41st European Photovoltaic Solar Energy Conference and Exhibition

This presentation was selected by the Sc. Committee of the EU PVSEC 2024 for submission of a full paper to one of the EU PVSEC's collaborating peer-reviewed journals.

SKILL-DRIVEN MODEL TRAINING FOR SOLAR FORECASTING WITH SKY IMAGES

Amar Meddahi[1,2], Arttu Tuomiranta[2], and Sebastien Guillon[2]
[1]O.I.E. – Mines Paris – PSL University, 1 rue Claude Daunesse, 06904 Sophia-Antipolis Cedex
[2]TotalEnergies, Le Next 7-9 Boulevard Thomas Gobert, 91120 Palaiseau
Contact: amar.meddahi@minesparis.psl.eu

ABSTRACT: Accurate short-term solar irradiance forecasting is critical for optimizing solar energy integration into power systems. This study presents an image-based deep learning framework for minute-scale solar irradiance prediction. Our model, developed locally, was benchmarked against two commercial forecasting solutions at the same experimental site, demonstrating superior accuracy and adaptability. A key innovation is the introduction of a skill-driven sampling algorithm, based on clear sky index persistence error, which optimizes the training dataset by excluding low-utility samples while preserving essential physical features, such as solar zenith and azimuth angles. This approach enables the removal of up to 30% of the original training data, leading to approximately 16% savings in computational resources without compromising forecast accuracy. Using a test set of 324,991 observations, our model achieved a skill score of 7.63%, significantly outperforming commercial models, which showed negative skill scores under the same conditions.
Keywords: Solar Forecasting, Sky Imager, Deep Learning, Data-centric

1 INTRODUCTION

1.1 Context

Solar energy plays an increasingly important role in the global energy landscape, driven by rapid advancements in photovoltaic (PV) technologies [1]. However, its availability is influenced by weather conditions that alter the interaction of solar radiation with the atmosphere. These variations in atmospheric optical properties—reflection, absorption, and scattering—create challenges for maintaining consistent energy output from large-scale PV installations, complicating their integration into the energy grid.

To mitigate these challenges, PV systems are often coupled with Energy Storage Systems (ESS), which balance fluctuations by storing and releasing energy as needed [2]. Solar forecasting tools have further enhanced the ability to predict surface solar irradiance (SSI) and PV output across various time scales, optimizing system performance and reducing financial penalties due to discrepancies in energy supply [3].

1.2 Background

Deep learning has advanced solar forecasting by integrating data from sources such as pyranometers and sky images from ground-based and satellite systems. For very short-term forecasts—ranging from minutes to hours—fisheye sky imagers provide high-resolution, wide-angle views of the sky, which are valuable for predicting cloud movements and solar irradiance variations, as shown in Figure 1.

Figure 1: Examples of sky images captured using different fisheye cameras

Deep learning models detect patterns between sequential sky observations and corresponding changes in solar irradiance or PV output. Trained on extensive historical data, these models have demonstrated strong predictive performance across various architectures [4]. In short-term solar forecasting using sky images, the goal is to predict future irradiance by leveraging historical sky images and irradiance data. The training process optimizes the model using historical datasets and evaluates its ability to generalize on unseen data, with performance metrics assessing accuracy [5].

1.3 Data-centric vs. Model-centric Approach

Traditionally, solar forecasting research has focused on improving neural network architectures to enhance predictive accuracy. However, a growing shift toward data-centric approaches prioritizes dataset quality over model refinement. Data augmentation and resampling techniques have shown promise in increasing the representativeness of training data and boosting model performance [6-9].

1.4 Problem statement

This research focuses on optimizing the selection of training samples for deep learning models used in intra-hour solar forecasting. Specifically, it aims to identify the most relevant subset of training data that enhances model performance without requiring the full dataset. The challenge lies in developing a method that selects a data subset yielding comparable or better performance than the full dataset, thus improving model training efficiency.

2 METHODOLOGY

2.1 Irradiance data

Global Horizontal Irradiance (GHI) measurements used for model development and validation were collected at an acquisition station in La Tour-de-Salvagny, France (latitude: 45.815, longitude: 4.726). The data were recorded every 10 seconds using a standard class A pyranometer and aggregated into 30-second averages.

2.2 Sky image data

Sky images were captured using a sky imager installed adjacent to the GHI measurement site in La Tour-de-Salvagny, operational since July 2019. The imager includes a visible spectrum camera equipped with a fisheye lens, capturing images at 30-second intervals. An example image from this imager is shown in the central panel of Figure 1.

2.3 Commercial forecasts

10.4229/EUPVSEC2024/4CO.9.2
020389-001

To validate the proposed forecasting model, data from two commercial forecasting solutions were used for comparison. Both systems use a proprietary sky imager with an embedded forecasting model, one operating in the visible spectrum and the other in the near-infrared spectrum. The specific details of these forecasting models are proprietary and not publicly available. Both systems are installed at the same site in La Tour-de-Salvagny as the validation measurements. The visible spectrum system was operational from September 1, 2020, to November 8, 2020, while the infrared system was active from November 15, 2022, to March 6, 2023. Each system provides forecasts with a 5-minute horizon; the visible system updates every 30 seconds, and the infrared system every 60 seconds.

2.3 Data preprocessing

Irradiance data with solar elevation angles below 15 degrees were systematically excluded due to increased uncertainty in clear sky models during these periods, which can lead to significant forecast errors in the clear sky index [10,11]. GHI measurements exceeding established physical limits were also removed [12]. For time series detrending, , the clear sky index (k) was computed using McClear [13].

Sky image preprocessing involved circular cropping at a 10-degree elevation to remove obstructions such as trees and buildings, ensuring consistency with irradiance data. Distortions caused by the camera and fisheye lens were corrected through checkerboard calibration, which facilitates accurate tracking of cloud movements and sizes. Finally, the images were downsampled to a lower resolution (e.g., 64x64 pixels) to prepare them for deep learning model training.

2.4 Skill-driven sampling

The skill-driven sampling algorithm refines the training dataset by excluding samples where the persistence model performs well (i.e., low prediction difficulty). It focuses on selecting samples with higher persistence error, where advanced forecasting models can demonstrate their capabilities.

The algorithm is outlined as follows:

- **Input**: The training dataset D_{train}, forecasting horizon h, and persistence error threshold τ.
- **Output**: A refined training dataset D'_{train}.

Steps:

1. For each sample in D_{train}, calculate the persistence error $\varepsilon_{persistence}(h)$.
2. If the error exceeds the threshold τ, include the sample in the refined dataset.
3. Return the refined dataset D'_{train}.

The persistence error $\varepsilon_{persistence}$ is defined as $\varepsilon_{persistence}(h) = |k_{t+h} - k_t|$, representing the forecasting error of the persistence model based on the clear sky index. The use of the L1 norm provides robustness against outliers and extreme deviations, common in minute-scale irradiance variability.

By estimating the error based on the clear sky index k, rather than the absolute GHI error, the algorithm avoids biases related to varying GHI levels. In lower solar elevation angles, surface irradiance decreases, potentially leading to misleadingly low error values if GHI were used. The clear sky index-based error allows for a more accurate assessment of forecasting challenges based on varying sky conditions, independent of the absolute GHI level. In this study, persistence error serves as a proxy for the predictability of the scenario.

2.5 Model

The proposed model integrates three neural network modules into a unified architecture, inspired by successful benchmarks in previous studies [14-18]. This design facilitates end-to-end learning and supports a multi-modal approach by processing sequences of past sky observations and GHI measurements to predict future GHI levels. Figure 2 provides a general overview of the model, highlighting key components and parameters.

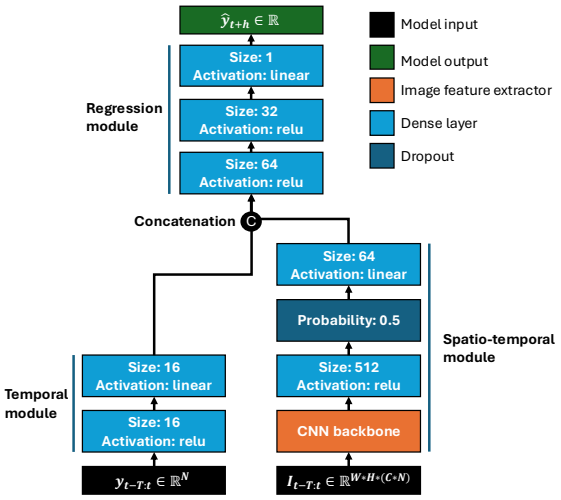

Figure 2: Proposed forecasting deep learning architecture. Each module is depicted with its specific neural operators and associated parameters. The "CNN backbone" refers to the 50-layer ResNet model [19].

3 RESULTS

3.1 Model architecture validation

Objective: The validation of our model architecture serves two main purposes: ensuring the control model performs adequately for subsequent experiments, and directly comparing it with commercial forecasting solutions deployed at the same site. This study is the first to directly compare an on-site developed deep learning model with commercial sky imaging solutions. While previous studies typically validate models against observations or persistence baselines, this work incorporates both commercial imagers and forecasting algorithms, providing insights into the relative performance of locally developed models versus off-site commercial solutions.

Global Performance Assessment: Table I compares our model's performance with two commercial systems (visible and infrared spectrum systems) using key error metrics. Our model consistently outperformed both commercial systems across all metrics:

- In cross-validation, our model achieved an RMSE skill score of 7.63%, a strong result for very short-term forecasts.
- Compared to the visible spectrum solution, our model significantly reduced RMSE and achieved a higher skill score (9.94% vs. -11.20%).

10.4229/EUPVSEC2024/4CO.9.2
020389-002

- Similarly, the infrared solution underperformed, yielding a negative skill score (-28.75%), while our model maintained a positive result (6.11%).

These findings highlight the advantage of developing site-specific models, which are better at capturing local conditions compared to off-site solutions.

Limitations: One limitation is that our model was optimized using site-specific data, allowing it to adapt to local characteristics, such as systematic biases and recurring weather patterns. In contrast, the commercial models were developed off-site, making them less tailored to the specific environment. Furthermore, rapid advancements in deep learning have likely contributed to our model's superior performance, underscoring the importance of incorporating the latest technologies to enhance forecasting accuracy.

Table I: Statistical error comparison between the proposed model and commercial solutions. Metrics labeled with ↓ indicate that lower values are better; metrics labeled with ↑ indicate that higher values are better.

Model	↓ MBE Wm^{-2} (%)	↓ MAE Wm^{-2} (%)	↓ RMSE Wm^{-2} (%)	↑ RMSE Skill Score %
10-fold cross-validation (from 2019-07-09 to 2023-06-01)				
Ours	0.10 (0.02)	38.10 (9.63)	85.84 (21.70)	7.63
Observation Mean: 395.54 Wm^{-2} - Observation Number : 324991				
Visible commercial solution (from 2020-09-01 to 2020-11-08)				
Visible	6.35 (1.99)	49.48 (15.52)	89.96 (28.21)	-11.20
Ours	-1.24 (-0.39)	35.30 (11.07)	72.86 (22.85)	9.94
Observation Mean: 318.89 Wm-2 - Observation Number: 6872				
Infrared commercial solution (from 2022-11-15 to 2023-03-06)				
Infrared	12.92 (5.46)	34.09 (14.40)	71.81 (30.32)	-28.75
Ours	0.21 (0.09)	27.20 (11.48)	52.37 (22.11)	6.11
Observation Mean: 236.84 Wm^{-2} - Observation Number: 7835				

3.2 Skill-driven validation

Objective: The objective of validating the skill-driven sampling approach is to assess whether it improves model performance and computational efficiency. Specifically, we aim to determine if the refined dataset reduces training time while maintaining or enhancing predictive accuracy compared to the full dataset.

Impact on Forecasting Performance: Table II presents the impact of skill-driven sampling on the model's performance. The model was trained on progressively refined datasets using different thresholds τ, while keeping the architecture unchanged. A 10-fold cross-validation was performed to evaluate the model, with the performance metrics averaged.

The results highlight three key trends:

- Aggressive Sampling (30-40% of data retained): High thresholds (τ=0.061 and τ=0.038) result in reduced performance, as seen in higher RMSE and negative skill scores. This suggests that removing too much data reduces the variability needed for model learning, especially in complex scenarios.
- Moderate Sampling (50-60% of data retained): At thresholds τ=0.023 and τ=0.014 model performance improves, with positive skill scores surpassing the persistence baseline. Training time was reduced by approximately 20%, without compromising predictive accuracy.
- Conservative Sampling (70-90% of data retained): At thresholds τ=0.007 and τ=0.002, model performance is nearly identical to the control model trained on the full dataset, while training times were reduced by 8-16%. This suggests that up to 30% of the original dataset is redundant, validating the hypothesis that simpler scenarios offer minimal value for model learning.

Overall, skill-driven sampling achieved up to 16% savings in computational resources without degrading model performance. This efficiency is significant given the increasing computational demands of deep learning models in energy forecasting.

Table II: Deep learning model forecasting performance for different skill-driven sampling levels. The bottom row, representing a control sampling scenario with τ = 0.000 (0%), assesses the model where the dataset is not refined. Each subsequent row evaluates the model with datasets refined using different levels of the skill-driven sampling algorithm. For each τ, a 10-fold validation was performed and the average across the folds is reported. ↓: the lower the better; ↑: the higher the better.

τ (%)	↓ MBE Wm^{-2} (%)	↓ MAE Wm^{-2} (%)	↓ RMSE Wm^{-2} (%)	↑ RMSE Skill Score %	↓ Training Time %
0.061 (30)	13.02 (3.29)	66.35 (16.77)	132.06 (33.39)	-42.11	69.83
0.038 (40)	3.48 (0.88)	50.64 (12.80)	127.34 (32.19)	-37.02	71.29
0.023 (50)	3.04 (0.77)	43.85 (11.09)	89.68 (22.67)	3.49	81.72
0.014 (60)	2.61 (0.66)	41.19 (10.41)	89.24 (22.56)	3.97	81.75
0.007 (70)	3.37 (0.85)	39.51 (9.99)	85.67 (21.66)	7.81	84.08
0.004 (80)	1.13 (0.28)	38.82 (9.81)	86.08 (21.76)	7.37	86.66
0.002 (90)	0.64 (0.16)	38.85 (9.82)	86.87 (21.96)	6.52	92.41
0.000 (100)	0.10 (0.02)	38.10 (9.63)	85.84 (21.70)	7.63	100
Observation Mean: 395.54 Wm^{-2} - Observation Number : 324991					

Perspectives: This study demonstrates that persistence error, based on the clear sky index, serves as a valid proxy for assessing the informativeness of training samples. By applying the optimal threshold (τ = 0.007), we reduced the dataset by 30%, leading to a 16% reduction in computational resources. However, further research is needed to generalize this approach to other datasets, particularly in environments where predictable conditions dominate. Combining this algorithm with data augmentation techniques or extending data collection in these regions may be necessary to avoid overfitting.

4 CONCLUSION

This study introduced an image-based deep learning framework for very short-term solar irradiance forecasting. By benchmarking our model against two

commercial forecasting solutions, we demonstrated its superior accuracy and adaptability to site-specific conditions.

A key innovation of this work is the development of a skill-driven sampling algorithm based on persistence error. This algorithm optimizes the training dataset by excluding low-utility samples that do not significantly contribute to model learning. Importantly, it preserves critical physical attributes, such as solar zenith and azimuth angles, even at high sampling rates.

Our findings show that the proposed sampling strategy allows for the exclusion of up to 30% of the original dataset, leading to approximately 16% savings in computational resources without compromising forecast performance.

5 REFERENCES

[1] International Energy Agency, World Energy Outlook, 2023.
[2] Li, Yaze, and Jingxian Wu. "Optimum integration of solar energy with battery energy storage systems." *IEEE Transactions on Engineering Management* 69.3 (2020): 697-707.
[3] Yang, Dazhi, et al. "History and trends in solar irradiance and PV power forecasting: A preliminary assessment and review using text mining." *Solar Energy* 168 (2018): 60-101.
[4] Paletta, Quentin, et al. "Advances in solar forecasting: Computer vision with deep learning." Advances in Applied Energy (2023): 100150.
[5] Yang, Dazhi, et al. "Verification of deterministic solar forecasts." *Solar Energy* 210 (2020): 20-37.
[6] Nie, Yuhao, Ahmed S. Zamzam, and Adam Brandt. "Resampling and data augmentation for short-term PV output prediction based on an imbalanced sky images dataset using convolutional neural networks." Solar Energy 224 (2021): 341-354.
[7] Paletta, Quentin, et al. "SPIN: Simplifying Polar Invariance for Neural networks Application to vision-based irradiance forecasting." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[8] Fabel, Yann, et al. "Combining Deep Learning and Physical Models: A Benchmark Study on All-Sky Imager-Based Solar Nowcasting Systems." Solar RRL 8.4 (2024): 2300808.
[9] Liu, Ling-Man, et al. "Dual-dimension Time-GGAN data augmentation method for improving the performance of deep learning models for PV power forecasting." Energy Reports 9 (2023): 6419-6433.
[10] Sengupta, Manajit, et al. Best practices handbook for the collection and use of solar resource data for solar energy applications. No. NREL/TP-5D00-77635. National Renewable Energy Lab.(NREL), Golden, CO (United States), 2021.
[11] Yang, Dazhi. "Choice of clear-sky model in solar forecasting." Journal of Renewable and Sustainable Energy 12.2 (2020).
[12] Urraca, Ruben, et al. "Quality control of global solar radiation data with satellite-based products." Solar Energy 158 (2017): 49-62.
[13] Lefèvre, Mireille, et al. "McClear: a new model estimating downwelling solar radiation at ground level in clear-sky conditions." *Atmospheric Measurement Techniques* 6.9 (2013): 2403-2418.
[14] Sun, Yuchi, Vignesh Venugopal, and Adam R. Brandt. "Short-term solar power forecast with deep learning: Exploring optimal input and output configuration." Solar Energy 188 (2019): 730-741.
[15] Feng, Cong, et al. "Convolutional neural networks for intra-hour solar forecasting based on sky image sequences." Applied Energy 310 (2022): 118438.
[16] Zhang, Jinsong, et al. "Deep photovoltaic nowcasting." Solar Energy 176 (2018): 267-276.
[17] Paletta, Quentin, et al. "ECLIPSE: Envisioning cloud induced perturbations in solar energy." Applied Energy 326 (2022): 119924.
[18] Sun, Yuchi, Gergely Szűcs, and Adam R. Brandt. "Solar PV output prediction from video streams using convolutional neural networks." Energy & Environmental Science 11.7 (2018): 1811-1818.
[19] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

10.4229/EUPVSEC2024/4CO.9.2
020389-004