

# Research Notes

## Geospatial data systems for environmental risk and decision support

**Jose Escobar** · Independent geospatial data engineer · Mexico  
jaemaciel.55@gmail.com · github.com/Goheist01 · linkedin.com/in/jose-escobar-cid3000 · goheist01.github.io/jose-escobar-cv  
Spanish (native) · English (fluent, C1)  
Version 1.0 · September 2026

### 1 Purpose and scope

This document is written for research supervisors, principal investigators and technical leads evaluating whether my work is relevant to their group. It is not a CV. It states, for each flagship system, the question it was built to answer, the data behind it, the method, how the result was checked, what the result does and does not support, and what I would test next.

All four systems are self-directed. None has been operated in production with real users. Each was built to be verifiable rather than impressive: source datasets are public and cited, code is open, and outputs are reproducible from the documented commands. Where a component is modelled rather than measured, it is labelled as such – including in the one case where that labelling materially reduces what the system can claim.

I hold a Bachelor of Industrial Engineering (TecNM – Instituto Tecnológico Superior de Poza Rica, 2022) and a Diploma of Information Technology (Laneway Education, Melbourne, 2026). I do not have a formal background in remote sensing, hydrology or computer science. The work below is how I have been building that foundation in public, and the limitations sections are the part I would most like to discuss.

### 2 Data provenance

Research credibility depends more on knowing which numbers are measurements than on how many numbers there are. The table below separates the two across all four systems.

| System                                     | Measured / real public data                                                                       | Modelled or synthetic                                                                                                                                 |
|--------------------------------------------|---------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------|
| Structural Triage Engine                   | Airborne LiDAR point clouds (OpenTopography); OSM building footprints via Overpass                | None. All indicators computed from returns.                                                                                                           |
| Hydrological Terrain Simulation Engine     | USGS 3DEP DEM; NHD flowlines; SSURGO soils; NOAA Atlas 14 precipitation frequency; OSM footprints | Design storm is a synthetic hyetograph constructed from real Atlas 14 depth-duration-frequency statistics – standard practice, not an observed event. |
| Amazon Deforestation Intelligence Platform | Hansen GFC v1.11 loss pixels; MODIS MCD12Q1 post-loss land cover; WDPA v2024 boundaries           | None. Driver attribution is a published proxy method applied to real land-cover data, not a label I invented.                                         |
| Mekong Delta Agricultural Intelligence     | Five Sentinel-2 SR Harmonized composites (Nov 2022 – Apr 2023, SCL cloud masking)                 | <b>The parcel layer and the per-parcel spectral series</b>                                                                                            |

| System   | Measured / real public data | Modelled or synthetic                                 |
|----------|-----------------------------|-------------------------------------------------------|
| Platform |                             | <b>are both synthetic.</b> See the note below and §6. |

#### MATERIAL DISCLOSURE — MEKONG DELTA PLATFORM

No public cadastral registry exists for the Mekong Delta at individual-parcel level. The 499 parcel polygons are a synthetic grid placed on real province coordinates, and the 4,981 spectral observations attached to them are derived from published spectral profile ranges rather than sampled per polygon from the imagery. The satellite imagery layer is real. The parcel-level analytics layer is a simulation. That system therefore demonstrates an interface and a data architecture; it does not measure conditions in any actual field, and no inference about real crop or flood status in the delta should be drawn from it. §6 states what would have to change for that to no longer be true.

### 3 Structural Triage Engine

Post-disaster inspection prioritisation from airborne LiDAR · Haiti 2010, Kumamoto 2016

[github.com/Goheist01/structural-triage-engine](https://github.com/Goheist01/structural-triage-engine) · MIT-licensed, research-level README

|                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
|--------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Question           | Can deterministic geometric analysis of a post-event airborne LiDAR point cloud produce a defensible building-level inspection ordering – with no machine learning, no GPU, no cloud inference and no internet connectivity at runtime?                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| Why it matters     | In the first days after an earthquake, inspection capacity is the binding constraint, and connectivity is typically the first infrastructure to fail. Post-event LiDAR is often flown within days. If a usable ordering can be extracted from that scan on a laptop in the field, the ordering becomes available at the moment it is worth the most.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
| Data and selection | <p>Two independent post-disaster surveys were chosen specifically because they disagree on almost every axis except magnitude:</p> <ul style="list-style-type: none"> <li>• <b>Haiti 2010</b> – World Bank / ImageCat / RIT survey via OpenTopography, flown 21–27 January 2010 (9–15 days post-event). EPSG:32618. 2 × 2 km AOI over Carrefour-Feuilles, Port-au-Prince. 17,829,030 filtered returns. Informal reinforced concrete block, high collapse rate.</li> <li>• <b>Kumamoto 2016</b> – Asia Air Survey Co. via OpenTopography, flown 23 April 2016 (7 days post-event). EPSG:2444 (Japan Plane Rectangular CS II). ~1.5 × 1.5 km over Mashiki town centre. 8,057,332 returns. Japanese timber frame, documented damage along the Futagawa fault zone.</li> </ul> <p>Building footprints from OpenStreetMap via the Overpass API, filtered to 20–5,000 m<sup>2</sup> to remove mapping artefacts. This is the only step requiring connectivity; everything after it runs offline.</p> |
| Method             | Returns are separated by classification – above-ground (class 1) scored, ground (class 2) retained for context, noise (class 7) discarded. A <i>scipy.cKDTree</i> over the horizontal coordinates of above-ground returns makes per-building radius queries tractable across 17M+ points; index build is under five seconds for a neighbourhood-scale AOI. Footprints with fewer than 50 returns are flagged No Data rather than scored.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |

Four physically-motivated geometric indicators are computed per footprint and combined with fixed weights:

| Indicator                     | w    | Computation and physical reasoning                                                                                                                                                                                                                              |
|-------------------------------|------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Vertical distribution entropy | 0.40 | Shannon entropy of returns binned into 10 equal-width height bands, normalised by the uniform-distribution maximum. An intact structure concentrates returns at ground and roof planes with a void between; debris fills that void and smears the distribution. |
| Point density irregularity    | 0.30 | Coefficient of variation of return counts across a $4 \times 4$ grid over the footprint bounding box. Intact roofs return near-uniform density; debris scatter is spatially irregular.                                                                          |
| Roof planarity RMSE           | 0.20 | Least-squares plane fitted to the top 20th percentile of Z returns; RMSE of residuals scaled $\times 20$ and capped at 100. Deformed roof geometry scatters away from any fitted plane.                                                                         |
| Point count deficit           | 0.10 | Shortfall against expected returns at the dataset mean density ( $8.0 \text{ returns/m}^2$ ) times footprint area. Rubble absorbs and scatters pulses differently from a specular roof surface.                                                                 |

Tiers are assigned by percentile within the scan – Critical (top 25%), High (50–75th), Moderate (25–50th), Low (bottom 25%) – deliberately rather than by absolute cutoff. In a heavily damaged neighbourhood, absolute thresholds compress every score into a narrow band and lose discrimination; the percentile formulation answers the operationally correct question of which buildings are most anomalous relative to their neighbours in this scan. Field routing uses greedy nearest-neighbour ordering within each tier, Critical first.

|                     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|---------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Validation approach | A cross-context transfer test rather than an accuracy test. The same codebase, unchanged, was run against both datasets. What this establishes is that the method is not tuned to one coordinate system, point density, construction typology or seismic context. What it explicitly does not establish is whether the ordering is correct.                                                                                                                                                                                                                                         |
| Results             | Haiti: 223 of 250 footprints scored. Kumamoto: 222 of 3,595 scored. The large No Data count in Kumamoto is a property of the source survey, not a failure of the engine – that dataset was flown as a fault-rupture scan rather than an urban building survey, so return density over structures is substantially lower than in Haiti. Runtime is under 60 seconds for a neighbourhood-scale scan on a laptop, producing a self-contained HTML priority map with all data embedded as JSON, plus a GPX route carrying both route and waypoint elements for OsmAnd, Garmin and QGIS. |
| Limitations         | <ul style="list-style-type: none"> <li>• <b>The central one:</b> the ordering has never been scored against engineer-assigned damage grades. Until it is checked against ATC-20 or EMS-98 style field tags, the indicators remain physically-motivated heuristics, not validated predictors.</li> </ul>                                                                                                                                                                                                                                                                             |

- **No pre-event baseline.** Every indicator measures post-event geometry alone. A building that was always geometrically irregular – informal construction, rooftop additions – can score elevated without being damaged.
- **Single-pass occlusion.** Structures shadowed by taller neighbours receive few or no returns and are flagged No Data rather than scored, so the sample is not random with respect to urban form.
- **No material classification.** Concrete, timber and masonry fail differently and leave different LiDAR signatures; the engine treats all typologies identically with one fixed weight vector.
- **Temporal gap.** Surveys are flown days to weeks post-event. Debris clearance and aftershock damage in that window both contaminate the geometric record.
- **Relative thresholds.** A Critical building in a lightly damaged area may carry only minor irregularity. The tier label is not portable between scans.

---

|                        |                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| What I would test next | In order of value: score the ranking against field-verified damage tags for a subset of Haiti or Kumamoto buildings and report precision at the top decile; fit per-typology weight vectors rather than one global set; add pre/post differencing where a matching pre-event survey exists; and attach an occlusion-aware confidence value so No Data and low-return buildings are distinguishable from genuinely low-anomaly ones. |
|------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

## 4 Hydrological Terrain Simulation Engine

DEM-to-flood-screening pipeline · Houston Ship Channel, Buffalo National River

[github.com/Goheist01/hydro-sim-engine](https://github.com/Goheist01/hydro-sim-engine) · MIT · LIMITATIONS.md documents every bug found during development

---

|                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
|--------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Question           | Can an automated open-source pipeline turn an arbitrary US digital elevation model into a defensible first-pass flood screening product – and can it be made bit-for-bit reproducible?                                                                                                                                                                                                                                                                                                                  |
| Why it matters     | A permitted hydraulic study costs weeks and thousands of dollars, so the decision about which sites justify one is usually made without hydrology. A reproducible screening layer changes what that decision is made on. The reproducibility half turned out to matter more than expected – see the note below.                                                                                                                                                                                         |
| Data and selection | USGS 3DEP elevation via OpenTopography; NHD flowlines for channel geometry; SSURGO soil survey for infiltration parameters; NOAA Atlas 14 for design-storm depth-duration-frequency statistics; OSM building footprints for exposure. All fetched automatically for the DEM footprint. The two test regions were chosen to be physiographic opposites: a flat coastal corridor (Galena Park / Pasadena / Deer Park, Texas) and dissected mountainous terrain (Buffalo National River, Arkansas Ozarks). |

|                     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
|---------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Method              | <p>Stream burning conditions the surface against known NHD channels, which on the demo DEM reduced unsolved depressions from 21,129 to 1,423. Depression filling and flat resolution then use deterministic priority-flood in pysheds; D8 pointer and accumulation use WhiteboxTools. Sub-basins are delineated around a stream network defined by an absolute upstream drainage-area threshold, default 1 km<sup>2</sup>.</p> <p>Per-basin rainfall-runoff uses Green-Ampt infiltration with SSURGO-derived soil parameters after Rawls, Brakensiek &amp; Miller (1983), cross-validated against the SCS Curve Number method. Routing is Muskingum-Cunge kinematic wave through compound channels whose floodplain geometry is measured from the DEM itself using the divided-channel Manning method. Flood depth per cell comes from HAND – height above nearest drainage – after Rennó et al. (2008) and Nobre et al. (2011), including per-building peak-depth exposure. Every run writes a provenance JSON recording sources, parameters and versions.</p>                                                                                                                                                                                                                                                                                              |
| Validation approach | <p>Five independent checks, none of which is a gauge calibration:</p> <ul style="list-style-type: none"> <li>• <b>Reproducibility.</b> A regression gate runs the pipeline twice and fails if the basin summaries differ by a single byte.</li> <li>• <b>Method convergence.</b> Green-Ampt (physics-based, SSURGO parameters) and SCS Curve Number (empirical) agree within roughly 5% on the same basins.</li> <li>• <b>Input corroboration.</b> The Atlas 14 100-year / 24-hour depth used for Houston, 17.7 in, independently matches Harris County Flood Control District's own post-Harvey design assumption.</li> <li>• <b>Physical plausibility of routing.</b> Replacing simplistic channel geometry with DEM-derived compound cross-sections reduced that configuration's terminal peak from ~209,000 to ~131,000 cfs – attenuation produced by measured floodplain storage rather than by tuning a parameter.</li> <li>• <b>Generalisation.</b> The Ozarks run used the same command with zero code changes. The engine fetched region-correct inputs automatically (8.4 in design storm against Houston's 17.7 in) and resolved the opposite drainage structure: 64 basins organised as one nested watershed draining 33.5 of 48.5 sq mi through a single terminal outlet, against the coastal corridor's 163 independent catchments.</li> </ul> |
| Results             | <p>Canonical demo run: 163 sub-basins at the 1 km<sup>2</sup> threshold, largest routed peak discharge 33,266 cfs from the corridor's largest independent catchment of roughly 14 sq mi. No basin aggregates the whole domain, correctly – the coastal corridor drains through many independent outlets to the Ship Channel and bay rather than as a single nested watershed. Outputs are an animated HTML simulation, a written PDF drainage report, a QGIS-ready GeoPackage of sub-basins and drainage network, per-basin summary and hydrograph CSVs, a peak flood depth raster, and the provenance record.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |

## METHODOLOGICAL NOTE — NONDETERMINISM IN TERRAIN CONDITIONING

Repeated runs on an identical DEM were producing different basin counts. Hashing every pipeline intermediate across runs localised the cause to terrain conditioning: WhiteboxTools' least-cost breaching and flat resolution break ties nondeterministically, through thread scheduling and Rust's per-process hash randomisation. The resulting sub-millimetre elevation noise is enough to flip D8 flow directions on flat ground and shift drainage divides between otherwise identical runs.

The fix moved conditioning to pysheds' deterministic priority-flood filling, retaining WhiteboxTools for D8 pointer and accumulation, which are deterministic given identical input. The same investigation exposed a second, independent design flaw: the stream threshold had been defined as a fraction of the DEM's own maximum flow accumulation, making basin count hostage to whatever the largest stream in the extent happened to carry. It was redefined as an absolute drainage area in km<sup>2</sup> – physical, and stable across both conditioning methods and DEM extents.

This is worth flagging beyond this project. Any published basin statistic derived from WhiteboxTools breaching-based conditioning on low-relief terrain may not be reproducible from the same DEM, and the discrepancy is small enough to pass unnoticed.

### Limitations

- **Screening, not engineering.** This is not a substitute for a permitted hydraulic study and should never be represented as one.
- **Kinematic wave routing cannot represent backwater effects**, which is a material omission in exactly the flat coastal settings where the tool is most tempting to use.
- **HAND flood mapping is a geometric approximation**, not 2D hydraulics.
- **Infrastructure is unmodelled** – reservoirs, levees, pumps and storm sewers do not exist in the simulation.
- **The design storm is spatially uniform** across the domain.
- **Threshold sensitivity is real.** The demo AOI yields 163 basins at a 1 km<sup>2</sup> threshold and 60 at 2 km<sup>2</sup>. Basin count is a parameter choice in low-relief terrain, not a property of the landscape, and should be reported as such.
- **Nothing is calibrated against stream gauge records**, and the national dataset integrations are US-only.

### What I would test next

Calibrate routed hydrographs against USGS gauge records for a basin where a real storm and a gauge record coincide – that is the check that converts plausibility into accuracy. Then: report basin statistics as an ensemble across a threshold sweep rather than at a single default; add 2D hydraulics where backwater dominates; and build data adapters so non-US regions are reachable, since the hydrology is general and only the data plumbing is American.

## 5 Amazon Deforestation Intelligence Platform

Forest loss against protected-area boundaries · Pará, Amazonas, Mato Grosso, Rondônia · 2001–2023

[goheist01.github.io/uav-mission-intelligence/outputs/amazon\\_platform.html](https://goheist01.github.io/uav-mission-intelligence/outputs/amazon_platform.html) · live, database-backed

### Question

Which satellite-detected forest-loss pixels fall within registered protected-area boundaries, and can that intersection be served as a queryable system where every result traces back to its original public source?

|                     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
|---------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Why it matters      | Forest-loss products and protected-area registries are maintained by different institutions and are rarely joined. An organisation that wants to act on a loss signal needs to know whether it sits inside a boundary before anything else, and needs the provenance to accompany the claim.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| Data and selection  | 1,121 forest-loss pixels from Hansen Global Forest Change v1.11 (Hansen et al. 2013), exported through Google Earth Engine across four Brazilian states for 2001–2023. Loss drivers derived from MODIS MCD12Q1 post-loss land cover using the published proxy methodology employed by INPE and Global Forest Watch – forest degradation 569, cattle ranching 541, fire 5, agriculture 4, mining 2. Protected-area geometry from WDPA v2024: 370 boundaries covering indigenous territories, national parks and extractive reserves.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| Method              | Loss pixels and boundaries are loaded into PostgreSQL 17 with PostGIS 3.6 and GIST spatial indexes, and intersected with ST_Within. The schema was migrated to Supabase for live hosting; the Leaflet frontend queries the Supabase REST API directly, with year, driver and overlap filters and a click-to-detail sidebar, so the map is a view onto a live database rather than a static export.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| Validation approach | Provenance rather than accuracy. Every displayed point resolves to its source dataset, loss year, canopy cover and driver classification, so any individual claim on the map can be checked against the original public product. The spatial join itself is exact given the inputs; what is uncertain is the inputs.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| Result              | 66 of the 1,121 loss detections fall inside a registered protected-area polygon under ST_Within – <b>potential overlaps</b> , and deliberately not called violations. See the limitation immediately below, which is the reason for the wording.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| Limitations         | <ul style="list-style-type: none"> <li>• <b>Hansen detects canopy loss, not legality.</b> Loss inside a protected boundary can be legal – permitted extraction in a sustainable-use reserve, natural disturbance, or a boundary whose designation postdates the loss year. An overlap is a lead for investigation and nothing stronger. Any framing of these 66 as confirmed violations would be an overstatement, and earlier versions of my own documentation made exactly that error.</li> <li>• <b>Scale mismatch in the driver attribution.</b> Hansen loss is 30 m; the MODIS land-cover proxy that labels its driver is 500 m. A driver label is therefore a neighbourhood-scale inference attached to a pixel-scale detection.</li> <li>• <b>Boundary uncertainty.</b> WDPA polygons carry their own positional error and currency lag, and no buffer sensitivity analysis has been run on the overlap count – an edge detection could flip in or out under a plausible boundary displacement.</li> <li>• <b>Annual temporal resolution.</b> Hansen resolves loss to a year, which is far too coarse for any enforcement response.</li> <li>• <b>No ground truth.</b> None of the 66 has been checked against an independent record.</li> </ul> |

|                        |                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
|------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| What I would test next | Run a boundary-buffer sensitivity analysis and report the overlap count as a range rather than a point estimate – that is the honest presentation of this number and I would do it before the platform were used for anything. Then move ingestion from annual Hansen to near-real-time GLAD or RADD alerts, which is what makes the system operationally rather than retrospectively useful, and add scheduled updates with an audit trail on exports. |
|------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

## 6 Mekong Delta Agricultural Intelligence Platform

Parcel-level spectral monitoring interface · 13 Vietnamese provinces, Kandal and Prey Veng, Cambodia

[goheist01.github.io/uav-mission-intelligence/outputs/mekong\\_platform.html](https://goheist01.github.io/uav-mission-intelligence/outputs/mekong_platform.html) · live

### READ THIS BEFORE THE REST OF THE SECTION

The parcel boundaries and the per-parcel spectral observations in this system are synthetic. It is an interface and data-architecture demonstration built over real imagery, not a measurement system. The section is written on that basis.

|                            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|----------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Question                   | Can a parcel-level spectral monitoring interface be built and served for a region that has no public parcel cadastre – and what would it take to make the analytics layer real?                                                                                                                                                                                                                                                                                                           |
| Why it matters             | An agricultural insurer assessing flood crop loss cannot inspect hundreds of parcels, and in the Mekong Delta a single image cannot separate loss from the flooding that the three-crop calendar produces by design. The unit of decision is the parcel across time. The obstacle is that the parcel layer, in this region, does not publicly exist.                                                                                                                                      |
| Data — real component      | Five Sentinel-2 SR Harmonized composites exported from Google Earth Engine for November 2022 – April 2023 with SCL cloud masking and a fallback composite: true colour, NDVI, NDWI, NDRE and false-colour NIR. This layer is genuine satellite observation.                                                                                                                                                                                                                               |
| Data — synthetic component | 499 parcel polygons generated as a synthetic grid on real province coordinates, and 4,981 spectral observations derived from published spectral profile ranges rather than sampled from the imagery per polygon. No public cadastral registry exists for the delta at individual-parcel level, so the parcel layer was constructed to exercise the architecture. Both are labelled as synthetic in the repository documentation.                                                          |
| Method                     | Composites are converted to Cloud Optimized GeoTIFF with GDAL and then to 23,245 XYZ tiles at zoom levels 7–13 via gdal2tiles, served from GitHub Pages so band switching is instant with no download wait; the COGs are held on AWS S3 (us-east-1). A PostGIS schema on Supabase holds the parcel geometry and the spectral series. The Leaflet and Chart.js frontend renders a temporal NDVI and NDWI chart on any parcel click, with land-use and country filters and a flood overlay. |

|                        |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
|------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Validation approach    | Internal consistency only. The generated series was checked against the delta's published three-crop calendar to confirm the simulation is not producing a phenologically impossible signal. This is a check on the simulation, not evidence about the delta.                                                                                                                                                                                                                                                                                                                                                                                                                            |
| Result                 | The synthetic NDWI series reproduces the expected paddy flood cycle – December 0.464, May 0.324, September 0.185 – consistent with the published three-crop calendar. The engineering result is the one that stands on its own: a tiled, database-backed, parcel-queryable interface over a large delta serving instantly from static hosting.                                                                                                                                                                                                                                                                                                                                           |
| Limitations            | <ul style="list-style-type: none"> <li>• <b>The analytics layer is modelled.</b> Nothing in the per-parcel charts is an observation of any real field, and no conclusion about actual crop or flood conditions in the Mekong Delta is supported by this system.</li> <li>• <b>The parcel geometry is not a cadastre</b> and does not correspond to real landholdings.</li> <li>• <b>The imagery is optical-only</b>, so monsoon cloud gaps would be unavoidable even with real sampling.</li> <li>• <b>No ground truth of any kind</b> – no yield records, no verified claims, no field survey.</li> </ul>                                                                               |
| What I would test next | One change dominates everything else: compute zonal statistics per polygon directly from the real composites and replace the synthetic series with measured values. That single step converts the system from an interface demonstration into a measurement system, and it is a small amount of work relative to what is already built – the tiling, schema and frontend all remain. After that: substitute a real parcel layer wherever one can be licensed or derived by segmentation, add Sentinel-1 SAR for cloud-independent water detection, and validate flood detection against an actual claims or yield dataset before the system is described as decision-support for anyone. |

## 7 Engineering practice

Three habits run through the work, and they are the part I would most want a group to interrogate.

### Determinism treated as a testable property

The hydrological engine ships a regression gate that runs the full pipeline twice and fails the build if the basin summaries differ by one byte. That gate exists because the pipeline was silently nondeterministic for a period and the discrepancy was small enough to look like nothing. Reproducibility that is asserted rather than tested is not reproducibility.

### Verification over assumption

A separate project – a real-time Earth-observation satellite visualiser using client-side SGP4 propagation – produced a set of errors worth recording because each was invisible to a spot check and only surfaced under a designed test:

- Raw SGP4 output is in an inertial frame that does not rotate with the Earth. Converting it directly to longitude produced positions that looked correct in any single check. The error only appeared when the same satellite was propagated across a six-hour gap and showed none of the roughly 90° of apparent longitude drift that Earth's rotation requires.
- Sentinel-2 STAC metadata tracks cloud cover and no-data percentage as independent dimensions. A real scene returned 0.4% cloud cover and 77% no-data simultaneously – an edge-of-swath tile that passes a naive cloud filter while being almost entirely empty.
- Landsat's thumbnail asset is flagged requester-pays and cannot be fetched from a browser, which is only discoverable by reading the asset metadata rather than the documentation.

None of these is profound. Collectively they are the argument that I check rather than assume, which is the habit that transfers to research work.

### Explicit labelling of modelled data

The Mekong disclosure in §2 and §6 costs that project most of its analytical claim. Stating it is not a formality – it is the difference between a portfolio that survives scrutiny and one that does not, and I would rather lose the claim than have someone else find it.

## 8 Background and interests

---

My formal background is industrial engineering. I moved into software and data engineering by building the evidence in public: working systems, reproducible pipelines, documented validation and explicit limitations. Alongside the four systems above there are nine further documented sprints covering UAV failsafe validation in PX4 SITL, photogrammetric orthomosaic and NDVI production, Sentinel-2 multispectral analysis, thermal anomaly detection and severity scoring on a 20,000-image benchmark, LiDAR DTM/DSM differencing, Random Forest land-cover classification on Earth Engine, and YOLOv8 transfer learning for solar panel fault detection. All are published with their outputs and their failure modes.

I am looking for a research group, postgraduate programme or specialist team where I can contribute as a technical builder while building a proper research foundation. The areas that interest me most are the ones where these projects keep running into their own limits: validation of remote-sensing-derived indicators against field records, reproducibility in DEM-derived hydrology, uncertainty propagation through spatial joins on administrative boundaries, and the general problem of stating honestly what an Earth-observation product does and does not license someone to conclude.

I read and write English fluently, and would relocate.

## 9 Data sources and references

---

Hansen, M. C. et al. (2013). High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science* 342(6160), 850–853. Global Forest Change v1.11.

Rawls, W. J., Brakensiek, D. L. & Miller, N. (1983). Green-Ampt Infiltration Parameters from Soils Data. *Journal of Hydraulic Engineering* 109(1), 62–70.

Chow, V. T., Maidment, D. R. & Mays, L. W. (1988). *Applied Hydrology*. McGraw-Hill. – Muskingum-Cunge routing.

Rennó, C. D. et al. (2008); Nobre, A. D. et al. (2011). HAND – Height Above Nearest Drainage terrain descriptor.

---

UNEP-WCMC & IUCN. World Database on Protected Areas (WDPA), v2024.

---

Friedl, M. & Sulla-Menashe, D. MODIS MCD12Q1 Land Cover Type. Driver attribution follows the published INPE / Global Forest Watch post-loss land-cover proxy methodology.

---

Copernicus Sentinel-2 Surface Reflectance Harmonized, accessed via Google Earth Engine.

---

USGS 3D Elevation Program (3DEP); USGS National Hydrography Dataset (NHD); USDA-NRCS Soil Survey Geographic Database (SSURGO); NOAA Atlas 14 Precipitation-Frequency Atlas.

---

Haiti 2010 LiDAR: World Bank / ImageCat / RIT, via OpenTopography – doi.org/10.5069/G96Q1V50. Kumamoto 2016 LiDAR: Asia Air Survey Co., via OpenTopography.

---

Building footprints: OpenStreetMap contributors, via the Overpass API. ODbL.

---

Software: pysheds (Bartos); WhiteboxTools (Lindsay); GDAL/OGR; rasterio; GeoPandas; Shapely; pyproj; laspy; SciPy; Leaflet.js.

---

Jose Escobar · jaemaciel.55@gmail.com · github.com/Goheist01 · goheist01.github.io/jose-escobar-cv

All code MIT-licensed and public. All datasets public. Corrections welcome — if something here does not hold up, I would rather know.