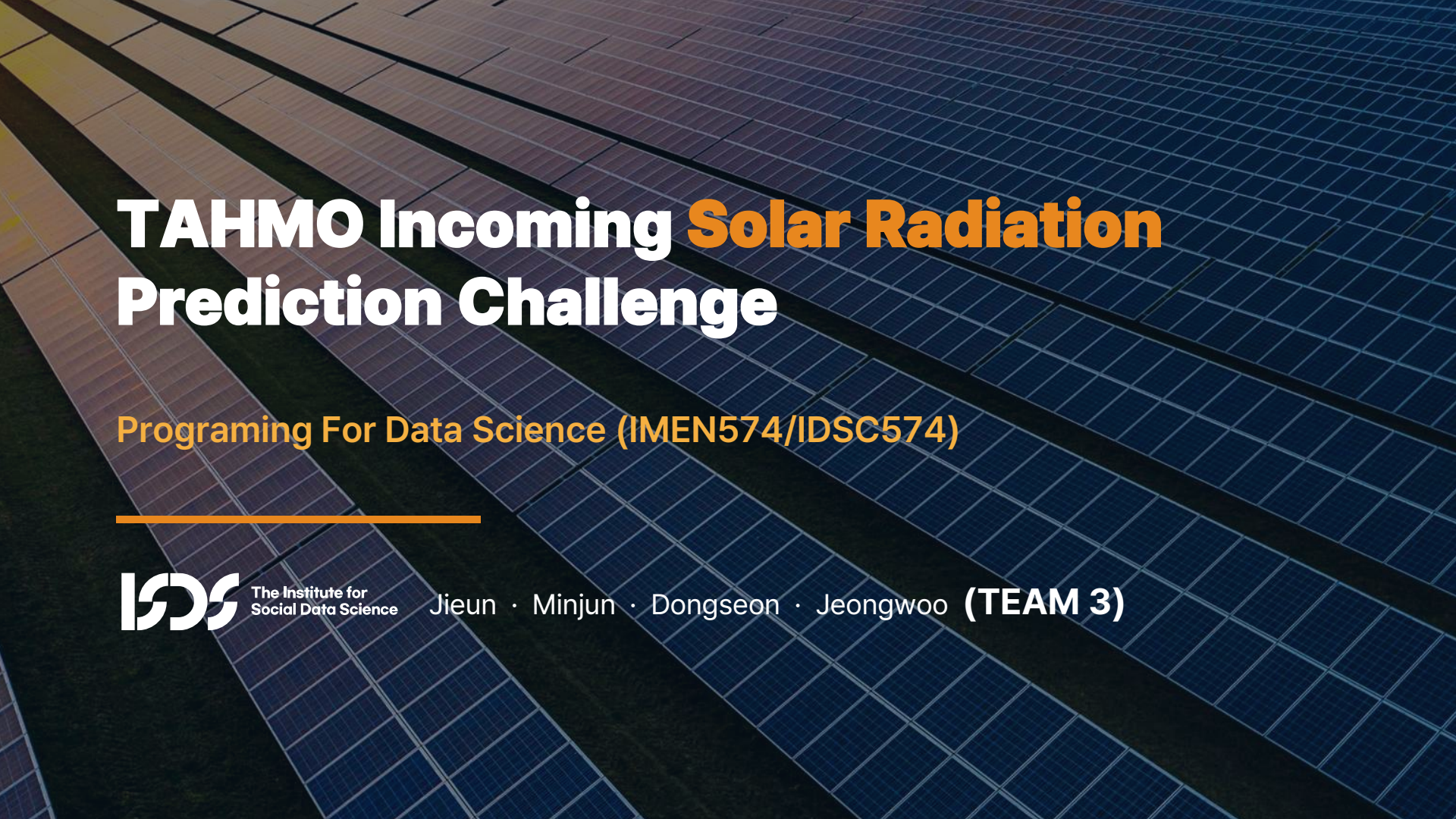


# TAHMO Incoming Solar Radiation Prediction Challenge

## Programing For Data Science (IMEN574/IDSC574)

The Institute for Social Data Science

Jieun · Minjun · Dongseon · Jeongwoo **(TEAM 3)**

# Project overview

| ORGANIZER | PARTICIPANTS | PRIZE | DURATION |
| --- | --- | --- | --- |
| TAHMO DeepMind | 1,529 joined<br>761 active | $10,000 USD | Apr 1 – May 25 2026 |

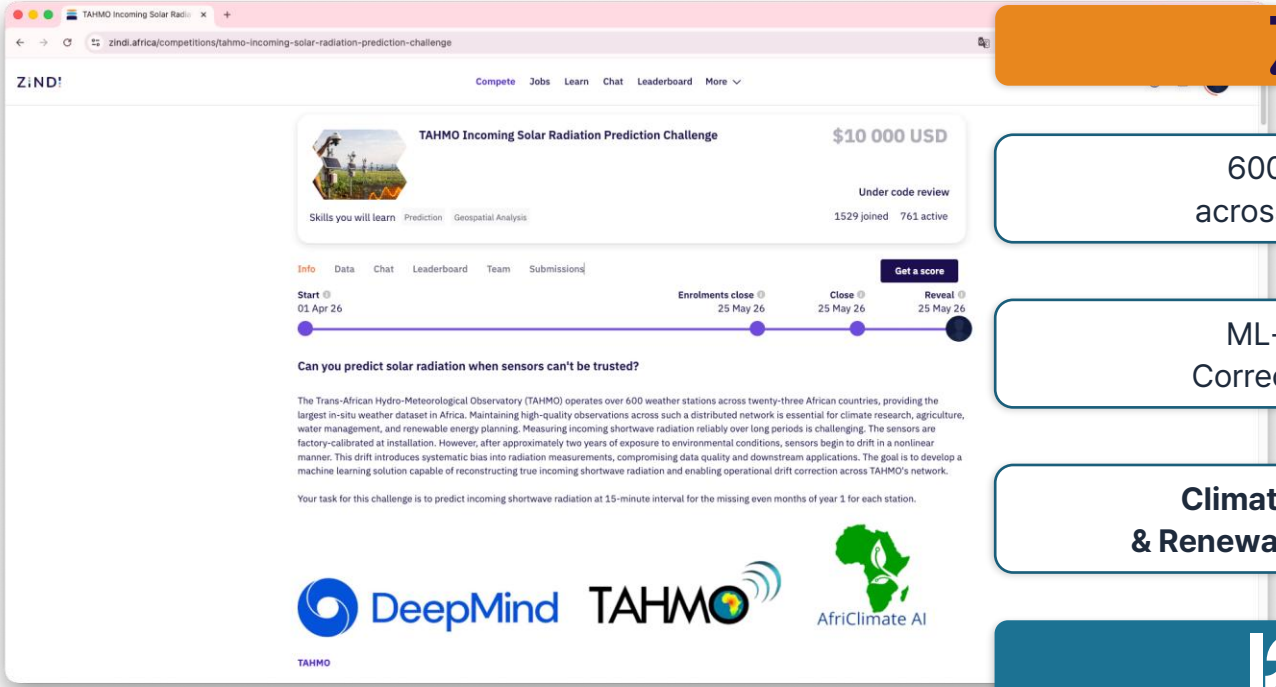
## What is TAHMO?

The Trans-African Hydro-Meteorological Observatory operates 600+ weather stations across 23 African countries — the largest in-situ weather dataset in Africa.

## The problem: sensor drift

After ~2 years of exposure, solar radiation sensors begin to drift in a nonlinear manner. This introduces systematic bias, compromising data quality for climate research, agriculture, and renewable energy planning. The goal is to build an ML model that can reconstruct the true incoming shortwave radiation.

# Why This Challenge Fits ISDS




ZiNDI

600+ Weather Stations
across 23 African Countries

↓

ML-based **Sensor Drift**
Correction & Reconstruction

↓

**Climate · Agriculture · Water**
**& Renewable Energy Infrastructure**

ISDS The Institute for Social Data Science

Solving **Real Social Infrastructure Problems** with Data Science

# Task: predict even months from odd months

**Predict** 15-min incoming shortwave radiation for **missing even months** using odd-month observations.

| Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|---|---|---|---|---|---|---|---|---|---|---|---|
| TRAIN | TEST | TRAIN | TEST | TRAIN | TEST | TRAIN | TEST | TRAIN | TEST | TRAIN | TEST |

Odd months (observed)  **Even months (predict)**

| | |
|---|---|
| STATIONS | 40 TAHMO stations across Africa |
| GRANULARITY | 15-minute intervals |
| TRAIN ROWS | 642,175 |
| TEST ROWS | 683,353 |



## Evaluation

Abs MBE (bias) × 50%  +  RMSE (accuracy) × 50%

# What We Tried
# More than 86 times

# EDA of the Provided Base Data

TAHMO Solar Radiation Prediction | Based on Train.csv / Test.csv

**Core task:** reconstruct 15-minute radiation for even months from odd-month observations

| 642,175 rows | 683,353 rows | 40 stations · 7 countries | 15-minute grid |
|---|---|---|---|
| Train: odd months / target available | Test: even months / target hidden | Same stations in Train and Test | 0% missing in basic variables |

## Key Findings from EDA

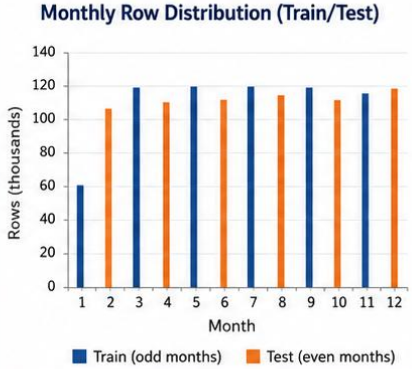
**Split Structure**
Train uses months 1·3·5·7·9·11, while Test uses 2·4·6·8·10·12. The model reconstructs even months for the same 40 stations.

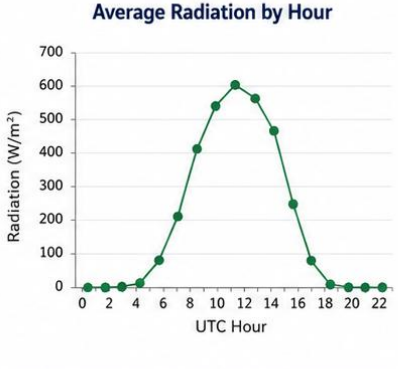
**Target Distribution**
Mean radiation is 190.4 W/m² and the median is 0. Since 50.3% of values are zero, separating day and night is the first step.

**Source of Variability**
Midday radiation is high, but the standard deviation is about 270 W/m². Time pattern alone cannot explain clouds, humidity, and regional differences.

**Data Quality Signal**
There are 6,067 daytime cases (UTC 08–14) with radiation = 0. These likely reflect sensor downtime and should be removed or flagged.

**Modeling Implication**
Go beyond UTC hour: use solar zenith angle or local hour together with station encoding, humidity, and temperature.




**Summary**
The overall monthly mean looks flat, but seasonal patterns differ by country and station. Therefore, time features + station features + weather variables matter more than simple month interpolation.

Station mean radiation range: 26.0–259.2 W/m²

**Takeaway |** **This is not a simple 'monthly interpolation' problem; it is a radiation reconstruction problem that must jointly model solar geometry, local climate, and sensor anomalies.**

# Data: 6 external sources, 84 columns

| | |
|---|---|
| **TAHMO**<br>Ground observations | Temperature, humidity, precipitation, measured radiation |
| **LSA SAF**<br>European satellite | DSSF, diffuse fraction, AOD, opacity index |
| **CAMS**<br>Atmospheric analysis | All-sky / clear-sky GHI, cloud info, diffuse fraction |
| **NASA POWER**<br>Reanalysis (hourly) | Radiation, cloud amount, temperature, wind, humidity |
| **ERA5**<br>ECMWF reanalysis | Cloud cover, pressure, dew point, wind, precipitation |
| **AOD/Albedo/MLST**<br>Atmos. & surface | Aerosol optical depth, surface reflectance, land temp |

Key: We used LLMs to search papers and product docs for variable candidates — no geoscience expert on the team.

# Preprocessing Strategy

From heterogeneous external sources to one leakage-safe model table

**Raw inputs**

- **TAHMO** — 15-min station data
- **LSA-SAF / DSSF** — satellite radiation
- **CAMS / POWER** — external weather & GHI

→ **Model-ready table**
one row = one station × one timestamp

## Preprocessing flow

**1 Align**

- Match all sources to TAHMO 15-min timeline
- Extract nearest grid value for each station lat/lon

→

**2 Standardize**

- Convert raw satellite values to W/m²
- Set night radiation to 0 and clip impossible ranges

→

**3 Validate**

- Remove target-leaking columns before prediction
- Keep only available information

→ **Ready for Model**

### Quality layer: do not hide data problems - make them learnable signals

| Missing flags | Quality flags | Source gaps | Night / physical rules |
|---|---|---|---|
| missingness kept as information | LSA-SAF bits split into features | DSSF, CAMS, POWER disagreement | night = 0; impossible values clipped |

# Algorithms & ensemble strategy

### LightGBM

Main model

Gradient boosting — strongest for tabular data

### LightGBM Tweedie

Zero-inflated

For data with many zeros (nighttime = 0)

### XGBoost

Diversity

Alternative gradient boosting for blending

### CatBoost

Diversity

Handles categorical features natively

### Ridge (Meta)

Blender

Weighted average — the final arbiter

## Target transforms

Direct (raw W/m²) · Clearness Index (kt) · Log1p (log-transformed)

## Validation

Month-held-out CV within odd months · 3-seed ensemble to reduce variance

# How we used AI tools

| | | |
|---|---|---|
| Early use | GPT<br>as a code helper | Asked for notebook cells one by one / Useful for baseline, preprocessing, and starter models<br>Slow once the work became multi-file / Error handling and lineage stayed manual |
| Later use | Codex<br>as a project agent | Read project files and prior runs / Edited scrips, executed check, fixed errors<br>Compared candidate submissions and diagnostics / Wrote method summaries and sharing pack |
| Harness Engineering | | Versioned runs / Smoke vs full validation gates<br>Station-month and daylight diagnostics / Submission format guard and input lineage |
| RALPH Loop | | Read problem/artifacts → Argue the physical hypothesis → Learn from validation → Patch code → Harden reproducibility |

Lesson: AI was not the model, AI was the experiment operating system.

# What Actually Worked

# Feature engineering 1

Solar Position and Physical Radiation Potential

## Time Cyclicity

**Converted timestamps** (hour, month, day-of-year) into sin/cos pairs so the model understands that 23:45 and 00:00 are neighbors - not 24 hours apart.

> **Why:** Radiation follows daily and seasonal cycles. Without cyclic encoding, the model thinks December and January are 11 months apart instead of 1.

## Solar Geometry

**Calculated the sun's position** from each station's coordinates and timestamp. Two stations at the same UTC time can have completely different sun positions.

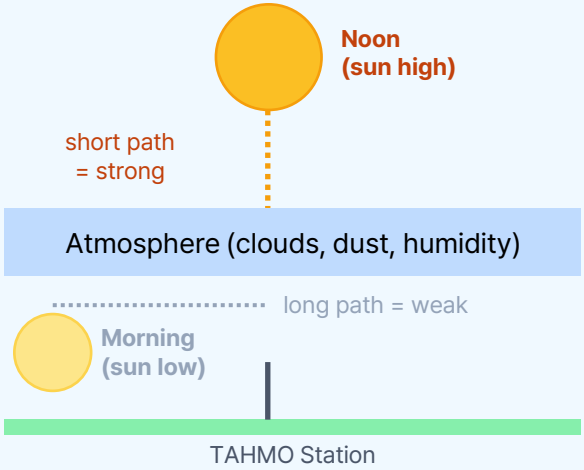
> **Why:** Sun height is the #1 driver of radiation. When the sun is low, light passes through more atmosphere and weakens. This single feature family explains most of the variation.

## Physical Radiation Baseline

Estimated the **maximum possible sunlight under a perfectly clear**. This sets the ceiling - actual radiation is always lower due to clouds, dust, and humidity.

> **Why:** To predict how much sunlight actually arrives, the model needs to know how much could arrive. The gap tells us how much was blocked.

### How sun height affects radiation




**(1) Time cyclicity**
= when to expect sunlight
**(2) Solar geometry**
= where the sun is right now
**(3) Radiation baseline**
= max possible sunlight (clear sky)

# Feature engineering 2

Atmospheric Attenuation & Short-term Variability

## Clearness Index

**Divided actual radiation** by the clear-sky baseline to get a ratio between 0 and 1. A value of 0.3 means clouds or dust are blocking 70% of the sunlight.

> Why: Raw radiation changes with time of day and season. The clearness index removes that variation, so the model can focus purely on how much the atmosphere is blocking.

## Source Disagreement

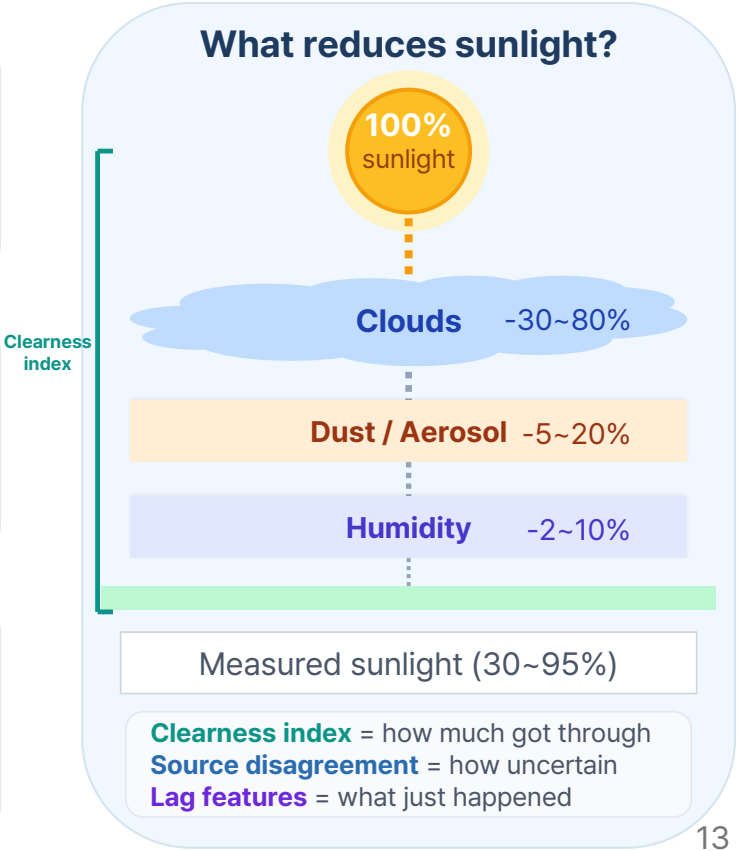
**Compared radiation estimates.** When they agree, conditions are stable. When they disagree, it signals cloudy, dust events, or sensor issues.

> Why: No single source is perfectly accurate. Disagreement itself is information - it tells the model that this moment is hard to predict and that some inputs may be unreliable.

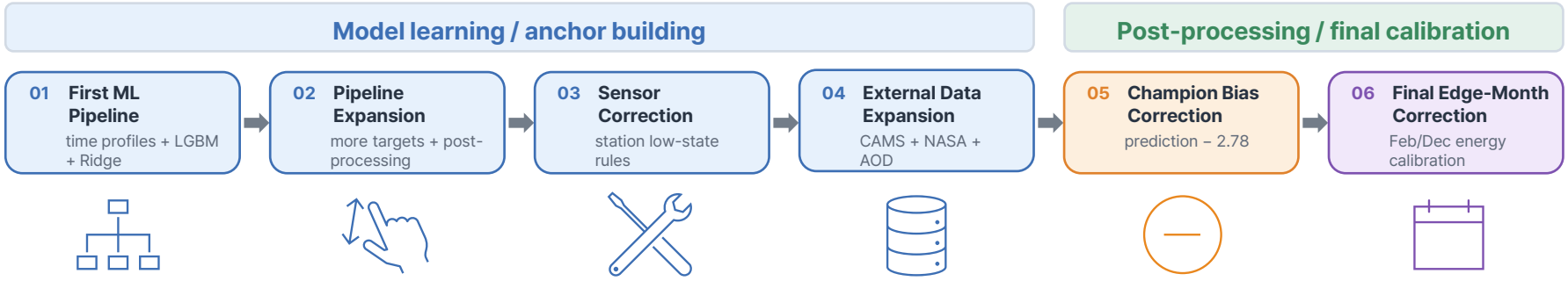
## Lag / Rolling Features

Added **"what happened 15 and 30 minutes ago"** as features , plus rolling averages and standard deviations over recent windows.

> Why: Clouds move continuously. A single snapshot can't capture whether radiation is rising or falling. Recent history helps the model detect short-term weather changes.



# Full Project Flow




**What changed over time?**

Early versions built model capacity. Later versions did not simply add complexity; they corrected specific bias and sensor patterns.

**Main lesson**

Complex ML created the foundation, but small calibrated corrections produced the biggest final gains.

# 01. First ML Pipeline

The first version already combined initial satellite variables, temporal profiles, LightGBM, and Ridge blending

## Initial feature set used

### TAHMO station observations

Station-time records used as the base table

- station and timestamp
- observed weather context
- target radiation in train

### LSA SAF satellite features

Core external radiation variables in the first model

**DSSF**
satellite-based surface shortwave radiation baseline

**Fraction Diffuse**
diffuse-light ratio; cloud / scattering signal

### Other engineered features

Expanded the raw table into model inputs

- time and station features
- solar-geometry features
- ~28 raw cols → ~70 features

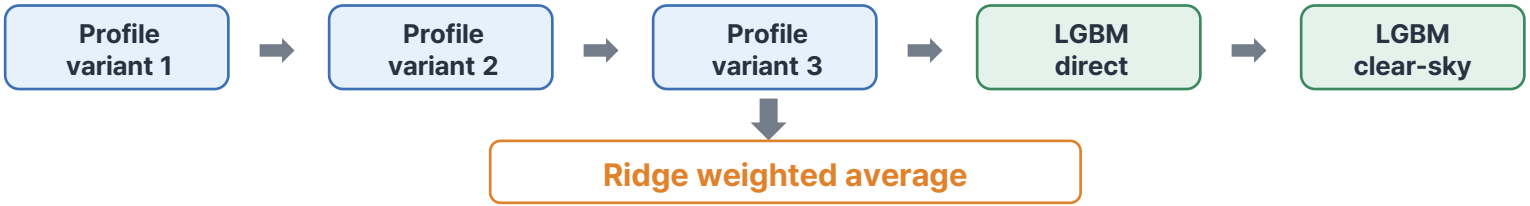
## These features fed two first-stage prediction branches

### Branch A: Temporal Profile Interpolation

Uses nearby months for the same station and 15-minute time slot.
Example: if March 2pm looked like this, April 2pm may be similar.

### Branch B: LightGBM Models

Two initial model views: direct radiation prediction and clear-sky-index prediction, both trained on the same feature set.



# 02. Pipeline Expansion

Expanded model views and added layered post-processing



## More LightGBM targets

| | |
|---|---|
| **direct** | raw radiation |
| **clear-sky index** | normalized ratio |
| **log1p** | skewed target stabilization |
| **daytime-only** | focus on daylight rows |

## Layered correction after prediction

| | |
|---|---|
| **Global bias correction** | subtract overall OOF MBE |
| **Station bias correction** | shrink station-level residual means |
| **Station × time residual profile** | fix repeated local residual patterns |
| **Champion drift guard** | pull risky predictions toward the previous best |

# 03. Defective Sensor Correction

Some patterns looked like station sensor problems, not general weather patterns

**Satellite sources**
say radiation is high

→

**Ground pattern**
is near zero

→

**Same station-month**
repeats the pattern

→

**Rule-based fix**
scale down or zero out

**Examples of conservative station rules**

**TA00338**
months 4/6/8/10
multiply by 0.25

**TA00123**
month 8
multiply by 0.35

**TA00122**
months 2/4
set prediction to 0

# 04. External Data Expansion

Added more atmospheric context and stronger LightGBM settings

### CAMS

atmosphere, aerosol, cloud-related signals

### NASA POWER

reanalysis weather and radiation variables

### AOD

aerosol optical depth

### Albedo

surface reflectance

### Feature expansion

External data created comparison features such as DSSF vs CAMS. The engineered feature set reached about 129 features.

### Model strengthening

A heavier LightGBM setup was used, with multiple seeds and stronger tree capacity. Auto sensor-state mining was also added.

### Safety strategy

New model outputs were blended with the previous champion to avoid large risky shifts.

# 05. Champion Bias Correction

The most effective single correction was surprisingly simple

**What leaderboard showed**

The champion prediction was, on average, too high by about 2.78 W/m².

**corrected = champion_prediction − 2.78**

| | |
|---|---|
| **No new learning** | This stage only adjusted an existing prediction file. |
| **Night guard** | solar elevation ≤ −2° → prediction = 0 |
| **Physical clipping** | final predictions clipped to [0, 1450] W/m² |
| **Best individual score** | Score 0.4402 in the shared notes |

# 06. Final Edge-Month Correction

The final submission reused the best anchor shape and corrected only February and December

**Strong anchor**

Use the stable 15-minute prediction shape as the backbone.

→

**Edge-month energy**

Estimate monthly clear-sky-index energy using adjacent observed odd months.

→

**Final submission**

Keep the anchor shape; adjust monthly energy by ratio scaling.

## How the edge months are inferred

**February**

January observed + March observed
→ February energy estimate

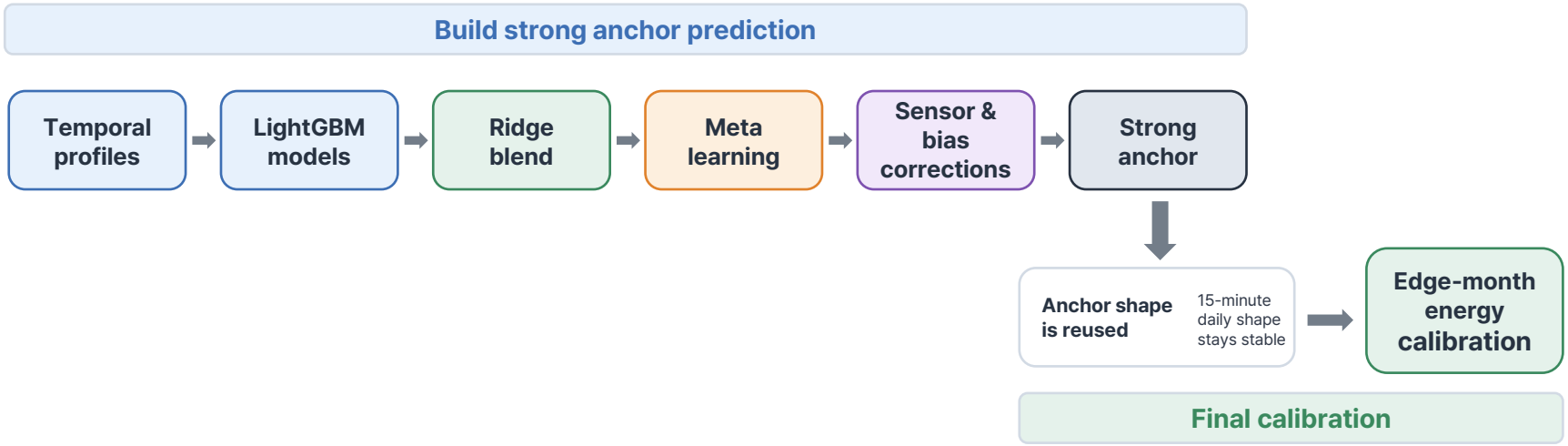
**December**

September observed + November observed
→ December energy estimate

**Mean-neutral constraint** The correction redistributes energy while keeping the global prediction mean almost unchanged.

# Final Prediction Structure

The final system separates model learning from final calibration




**Final lesson**

The pipeline's strength came from both sides: ML created a reliable anchor, and post-processing corrected the specific patterns that the model still missed.

# Final result

**38** th

out of 1,529 participants (top 5%)

| Private SCORE | ABS MBE | RMSE |
|---|---|---|
| 0.456 | 2.52 | 60.78 |

TOP 4 COMPARISON

| RANK | TEAM | Private SCORE | ABS MBE | RMSE |
|---|---|---|---|---|
| 1 | Brainiac | 0.630 | 0.78 | 55.82 |
| 2 | eugenius | 0.625 | 0.87 | 55.19 |
| 3 | JZ_4 | 0.625 | 0.82 | 56.10 |
| 4 | fgbfgb | 0.610 | 0.97 | 56.59 |
| | | . . . | | |
| 38 | Postech ISDS | 0.456 | 2.52 | 60.78 |

Gap analysis: RMSE differs by ~5, but MBE differs by ~1.7 — bias control was the decisive factor.

# What we learned

**01** **Data matters more than models**

Adding CAMS + NASA POWER gave the largest RMSE improvement (~13 pts). No model tuning matched that.

**02** **LLMs can fill the domain-knowledge gap**

No geoscience expert on the team — LLMs searched papers, extracted variable candidates, and designed preprocessing.

**03** **Coding agents: powerful but not infallible**

Codex and Claude accelerated workflow, but preprocessing logic must be verified by humans.

**04** **Never underestimate post-processing**

Our best score: a one-line bias correction. In MBE-weighted competitions, calibration is king.

# TAHMO Incoming Solar Radiation Prediction Challenge

# Thank you!

The Institute for Social Data Science

Jieun · Minjun · Dongseon · Jeongwoo