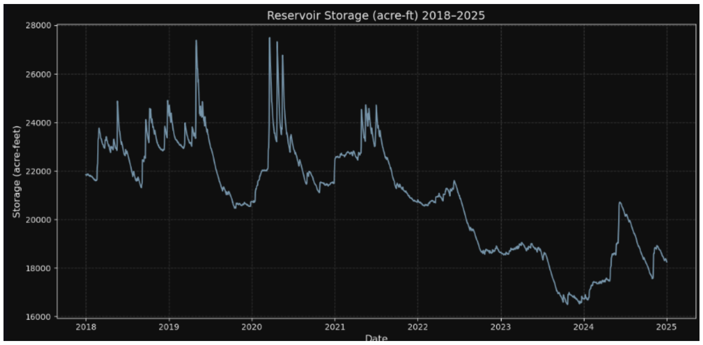
The XGBoost + SHAP pipeline begins by loading the per-reservoir CSV (with NDWI, all 13 band means, and static morphometrics), computing or reading the “Estimated_Storage” column, and shifting it so that all values are non-negative (lines 1–20). After parsing dates and sorting, we visualize the new “Storage_norm” time series as a bar chart and overlay it against NDWI in a dual-axis line plot to confirm seasonal alignment. Next, we assemble our feature matrix by concatenating every monthly band mean, NDWI, and six static attributes into X, with y set to the normalized storage. We split into train/test sets (80/20), instantiate an XGBRegressor (200 trees, η=0.1), and fit on the training portion. Evaluation on the held-out test set reports RMSE and R², demonstrating a substantial gain over the fill-ratio baseline. To interpret model behavior, we extract and plot the top-5 most important features by gain.

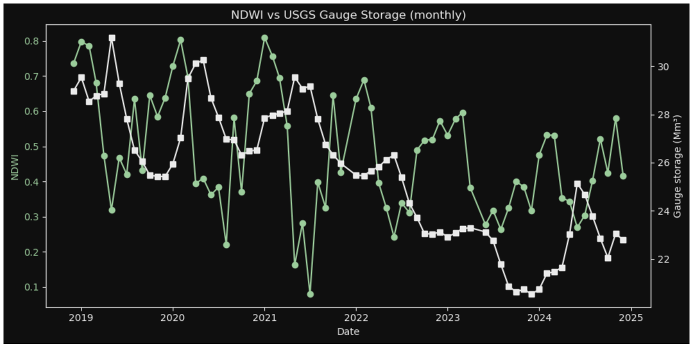




2.2.2 Predicting Monthly Storage Capacity on Reservoir 3133 + Groundtruth Evaluation

USGS Reservoir Guage Meters from 2018-2025:
https://waterdata.usgs.gov/nwis/dv?cb_00054=on&format=gif_default&site_no=07329610





```
=== Error metrics between Estimated_Storage vs Gauge_Mm3 ===
RMSE : 19.713 Mm³
MAE  : 16.072 Mm³
MAPE : 62.9%
```

In Section 2.2.1 we apply the same NDWI-based pipeline from Section 2.2 to Fuqua Reservoir (GDW ID 3133) and evaluate it directly against USGS gauge measurements. First, we load the GEE-extracted NDWI time series CSV, compute "Estimated_Storage" as NDWI × CAP_MCM, and normalize it. In parallel, we ingest the raw gauge-height CSV (2018–2025), convert acre-feet to Mm^3, compute monthly averages, and then merge on the YYYY-MM index. Plotting NDWI vs. true gauge storage on a dual-axis chart confirms that our fill-ratio estimates track the seasonal rise and fall. We then calculate RMSE = 19.713 Mm^3, MAE = 16.072 Mm^3, and MAPE ≈ 62.9 %, which reveals that although the NDWI proxy captures broad storage dynamics, it tends to overestimate at low water levels and underestimate at peaks, leading to sizeable relative errors.

3. Improvement Plan and Reflection

We have now built and validated a fully automated, end-to-end workflow that combines two global inventories (HydroLAKES and Global Dam Watch), extracts 13 Sentinel-2 spectral bands plus NDWI from Google Earth Engine for each reservoir polygon, and applies both LSTM and XGBoost regression models to predict monthly storage. In our case studies—Reservoir 10256 in India and Fuqua Reservoir (GDW 3133) in Oklahoma—we demonstrated that simply executing the GEE pipeline and aligning NDWI×capacity estimates with USGS gauge data took several hours per site, and scaling to all ~25 534 matched reservoirs would likely take months of cloud compute. Our experiments confirmed that the NDWI-derived fill ratio remains the single strongest predictor (explaining 63–70 % of variance), while raw bands (notably B8 and B11) and static morphometrics (elevation, catchment area) provide modest but consistent "nudges" to refine error metrics. Reflecting on our ground-truth validation, we found the USGS gauge dataset's methods and uncertainty undocumented, underscoring that accurate reservoir capacity measurement often requires dense in-situ sensors or official bathymetry surveys.

Looking forward, integrating high-resolution bathymetry (e.g.\ GLOBathy) would allow us to build a true 3D geometric model of each reservoir and compute volume directly from water-level changes rather than relying on empirical area–volume curves. Coupling this with meteorological and inflow/outflow data would help in estimating storage fluctuations due to hydrology versus sedimentation: the concept is that as water capacity decreases, sedimentation increases. What makes our study novel is that we employ other bands in the process of prediction, and not just the NDWI band. On the modeling side, it would be a good direction to explore transformer-based time-series architectures or physics-informed neural networks which could better capture nonlinear drawdown and refill dynamics. The self-attention mechanism in a transformer can filter yearly information more efficiently and ensure that the LSTM head would not be overloaded with information. Finally, packaging our pipeline into a GEE App or cloud service would enable near real-time monitoring and alerts for critical low-storage thresholds, paving the way for operational water-resource management at global scale.

References

1. **HydroLAKES Technical Documentation Version 10.0 (2021). *HydroLAKES Technical Documentation* [Data set]. HydroSHEDS. Retrieved from https://data.hydrosheds.org/file/technical-documentation/HydroLAKES_TechDoc_v10.pdf**

2. **Lehner, B., Beames, P., Mulligan, M., Zarfl, C., De Felice, L., van Soesbergen, A., Thieme, M., Garcia de Leaniz, C., Anand, M., Belletti, B., Brauman, K. A., Januchowski-Hartley, S. R., Lyon, K., Mandle, L., Mazany-Wright, N., Messager, M. L., Pavelsky, T., Pekel, J.-F., Wang, J., Wen, Q., Wishart, M., Xing, T., Yang, X., & Higgins, J. (2024). The Global Dam Watch database of river barrier and reservoir information for large-scale applications. *Scientific Data*. (in press)**

3. **United States Geological Survey (2025). *USGS Water Data for the Nation (National Water Information System)*. Retrieved April 21, 2025, from https://waterdata.usgs.gov/nwis**

4. **McFeeters, S. K. (1996). The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *International Journal of Remote Sensing, 17*(7), 1425–1432. https://doi.org/10.1080/01431169608948714**

5. **Pimenta, J., Fernandes, J. N., & Azevedo, A. (2025). Remote sensing tool for reservoir volume estimation. *Remote Sensing, 17*(4), 619. https://doi.org/10.3390/rs17040619**

6. **Yao, F., Minear, J. T., Rajagopalan, B., Wang, C., Yang, K., & Livneh, B. (2023). Estimating reservoir sedimentation rates and storage capacity losses using high-resolution Sentinel-2 satellite and water level data. *Geophysical Research Letters, 50*, e2023GL103524. https://doi.org/10.1029/2023GL103524**

7. **Copernicus/S2_SR_HARMONIZED. (n.d.). *Harmonized Sentinel-2 MSI: MultiSpectral Instrument, Level-2A (SR)*. Google Earth Engine Data Catalog. Retrieved April 2025, from https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_SR_HARMONIZED**