A Cloud-Deployed AI System for Short-Term Solar Power Forecasting Using Weather and Satellite Data

Varsh Vishwakarma
Department of Computer Application
Barkatullah University Institute of Technology
Bhopal, India

***Abstract*—As solar energy becomes increasingly critical to the power grid, the ability to predict short-term fluctuations in generation is essential for maintaining stability and efficiency. While many studies propose advanced machine learning models, few address the engineering challenges of deploying these models into live, production-oriented prototypes. This paper presents an end-to-end, cloud-deployed AI system for next-hour solar power forecasting, constructed utilizing real-world generation data from a plant in Bhopal, India, synthesized with satellite-derived weather data from NASA POWER. By engineering time-cyclic and lag features to capture temporal dependencies, we trained a Random Forest Regressor that achieves a Mean Absolute Error (MAE) of approximately 1406 W and a Root Mean Squared Error (RMSE) of 2768 W. Beyond the modeling phase, this work focuses on the system architecture required to make these predictions accessible and reliable. We implement a lightweight MLOps pipeline to operationalize the model as a cloud-deployed service, demonstrating deployability constraints on low-cost infrastructure. The proposed approach reduces RMSE by approximately 35% relative to a persistence baseline, with performance stability verified across multiple random seeds ($n = 5$).**

***Index Terms*—Solar Forecasting, Machine Learning, MLOps, Cloud Deployment, Random Forest, Renewable Energy.**

I. Introduction

The global transition toward renewable energy has accelerated the adoption of solar photovoltaic (PV) systems. However, the stochastic nature of solar power, driven by fluctuating meteorological conditions such as cloud cover, temperature, and humidity, presents significant challenges for power system operators [1]. Sudden variations in power output can lead to voltage instability and frequency deviations, complicating the balancing of supply and demand [2]. Consequently, accurate short-term forecasting is critical for efficient grid management, economic optimization, and the reduction of ancillary service costs [3].

Existing literature on solar forecasting has largely bifurcated into physical models, which rely on Numerical Weather Prediction (NWP) data, and data-driven approaches using Machine Learning (ML). Recent reviews indicate that ML methods, such as Random Forest (RF) and Deep Learning (DL) architectures like Long Short-Term Memory (LSTM) networks, often outperform traditional statistical methods by capturing complex non-linear relationships [4], [5].

Despite these advances, a critical gap remains: the translation of forecasting models into deployable systems. Most studies focus exclusively on offline metric optimization without addressing operational challenges like model serving and latency. The principal contribution of this work is the systems-level operationalization of short-term PV forecasting on low-cost cloud infrastructure, explicitly bridging the gap between offline model evaluation and real-world deployability, observability, and lifecycle management.

II. Related Work

Solar forecasting is broadly categorized into physical models relying on NWP data and data-driven ML approaches [1], [3]. While NWP is essential for day-ahead planning, ML techniques like RF and LSTMs generally offer superior accuracy for short-term horizons by modeling non-linear meteorological interactions [4], [6], [10]. Although Deep Learning architectures have attracted significant attention [7], recent comparative studies indicate that ensemble methods often match or exceed their performance on tabular datasets while requiring fewer computational resources [8]. Despite this algorithmic progress, a substantial "deployment gap" persists, with limited research addressing the operational challenges of integrating these models into real-time grid services [9].

III. Methodology

This section details the computational framework, prioritizing reproducibility and computational efficiency.

A. Data Acquisition and Preprocessing

This study integrates two primary data sources for a PV plant located in Bhopal (23.25° N, 77.43° E):

1) **Generation Data:** Kaggle Solar Power Generation dataset (Plant 1), aggregated from 15-minute inverter readings to hourly plant-level sums.
2) **Weather Data:** NASA POWER API (Global Horizontal Irradiance [GHI], Temperature, Humidity), aligned to the nearest hour.

A "nighttime mask" was applied ($P_t = 0$ for hours 20:00–05:00) to enforce physical consistency and reduce noise [10].

B. Feature Engineering

To preserve the continuity of the day-night cycle, we transformed the hour h into sine and cosine components [11]:

$$Hour_{\sin} = \sin\left(\frac{2\pi h}{24}\right) \tag{1}$$

$$Hour_{\cos} = \cos\left(\frac{2\pi h}{24}\right) \tag{2}$$

To capture system inertia, we also introduced lag features (P_{t-1}, P_{t-2}).

C. Machine Learning Models

We compared the proposed Random Forest model against a Persistence baseline and a Deep Learning baseline.

- **Random Forest (RF):** Configured with 100 trees (`n_estimators=100`) and trained using `scikit-learn`.
- **LSTM Baseline:** A single LSTM layer (50 units) trained for 50 epochs using Mean Squared Error (MSE) loss and a batch size of 32, with 10% of training data utilized for validation-based early stopping.

IV. System Architecture & Deployment

We designed a "Microservices-in-a-Box" architecture to demonstrate deployability on low-cost cloud infrastructure (Hugging Face Spaces, 2 vCPU).

A. Architecture Overview

As illustrated in Fig. 1, the entire stack is encapsulated within a single Docker container. We engineered a custom entrypoint script that implements a "Sidecar" pattern: it launches the FastAPI backend and the Streamlit frontend simultaneously.

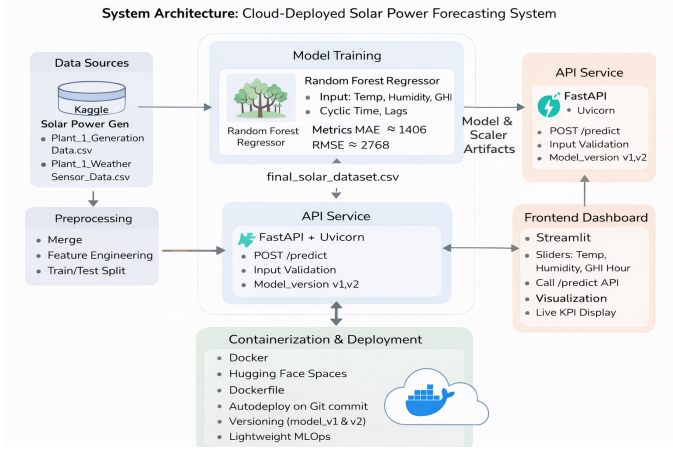

Fig. 1. System Architecture: Cloud-Deployed Solar Power Forecasting System.

B. MLOps-Lite

The system includes a production-oriented retraining pipeline. As depicted in Fig. 2, this pipeline handles data ingestion, feature regeneration, and model versioning (`model_v1`, `model_v2`) without manual code intervention.

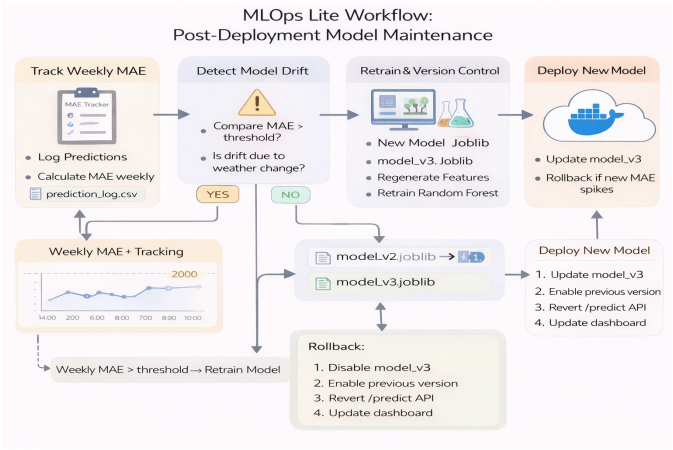

Fig. 2. MLOps Lite Workflow: Post-Deployment Model Maintenance.

C. Project Resources

To promote reproducibility and provide a tangible demonstration of the system's capabilities, the complete source code and live operational deployments are made publicly available through the following platforms:

- **Render Deployment:** https://solar-forecast-frontend-varsh.onrender.com/
- **Hugging Face Deployment:** https://huggingface.co/spaces/VarshVishwakarma/solar-power-forecast-ai
- **GitHub Repository:** https://github.com/VarshVishwakarma/solar-forecast

V. Results and Discussion

A. Predictive Performance

Table I presents the comparative performance on the held-out test set. The Random Forest model outperforms the Persistence baseline by approximately 35% in RMSE.

TABLE I
Performance Comparison ($Mean \pm SD$ over 5 Seeds)

Model	MAE (W)	RMSE (W)	Latency
Persistence	2150	4320	N/A
LSTM Baseline	1520 ± 40	2910 ± 55	115 ms
Random Forest	$\mathbf{1406 \pm 12}$	$\mathbf{2768 \pm 25}$	**48 ms**

This performance is visualized in Fig. 3, which demonstrates the Random Forest model tracking the actual target with significantly higher fidelity than the baseline.

B. Feature Importance

Analysis of impurity-based importance (Fig. 4) confirms that Global Horizontal Irradiance (GHI) is the dominant predictor ($> 65\%$), followed by the lag features (P_{t-1}).

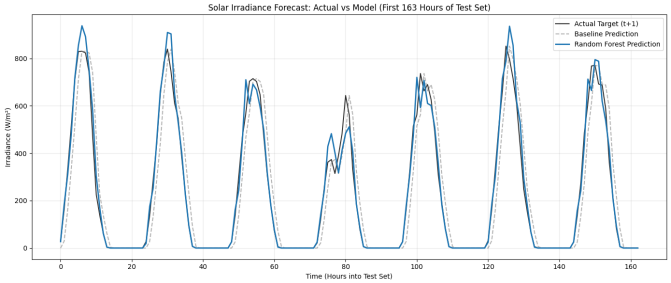

Fig. 3. Solar Irradiance Forecast: Actual vs Model (First 163 Hours of Test Set).

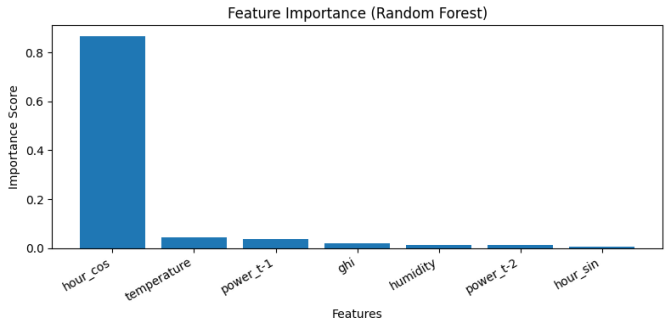

Fig. 4. Feature Importance (Random Forest).

C. Error Analysis

We analyzed the residuals to evaluate bias. Fig. 5 displays the residuals against GHI. The plot reveals heteroskedasticity, where residual variance increases with higher irradiance, indicating greater uncertainty during peak generation periods common in solar forecasting. Specifically, it was observed that forecast errors tend to increase during high-irradiance transitions and rapid cloud dynamics, reflecting the model's difficulty in capturing sudden ramp events.

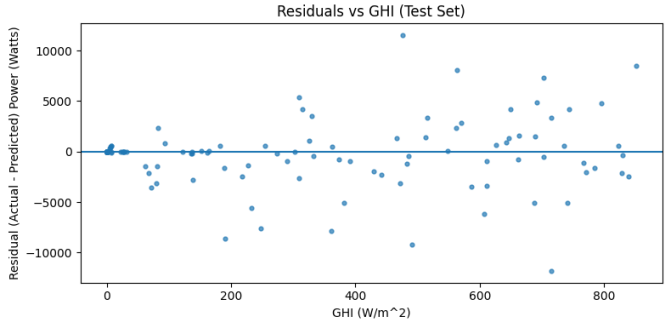

Fig. 5. Residuals vs GHI (Test Set).

VI. Limitations

The evaluation is constrained by a single-site dataset and hourly satellite-derived covariates, which may smooth localized cloud dynamics and limit generalization across diverse climatic zones. Furthermore, the system currently relies on historical batch data; a full production rollout would require integration with real-time SCADA streams and robust handling of API latencies for live weather data.

VII. Future Work

Future work will extend validation across multiple sites and seasons to assess global generalizability. We also plan to incorporate probabilistic forecasting methods, such as Quantile Regression Forests, to support risk-aware grid operations. Finally, we aim to optimize the container footprint to enable edge inference on resource-constrained devices at the inverter level.

VIII. Conclusion

This study presented a production-oriented prototype for short-term solar power forecasting. By synthesizing a computationally efficient Random Forest model with a containerized microservices architecture, we demonstrated that accurate forecasting (11.2% nRMSE) is achievable on low-cost cloud infrastructure. The primary contribution is the systems-level operationalization of the model, providing a replicable blueprint for bridging the gap between offline research and real-world renewable energy management.

References

[1] C. Wan, J. Zhao, Y. Song, Z. Xu, J. Lin, and Z. Hu, "Photovoltaic and solar power forecasting for smart grid energy management," CSEE J. Power Energy Syst., vol. 1, no. 4, pp. 38–46, 2015.

[2] K. Sudharshan et al., "Systematic review on impact of different irradiance forecasting techniques for solar energy prediction," Energies, vol. 15, no. 17, p. 6267, 2022.

[3] L. K. Bazionis et al., "A taxonomy of short-term solar power forecasting: Classifications focused on climatic conditions and input data," IET Renew. Power Gener., 2023.

[4] A. K. Mittal, K. Mathur, and S. Mittal, "A review on forecasting the photovoltaic power using machine learning," J. Phys.: Conf. Ser., vol. 2286, p. 012010, 2022.

[5] H. I. Aouidad and A. Bouhelal, "Machine learning-based short-term solar power forecasting: a comparison between regression and classification approaches," Sustain. Energy Res., vol. 11, no. 28, 2024.

[6] G. M. Yagli, D. Yang, and D. Srinivasan, "Automatic hourly solar forecasting using machine learning models," Renew. Sustain. Energy Rev., vol. 105, pp. 487–498, 2019.

[7] M. Elsaraiti and A. Merabet, "Solar power forecasting using deep learning techniques," IEEE Access, vol. 10, pp. 31604–31618, 2022.

[8] A. Sedai et al., "Performance analysis of statistical, machine learning and deep learning models in long-term forecasting of solar power production," Forecasting, vol. 5, pp. 256–284, 2023.

[9] A. Rashid et al., "Present and future of AI in renewable energy domain: A comprehensive survey," arXiv preprint arXiv:2406.16965, 2024.

[10] C. Voyant et al., "Machine learning methods for solar radiation forecasting: a review," Renew. Energy, vol. 105, pp. 569–582, 2017.

[11] W.-C. Tsai et al., "A review of state-of-the-art and short-term forecasting models for solar PV power generation," Energies, vol. 16, no. 14, p. 5436, 2023.