Predicting the Best Places to Build a Dam

A recreated English ArcGIS geospatial siting analysis of the Yarlung Tsangpo River, Tibet based on my work with Tsinghua University's Institute of Climate Change & Sustainable Development in SupermapGIS (unavailable outside of China). The way this document is written and formatted is to be understandable by both professionals and general reader.

Abstract

I rebuilt a map-based model that predicts the best locations to build a hydropower dam along a 60-kilometre stretch of the Yarlung Tsangpo River, the site of what will become the largest hydropower project in the world. The model uses only free, publicly available data. I then tested it against the location the Chinese government approved. The model independently identified the same corridor the real engineers chose, and it revealed the trade-off built into their exact site choice. The final product is an interactive three-dimensional map that anyone can explore, click through, and understand.

Introduction

The Yarlung Tsangpo River makes a dramatic horseshoe bend in the Himalayas, dropping roughly 2,000 metres over just 50 kilometres as it wraps around a 7,700-metre peak. That enormous drop is exactly what makes it one of the best hydropower sites on Earth. In my original work on Supermap I had access to personal and government collected vector data[1]. Unfortunately, as I no longer have access to the original data, the question I set out to answer was simple to state:

> *Using only open data, where would a model predict the best places to build a dam — and would those predictions match where a dam is actually being built?*

What this project demonstrates

- **Spatial analysis and modelling**: turning raw terrain and river data into a weighted, defensible prediction.
- **Working with real, imperfect data**: sourcing, cleaning, and combining datasets, and being transparent about their limits.
- **Judgement, not just tools**: choosing which factors matter, how to weight them, and how to interpret the result honestly.
- **Communication**: packaging a technical analysis into an interactive map with plain-language explanations that a non-expert can follow.

How the model works, in plain terms

A dam site is "good" or "bad" for several independent reasons at once. My model scored every stretch of the river against four factors, gave each a rating from 1 (poor) to 5 (excellent), and then combined those ratings into a single score. Think of it like rating a restaurant on food, service, price, and location, then blending those into one star rating, with some factors weighted more heavily than others.

The four factors, and why each matters:

1. The drop in the river (weighted most heavily). Hydropower comes from falling water, the steeper the river falls, the more energy it produces. I measured how much each stretch of river dropped over its length. This is the single most important factor, so it carried the most weight in the final score.

2. The size of the river. A steep river with barely any water in it produces little power. A steep and large river produces enormous power. Direct flow measurements weren't available, so I used the size of the land area draining into each stretch as a stand-in, a larger drainage area means more water.

3. How narrow the valley is. A dam is far cheaper and more practical to build in a tight, steep-walled gorge than in a wide, open valley. There's no direct "valley width" dataset, so I measured how steep the surrounding terrain was as a proxy: steep walls mean a confined gorge.

4. How close towns are (this one counts against a site). The first three factors reward a location. This one penalises it: building near towns means a reservoir would flood people out of their homes. The closer a stretch is to a settlement, the lower it scored. I weighted this factor lightly, reflecting a stated priority on energy potential over displacement, a deliberate choice I documented rather than hid.

I combined the four ratings into one final suitability score, with the weights adding up to a clean total so the result stays easy to read. Then I pulled out the highest-scoring stretches as the model's predicted best sites.

Method

The whole analysis was done in ArcGIS Pro, a professional mapping and spatial-analysis application. In broad strokes:

Stage	What I did
Set up	Downloaded elevation data (a digital elevation model[2], essentially a 3-D map of the terrain) and the river network; aligned everything to a common coordinate system[3] measured in metres so distances and areas would calculate correctly.
Isolated the river	The raw data included thousands of small tributaries. I filtered them out to keep only the main river, which the data conveniently splits into 145 short segments.
Scored each factor	For each of the 145 river segments, I derived the drop, the river size, the valley steepness, and the distance to the nearest town, then converted each into a 1–5 rating.
Combined the scores	Blended the four ratings into one weighted suitability score per segment.
Extracted predictions	Selected the top-scoring segments as the model's predicted best dam sites.
Validated	Overlaid the real, approved dam location and existing dams to see whether the model agreed with reality.
Presented	Built an interactive 3-D scene where the suitability score rises as a coloured wall along the river, with clickable information for every feature, and published it online.

Formal Breakdown

From the sourced data I reclassified[4]*Formal Breakdown*

raw value measurements such as metres per kilometre, square kilometres, or degrees into a common 1-5 scale. There were four criteria used for this calculation

1. Head (River Gradient)
2. Discharge (River Size)
3. Valley Confinement (via terrain-slope proxy)
4. Settlement Proximity

Head (River Gradient)

Purpose: Hydropower is fundamentally about Hydraulic/piezometric Head, the vertical distance that water drops (normalized by specific weight) and the liquid elevation above a vertical datum. Therefore, the more a river drops over a given length, then the more potential energy is available that an Hydro dam can use to generate electricity.

Head was approximated per reach by the channel's vertical drop: using Add Surface Information[5], each reach[6] was stamped with its maximum and minimum elevation (Z_Max, Z_Min)[7]; gradient[8] was then calculated as (Z_Max − Z_Min) / length, in metres per kilometre. This channel gradient is a proxy for the hydraulic head available to a dam, not an engineered head figure, which would depend on dam height and intake placement, but a defensible measure of which reaches offer the most vertical drop to exploit for a dam.

Thus, the gradient is the direct measure of this and is the entire reason the Great Bend is a world-class site for hydropower. The Great Bend's gradient reaches a fall of 60-100+ metres per kilometre of river length. Since the vertical drop is the single strongest physical driver of hydropower potential, Head carried the heaviest weight in the final model.

Calculation:

grad_m_km = (Z_Max − Z_Min) / LENGTH_KM

This is the elevation drop across a reach (its highest point minus its lowest point) divided by the reach's length in kilometres, giving gradient in metres per kilometre.

Geoprocesses used:

- Add Surface Information: with the digital elevation model as the input surface and the river reaches as the input features. This tool "drapes" each reach over the terrain and stamps its attribute table[9] with Z_Min and Z_Max, the lowest and highest elevations the reach touches. This is how a river line acquires elevation information from a raster[10].
- Calculate field: Used to compute grad_m_km from Z_Max, Z_Min and the existing Length_KM field.

Reclassification:

Gradient (m/km)	Head score
< 2	1

2-5	2
5-15	3
15-30	4
> 30	5

Discharge (river size): weighted 0.25, via a drainage-area proxy

Purpose: Power output depends on head *and* the volume of water flowing through. The physical relationship is $P \approx \rho \cdot g \cdot Q \cdot H \cdot \eta$, where Q is discharge (flow) and H is head, so a steep trickle generates very little, while a steep *and* large river generates enormous power.

The proxy: True measured discharge wasn't available in this build of the river dataset. Therefore, I used upstream drainage area (UPLAND_SKM[11], the size of the land area that drains into each reach, in square kilometres) as a stand-in. Logic being, a larger catchment collects more rainfall and snowmelt, so more drainage area means more water potential.

Geoprocesses used:

- Calculate Field — to produce dis_score by reclassifying UPLAND_SKM

Reclassification:

Upstream drainage area (km²)	dis_score
< 5,000	1
5,000 – 20,000	2
20,000 – 50,000	3
50,000 – 100,000	4
>100,000	5

Valley confinement

Purpose: A dam is far cheaper and more feasible to build where the valley is narrow and steep-walled. A tight gorge requires far less material to impound a given volume of water than a wide floodplain, where you'd need an enormous structure to hold back the same reservoir. Constructability matters, and valley confinement is the spatial factor that captures it.

Data Proxy: There is no accurate "valley width" data set, so the mean steepness of the terrain corridor around each reach was used instead. The steep surrounding terrain implies confining valley walls close to the channel. This criterion required the most processing because the value lived in a raster and had to be summarised into a score.

Geoprocesses used:

1. Slope[12]: run on the digital elevation model, producing a slope raster (slope_deg) where every cell holds the terrain steepness in degrees.
2. Pairwise Buffer[13]: run on the river reaches, buffering a few hundred metres to either side, producing one corridor polygon per reach.

3. Zonal Statistics as Table: with the buffer polygons as the zones and the slope raster as the values, computing the MEAN slope inside each buffer. Output is a table: one row per reach, holding its mean surrounding slope.
4. Add Join[15]Zonal Statistics as Table: with the buffer polygons as the zones and the slope raster as the values, computing the MEAN slope inside each buffer. Output is a table: one row per reach, holding its mean surrounding slope.
5. : to attach that statistics table back onto the reaches by a shared identifier, then Calculate Field to copy the mean-slope value into a permanent field (valley_slope) on the reaches, then Remove Join (a join is temporary; baking the value into a real field makes it permanent).
6. Calculate Field: to produce valley_score by reclassifying valley_slope.

Reclassification: After checking the mean slopes clustered in a 29–39° band, I set the breakpoints to spread the range across different classes.

Mean surrounding slope (degrees)	Valley_score
< 20	1
20 – 27	2
27 – 32	3
32 – 37	4
>37	5

Settlement Proximity

Purpose: Unlike the first three factors, settlement proximity is a cautionary constraint. Building a dam near town means the reservoir behind it floods people out of their homes causing displacement. However, it is important to note, that depending on priority, settlement proximity may be weighted positively as well. For example, settlements that would be affected by a new energy grid or opportunities for work. However, in this case, I decided displacement to be a topic of concern as in my original work with Tsinghua only glanced over this concern, because residents in the area were open to relocation and were provided with government financed housing, relocation and infrastructure development.

Calculation: Distance from each reach to its nearest settlement was computed with the Near tool[16], which measures the straight-line distance from every reach to its closest settlement point and writes it to a NEAR_DIST field. Because the data was in a projected coordinate system (UTM Zone 46N[17]), distances were returned in metres

Geoprocesses used:

1. Near — with the river reaches as input and the settlement points as the near features. This tool measures the straight-line distance from each reach to its closest settlement and writes it into a NEAR_DIST field (in metres, because the data was in a projected coordinate system). (Because the data was in projected metres, the distances came out correctly in metres; in a geographic coordinate system this would have produced meaningless "degrees," which is why projecting everything first mattered.)

2. Calculate Field — to produce set_score by reclassifying NEAR_DIST, inverted.

Reclassification:

Distance to nearest town (km)	set_score
< 1	1
1 – 3	2
3 – 8	3
8 -18	4
> 18	5

Weighted overlay (scores combined):

With all four criteria expressed as 1–5 columns, one final **Calculate Field** produced the combined suitability index[18]:

suit_v3 = (head_score × 0.40) + (dis_score × 0.25) + (valley_score × 0.25) + (set_score × 0.10)

Purpose: This is weighted overlay[19], the standard method for fusing multiple criteria. Each score is multiplied by a weight reflecting its relative importance, and the weights sum to 1.0 so the combined result stays on the same intuitive 1–5 scale. Head leads at 0.40 because vertical drop is the decisive physical factor; discharge and valley confinement share 0.25 each as strong secondary factors; settlement proximity contributes a light 0.10 as a constraint.

Results

The suitability model produced a clear spatial gradient across the 145 mainstem reaches. The Great Bend corridor scored consistently within the upper suitability classes, driven principally by its exceptional channel gradient and large upstream drainage area. Reaches in the top class (combined suitability index suit_v3 ⩾ 4.5) concentrated along the northern loop of the bend, where steep, confined terrain coincided with distance from settlements.

Overlaying the announced location of the Medog Hydropower Station, together with the existing upstream cascade, against the modelled surface provided an independent check on the results. Constructed solely from open terrain and hydrographic data, the model identified the same corridor selected for the approved project, indicating that the four physical and constraint criteria capture the dominant drivers of the real siting decision.

The approved site did not, however, coincide with the single highest-scoring reaches. It fell within the high-suitability class (suit_v3 4.15–4.50) rather than the top class, while the model's peak-scoring reaches lay slightly upstream in a more remote section of the loop. This offset is attributable to the settlement-proximity constraint: the approved site lies nearer to inhabited areas and therefore incurred a displacement penalty that the model applied but that the actual decision evidently weighed differently. The divergence is also consistent with siting factors beyond the scope of an open-data model, notably the tunnel-diversion geometry that routes flow across the neck of the bend, and proximity to existing road access, which plausibly favoured the selected reach on engineering and logistical grounds. Convergence at the corridor scale, combined with an explicable divergence at the reach scale, offers stronger evidence of a well-specified model than an exact match would, the latter potentially indicating overfitting to a known outcome.

Limitations

An honest analysis names its limits:

- **Open data only** — it cannot account for the underground rock conditions, earthquake fault lines, or political factors that also drive real dam siting.
- **Some inputs are approximations** — the river outline is generalised, dam coordinates are the publicly announced (not surveyed) locations, and a few town populations are estimated or dated.
- **The weightings reflect my judgement** — I stated them openly so anyone can see the model's priorities and test how sensitive the result is to them.

Being upfront about these limits is not a weakness in the work — it is what separates a credible planning analysis from a map that merely looks authoritative.

Tools and data

- **Software** — ArcGIS Pro (analysis and 3-D scene), ArcGIS Online (interactive web publishing).
- **Elevation** — Copernicus GLO-30 digital elevation model (European Space Agency / Copernicus, via OpenTopography).
- **River network** — HydroRIVERS (World Wildlife Fund / HydroSHEDS).
- **Settlements and population** — OpenStreetMap and China's 2020 census.
- **Dam locations** — published reporting.

Conclusion

This study demonstrated that a transparent, reproducible multi-criteria suitability model, constructed entirely from openly available terrain and hydrographic data, can independently identify the corridor selected for a major hydropower development and characterise the trade-offs embedded in the final siting decision. Using four weighted criteria, hydraulic head, discharge, valley confinement, and settlement proximity; the analysis reproduced the regional siting logic of the Medog Hydropower Station and situated the divergence between the modelled optimum and the approved site within an interpretable framework of displacement cost and unmodelled engineering constraints.

The methodology is generalizable, deriving physical and constraint criteria for each spatial unit, normalising them to a common scale, combining them through weighted overlay, and validating against a known outcome, transfers directly to other siting and suitability problems, including flood-risk zonation, protected-area prioritisation, renewable-energy siting, and emergency-facility placement. Its principal value lies not in prescribing a single answer, but in rendering the assumptions, weights, and limitations explicit and testable.

The results are subject to the limitations outlined above, chiefly the reliance on open-data proxies and the sensitivity of the outcome to the selected weights; they are therefore best interpreted as decision-support screening rather than a definitive siting recommendation. An interactive three-dimensional version of the analysis is available online, in which each reach can be queried for its component and composite scores.

Glossary

Spatial and GIS terms

[1] **Vector —** spatial data represented as points, lines, or polygons (for example, river lines and settlement points) rather than a grid.

[2] **Digital elevation model (DEM) —** a raster dataset in which every cell stores the ground elevation at that spot; effectively a three-dimensional model of the terrain surface.

[3] **Coordinate reference system / projection —** the framework used to locate features on the Earth. A projected system (such as Universal Transverse Mercator) measures in metres; a geographic system measures in degrees of latitude and longitude.

[4] **Reclassification —** converting a raw measurement (metres per kilometre, square kilometres, degrees, metres) into a common ranked scale — here 1 (poor) to 5 (excellent) — so that unlike factors can be combined.

[5] **Add Surface Information —** a tool that reads a surface (the DEM) beneath each feature and stamps the feature's attribute table with elevation statistics — here, the highest and lowest elevations each reach touches (Z_Max, Z_Min).

[6] **Reach —** a single segment of a river channel, usually running between confluences. The unit of analysis in this project; the mainstem was made up of 145 reaches, each scored individually.

[7] **Z_Max / Z_Min —** the highest and lowest elevations a reach touches, in metres above sea level.

[8] **Gradient (channel gradient) —** the change in elevation along a reach divided by its length, in metres per kilometre; used as a proxy for hydraulic head.

[9] **Attribute table —** the data table behind a spatial layer: one row per feature, one column (field) per piece of information.

[10] **Raster —** grid-based spatial data made of cells (pixels), each holding one value, such as elevation or slope.

[11] **Drainage area / catchment (UPLAND_SKM) —** the total land area that drains into a given point on the river, in square kilometres; used as a proxy for discharge.

[12] **Slope —** the steepness of the terrain surface, in degrees, derived from the DEM.

[13] **Buffer —** a zone of a set distance drawn around a feature; here, a corridor a few hundred metres wide around each reach.

[14] **Zonal statistics —** a tool that summarises raster values inside defined zones; here, the mean terrain slope inside each reach's buffer.

[15] **Join —** a temporary link between a separate table and a spatial layer, matched on a shared identifier.

[16] **Near —** a tool that measures the straight-line distance from each feature to the nearest feature in another layer.

[17] **Universal Transverse Mercator, Zone 46N (UTM 46N) —** the projected coordinate system used throughout this analysis. Its units are metres, which is what makes distance, length, and area calculations come out correctly.

[18] **Suitability index (suit_v3) —** the final combined 1–5 score for each reach, produced by the weighted overlay.

[19] **Weighted overlay —** the standard method for fusing multiple ranked criteria into one score, multiplying each by a weight that reflects its importance, with the weights summing to 1.0.

Symbols in the hydropower equation ($P \approx \rho \cdot g \cdot Q \cdot H \cdot \eta$)

ρ — the Greek letter "rho," the density of water (about 1,000 kilograms per cubic metre). This converts a volume of water into a mass, because it's the mass of the falling water that carries energy. It's essentially a constant here, water’s density doesn't change from site to site, so it doesn't affect where you'd site a dam.

g — *gravitational acceleration* that pulls the water down (~9.81 m/s^2).

Q — discharge, *the flow rate (m^3/s)*. This is how much water passes through per second, or the volume of water flowing through.

H —head, the total mechanical energy of water per unit weight expressed as height, the vertical drop (metres). This is how far the water falls. It is made up of components: elevation, water pressure and velocity.

η — the Greek letter "eta," *efficiency* (a fraction between 0 and 1, typically ~0.8–0.9) the ratio of useful work or energy output to total input energy in engines and thermodynamic systems. No turbine converts 100% of the water's energy into electricity, some is lost to friction and heat. This just scales the result down to how much energy might formally be captured.