Automated Solar Power Emission Projection Pipeline: A Machine Learning Approach to Climate Transition Risk Assessment

Yiwen Mai, Aya Bekhtiar, Fatma Braham, Apolline Hadjal
Research and emerging topics
https://github.com/yyypsycheguy/Final-Project-Research-Emerging-topics.git

January 3, 2026

1 Introduction

The energy transition requires robust quantitative frameworks to assess climate-related financial risks. As global solar capacity expands, accurately projecting emission reductions under divergent policy pathways becomes critical for investment decisions, policy evaluation, and climate risk disclosure. Traditional approaches rely on static spreadsheet models that lack reproducibility, scalability, and integration with real-time data sources.

1.1 Motivation

The financial sector increasingly demands automated, production-grade analytics to quantify transition risk—the financial exposure arising from misalignment between current assets and future low-carbon scenarios. Solar power, representing the fastest-growing renewable energy source, serves as an ideal case study for developing scalable emission projection infrastructure.

1.2 Contribution

We develop a fully automated, end-to-end machine learning pipeline that:

- Ingests and processes 75,671 solar projects from the Global Energy Monitor database
- Engineers 72 features capturing temporal, geographic, capacity, and risk dimensions
- Trains scenario-specific XGBoost models projecting emissions through 2050
- Quantifies transition risk via scenario divergence metrics
- Deploys as a reproducible, version-controlled system enabling continuous analytics

Our architecture advances prior work by integrating data engineering, feature engineering, statistical modeling, and risk analytics into a unified, production-ready framework suitable for institutional deployment.

2 Data and Methodology

2.1 Data Source

We used the Global Solar Power Tracker (February 2025 release), a comprehensive database maintained by Global Energy Monitor containing detailed information on utility-scale solar projects worldwide. The dataset encompasses:

- **Coverage:** 75,671 projects across 195 countries
- **Capacity Range:** 1 MW to 10,000 MW (total: 3.2 TW)

- **Status Categories:** Operational, Under Construction, Planned, Cancelled, Retired
- **Attributes:** 32 variables including capacity, location coordinates, technology type, commissioning dates, ownership, and operational status

2.2 IEA Scenario Framework

We model three standardized International Energy Agency scenarios representing divergent climate policy pathways:

Net Zero Emissions by 2050 (NZE): Aggressive decarbonization aligned with limiting global warming to 1.5°C. Parameters: 15% annual solar growth, $130/tCO carbon price by 2030, 85% electrification rate.

Announced Pledges Scenario (APS): Implementation of current national climate commitments and announced targets. Parameters: 10% annual solar growth, $75/tCO carbon price by 2030, 65% electrification rate.

Stated Policies Scenario (STEPS): Conservative baseline reflecting only existing and enacted policies. Parameters: 6% annual solar growth, $30/tCO carbon price by 2030, 50% electrification rate.

2.3 Pipeline Architecture

Our system implements a four-phase modular pipeline (Figure 1):

```
Raw Data → Ingestion → Feature Engineering → Modeling → Risk Analysis
 (Excel)     (Clean)      (72 features)      (XGBoost)    (Reports)
```

Figure 1: End-to-end pipeline architecture for emission projection

2.3.1 Phase 1: Data Ingestion

The ingestion module (`data_loader.py`) executes:

1. Load separate sheets for large-scale (20 MW) and small-scale (1-20 MW) projects
2. Validate capacity constraints (1-10,000 MW) and temporal bounds (2000-2050)
3. Remove duplicates via unique `GEM_phase_ID` identifier
4. Standardize column names and status categories
5. Export to compressed Parquet format for efficient downstream processing

Quality metrics: 99.2% data completeness on critical fields (capacity, location, status), zero duplicates after deduplication.

2.3.2 Phase 2: Feature Engineering

We engineer 72 features across six dimensions (`feature_engineering.py`):

Temporal Features (6): Years operational, age categories, construction decade, expected retirement year, time to retirement

Geographic Features (5): Regional grouping (Asia, North America, Europe, etc.), climate zone (tropical/subtropical/temperate/polar), hemisphere, solar irradiance proxy based on latitude

Capacity Features (5): Log-transformed capacity, capacity categories (small/medium/large/utility-scale), technology efficiency proxy

Risk Features (7): Age-based risk score, policy risk by region, technology obsolescence risk, market risk, composite stranded asset risk, retirement probability

Emission Features (6): Lifecycle emissions factor (45-50 kgCOe/MWh for solar PV/thermal), regional grid intensity (275-630 kgCOe/MWh), net emission reduction factor, capacity factor (15-30% based on latitude), annual generation estimate, annual emissions avoided

Scenario-Specific Features (4 per scenario): Growth rate adjustment, carbon price trajectory, retirement rate, competitive advantage score

Feature engineering transforms raw project attributes into predictive signals capturing physical, economic, and policy dynamics.

2.3.3 Phase 3: Emission Modeling

For each scenario, we train gradient-boosted regression models:

Algorithm: XGBoost with hyperparameters: 200 estimators, max depth 8, learning rate 0.05, subsample 0.8

Target Variable: Annual emissions avoided (tCOe) = Annual generation (MWh) × Emission reduction factor (kgCOe/MWh) / 1000

Training Protocol: 80/20 train-test split, 5-fold cross-validation, stratified by region

Projection Method: For year t and scenario s:

$$C_t^s = C_0 \times (1 + g_s)^{t-2024} \times (1 - r_t) \tag{1}$$

where C_t^s is projected capacity, C_0 is base capacity, g_s is scenario growth rate, r_t is retirement adjustment

Emission Calculation:

$$E_t^s = C_t^s \times 8760 \times CF \times (I_{grid} - I_{solar})/1000 \tag{2}$$

where CF is capacity factor, I_{grid} is grid intensity, I_{solar} is solar lifecycle emissions

2.4 Transition Risk Metrics

We quantify transition risk via scenario divergence:

Transition Risk Score:

$$TRS_{t,r} = \frac{|E_t^{NZE} - E_t^{STEPS}|}{E_t^{NZE}} \tag{3}$$

Policy Risk Score:

$$PRS_{t,r} = \frac{|E_t^{APS} - E_t^{STEPS}|}{E_t^{APS}} \tag{4}$$

Stranded Asset Exposure:

$$SAE_{t,r} = \max(0, E_t^{STEPS} - E_t^{NZE}) \tag{5}$$

where t is year, r is region, E is emissions avoided.

3 Results

3.1 Model Performance

Table 1 shows cross-validated performance metrics for scenario-specific models:

Table 1: Model Performance Metrics (5-Fold Cross-Validation)

Scenario	R^2 Score	RMSE	MAPE (%)
NZE	0.891	1,247 tCOe	12.3
APS	0.884	1,305 tCOe	13.1
STEPS	0.878	1,398 tCOe	14.2

Models achieve strong predictive performance (R^2 ¿ 0.87) with reasonable error margins relative to the scale of emissions avoided (mean: 10,200 tCOe per project annually).

3.2 Emission Projections

Global emission reductions show dramatic scenario divergence (Table 2):

Table 2: Global Emissions Avoided by Scenario (MtCOe)

Scenario	2030	2050	CAGR	Total (2025-2050)
NZE	58.7	635.7	15.0%	5,847
APS	45.0	200.1	10.0%	2,134
STEPS	36.0	76.4	6.0%	981
NZE-STEPS Gap	22.7	559.3	–	4,866
% Divergence	63%	732%	–	496%

By 2050, NZE delivers 8.3× the emission reductions of STEPS, accumulating 5.8 GtCOe avoided over 25 years compared to 1.0 GtCOe under current policies.

3.3 Regional Analysis

Asia dominates absolute capacity (45% global total) but faces highest transition risk. Regional breakdown for 2030:

Table 3: Regional Projections for 2030 (NZE Scenario)

Region	Capacity (GW)	Generation (TWh)	Emissions (MtCOe)
Asia	3,046	6,630	26.5
North America	1,523	3,312	13.2
Europe	1,218	2,648	10.6
South America	608	1,323	5.3
Other	373	812	3.2

3.4 Transition Risk Assessment

Transition risk scores escalate over time as scenario pathways diverge:

- **2030:** Average TRS = 0.41, indicating moderate divergence
- **2050:** Average TRS = 0.88, signaling severe misalignment
- **High-Risk Regions (2050):** Asia (TRS=0.88), North America (TRS=0.88), Other (TRS=0.88)

Policy risk (APS-STEPS gap) remains elevated throughout the projection period, averaging 0.42, indicating substantial uncertainty around policy implementation.

3.5 Feature Importance

Top predictive features from NZE model:

1. Log capacity (18.3% importance)
2. Annual generation estimate (15.7%)
3. Grid intensity (14.2%)
4. Years operational (11.8%)
5. Emission reduction factor (10.9%)
6. Solar resource proxy (8.4%)

Capacity and generation metrics dominate, confirming that project scale and geographic solar resource drive emission impact.

4 Discussion

4.1 Implications for Climate Finance

Our results quantify substantial transition risk: assets aligned with STEPS face potential stranding as policy shifts toward NZE. The 732% emission divergence by 2050 translates to massive revaluation risk for solar-exposed portfolios. Financial institutions require scenario analysis tools—like our pipeline—to stress-test holdings.

4.2 Production Deployment Considerations

The pipeline demonstrates key production-readiness attributes:

Modularity: Separation of ingestion, feature engineering, and modeling enables independent component updates and testing.

Reproducibility: Configuration-driven design (`config.yaml`) with fixed random seeds ensures deterministic outputs.

Scalability: Parquet storage, vectorized operations, and XGBoost enable processing of larger datasets (10M+ projects) with minimal code changes.

Maintainability: Comprehensive logging (`loguru`), type hints, and docstrings facilitate team collaboration.

Automation: Single entry point (`main.py`) with CLI arguments supports scheduled execution via cron jobs or workflow orchestrators.

4.3 Limitations and Extensions

Current limitations include:

- Simplified capacity factor model (latitude-based proxy vs. detailed meteorological data)
- Static grid intensity assumptions (no temporal decarbonization of grids)
- Absence of economic variables (LCOE, subsidy schedules, financing costs)
- No uncertainty quantification (point estimates without confidence intervals)

Future extensions could integrate:

- Probabilistic forecasting (quantile regression, conformal prediction)
- Real-time data APIs (automated monthly updates from GEM)
- Economic optimization (least-cost pathways under constraints)
- Web API deployment (FastAPI endpoint for on-demand projections)
- Dashboard interface (interactive Plotly/Streamlit visualization)

5 Conclusion

We present a production-grade automated pipeline for solar emission projections and transition risk analysis. Our system processes 75,671 projects through modular data engineering, feature engineering, and machine learning stages, achieving $R^2 = 0.89$ predictive accuracy. Scenario analysis reveals 732% emission divergence between NZE and STEPS by 2050, quantifying severe transition risk.

The pipeline's value proposition centers on **automated, scalable, reproducible climate analytics**. By encoding domain expertise into engineered features, scenario parameters, and risk metrics, we enable financial institutions, policymakers, and energy companies to continuously assess climate alignment. The fully version-controlled, documented codebase supports institutional deployment with minimal adaptation.

This work demonstrates that rigorous data engineering—modular architecture, automated workflows, reproducible execution—transforms ad-hoc climate analysis into systematic, production-ready infrastructure suitable for enterprise risk management.