SKYIMAGENET: TOWARDS A LARGE-SCALE SKY IMAGE DATASET FOR SOLAR POWER FORECASTING

Yuhao Nie*
Institute for Data, Systems, and Society
Massachusetts Institute of Technology
Cambridge, MA , USA
nieyh@mit.edu

Quentin Paletta*
ESA Φ-lab
European Space Agency - ESRIN
Frascati, Italy
quentin.paletta@esa.int

Sherrie Wang
Institute for Data, Systems, and Society
Massachusetts Institute of Technology
Cambridge, MA , USA
sherwang@mit.edu

ABSTRACT

The variability of solar photovoltaic (PV) output, particularly that caused by rapidly changing cloud dynamics, challenges the reliability of renewable energy systems. Solar forecasting based on cloud observations collected by ground-level sky cameras shows promising performance in anticipating short-term solar power fluctuations. However, current deep learning methods often rely on a single dataset with limited sample diversity for training, and thus generalize poorly to new locations and different sky conditions. Moreover, the lack of a standardized dataset hinders the consistent comparison of existing solar forecasting methods. To close these gaps, we propose to build a large-scale standardized sky image dataset — SkyImageNet — by assembling, harmonizing, and processing suitable open-source datasets collected in various geographical locations. An accompanying python package will be developed to streamline the process of utilizing SkyImageNet in a machine learning framework. We hope that the outcomes of this project will foster the development of more robust forecasting systems, advance the comparability of short-term solar forecasting model performances, and further facilitate the transition to the next generation of sustainable energy systems.

1 INTRODUCTION

Integrating renewable resources, such as solar photovoltaic (PV), into the electricity grid has been recognized as an important pathway towards a low-carbon energy system. However, large-scale integration of PV is challenged by its fluctuating power output, mainly caused by short-term cloud passage events (Nie et al., 2021). Current electricity systems contain a large amount of dispatchable resources (e.g., coal, natural gas, etc.) that can be ramped to fill in for the variability. In contrast, as future grids transition towards a significant share of PV, the rapid loss of power supply within minutes would pose a substantial challenge for grid management. Anticipating such events, even only 5 to 15 minutes in advance, would allow grid operators to efficiently adapt the response of the grid to incoming power supply fluctuations.

Short-term solar forecasting, defined as predicting either PV power generation or solar irradiance within a time horizon up to 30 minutes, has historically been challenging because of the complex cloud dynamics. Images taken by ground-level sky cameras (see the right column of Figure 1) contain abundant information of the sky and are capable of providing warning of approaching clouds from minutes to an hour ahead (Yang et al., 2018), making it increasingly popular in short-term solar forecasting (Paletta et al., 2023).

*Equal contribution

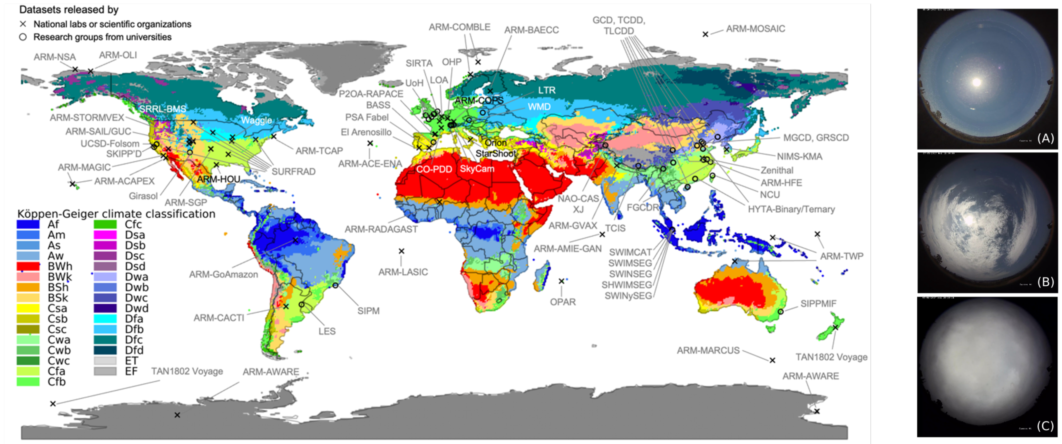

Figure 1: On the left, the geographic locations of 72 open-source sky image datasets annotated on the Köppen–Geiger climate map. Adapted from (Nie et al., 2024). On the right, sky images taken by a fish-eye camera under different sky conditions (a) clear sky, (b) partly cloudy, and (c) overcast.

In the recent 5 years, the use of deep learning (DL) models, such as CNNs (Sun et al., 2019; Venugopal et al., 2019; Feng & Zhang, 2020; Paletta & Lasenby, 2020a; Feng et al., 2022) and RNNs (Zhang et al., 2018; Paletta et al., 2021; 2022b) has seen a significant rise in image-based solar power modelling. These deep learning models achieve superior performance over traditional physics-based models (Chow et al., 2011; Marquez & Coimbra, 2013; Quesada-Ruiz et al., 2014) and machine learning models coupled with hand-crafted feature engineering (Fu & Cheng, 2013; Chu et al., 2013; 2015a;b; Pedro et al., 2019).

However, most existing DL-based solar forecasting models trained on datasets with limited spatial and temporal coverage, struggle to generalize effectively to new locations. In addition, these DL-based methods show poor anticipation skills under cloudy conditions, for which solar power generation exhibits higher levels of variability, partly due to the lack of diversified cloudy samples for model training (Paletta et al., 2021). The data limitation also hinders the further exploration of more advanced and promising deep learning approaches for solar forecasting, such as foundation models (Bommasani et al., 2021), which are trained on large-scale datasets and can be fine-tuned to a range of downstream tasks or to new locations (Paletta et al., 2024) with excellent generalization skills. Moreover, the lack of a standardized dataset hinders the consistent comparison of various existing solar forecasting methods (Nie et al., 2022a).

2 Related Work

The increasing release of sky image datasets in recent years has provided great opportunities to address these limitations. In a recent work by Nie et al. (2024), 72 open-source ground-based sky image datasets collected globally for research on solar forecasting and cloud modeling have been identified (see the left column of Figure 1). Utilizing such open-source datasets would save significant efforts in terms of in-situ data collection, which is expensive and time consuming especially when devices need to be deployed in multiple locations for multiple years to ensure broad spatial and temporal data coverage. Hence merging suitable open-source datasets to build one large and diversified dataset would benefit comparable and robust model development, which is of utmost importance in the solar forecasting community (Yang, 2019).

3 Objectives

In this project, we propose to:

1. Build a large-scale standardized ground-based sky image dataset — SkyImageNet, by assembling, harmonizing and processing suitable open-source datasets collected in diverse climate conditions. This comprehensive dataset would be valuable for short-term solar forecasting as well as other related areas such as cloud modeling.

2. Implement and compare various existing solar forecasting methods based on the SkyImageNet dataset. This would provide a comprehensive performance benchmark of the existing methods, enabling researchers to consistently evaluate and improve their models.
3. Develop a python package with pre-implemented functions for the dataset download, data processing, pre-trained model loading, and forecasting performance evaluation. This package would streamline the process of utilizing the SkyImageNet and accelerate the method development of solar forecasting.

4 Proposed methodology

This study follows the common procedures for machine learning pipeline development, which cover the following four stages: (1) data collection, (2) data processing, (3) model development and evaluation, and (4) deployment. The methodology adopted for each stage are described separately below:

Data collection The datasets suitable for short-term solar forecasting would be selected from the 72 open-source sky image datasets identified in previous study (Nie et al., 2024), based on attributes including label type (solar irradiance or PV power output), temporal resolution, and image pixel resolution. The raw data would be collected and stored locally, while the meta information for each dataset (e.g., the geographic locations of cameras/sensors, the camera model, the camera orientation, the time stamps, etc.) and the processed data will be centralized and made publicly available.

Data processing The main challenge of this project is to deal with the heterogeneity of the data, as different datasets have different data characteristics, e.g., image pixel resolution, temporal resolution and label categories (solar irradiance or PV power output). Specifically, the various high resolution of sky images will be down-sampled to the same lower resolution to facilitate model development. Furthermore, to standardize images taken by cameras from sensors with different celestial orientations, a clear transformation pipeline will be developed. For different data labels, proper normalization techniques will be proposed to handle the associated scale and unit diversity (Nie et al., 2022b; Paletta et al., 2024). After processing, valid samples formatted as $\{input, output\}$ will be formed based on the specific setup of the solar forecasting task of inerest (e.g., forecasting horizon, sampling interval, history length, etc.) (Paletta et al., 2023). The resulting data will then be split into model training/validation and testing subsets, where a selection of data points representing diverse weather conditions and locations will be isolated to constitute a fixed test set.

Model development and evaluation A typical set up for a short-term solar forecasting task is to predict the T-minute-ahead ($T \leq 30$) future solar irradiance or PV power output based on a sky image sequence and possibly together withauxiliary data such as sun angles, wind speed/direction, irradiance value and PV measurement as model input. In this project, diverse types of existing solar forecasting models would be implemented, such as statistical (Reikard, 2009), machine learning (Fu & Cheng, 2013; Chu et al., 2013), deep learning (Sun et al., 2019; Paletta et al., 2022b; Feng et al., 2022), or physics-based models (Marquez & Coimbra, 2013) to construct a performance benchmark. Following this, the training of more advanced large-scale deep learning models on the SkyImageNet, i.e., foundation models (Bommasani et al., 2021), would be explored.

Deployment The resulting processed dataset will be uploaded to public repositories (e.g., Zenodo, Mendeley data, Hugging Face) and a permanent URL will be generated to simplify its access. A python package will be developed to include pre-implemented dataset download functions, data pre-processing pipelines, typical image transform and augmentation functions (Paletta & Lasenby, 2020b; Julian & Sankaranarayanan, 2021; Paletta et al., 2022a; Nie et al., 2021; Terrén-Serrano & Martínez-Ramón, 2022), pre-trained benchmark models and forecasting performance evaluation metrics. This will enable researchers to accelerate their ML-based forecasting model development.

5 Pathway to impact

The SkyImageNet project aims at advancing the development of ML-based forecasting tools to facilitate the integration of a large share of solar power into the energy mix. Predicting the future energy yield of this intermittent source of energy at different time scales would indeed benefit diverse activities including energy trading, energy dispatch, frequency setting, hybrid power plant optimisation, smart grids, and storage management (Law et al., 2016b; Carriere & Kariniotakis, 2019). For PV systems, a recent study estimated that a solar forecasting improvement as small as 0.1% relative to the reference model could save about 5500 tonnes of CO_2 induced by gas spinning-reserves (Dixon et al., 2022) in the UK. It was also shown that short-term forecasts contribute substantially to both a higher economic profitability and a lower outage rate (Law et al., 2016a), thereby enabling an increased adoption rate of this low-carbon technology around the world.

The project expects to provide easy access to multi-location multi-year sky imagery data and other atmospheric observations, benefiting a wide community including grid operators, energy producers, solar forecasting companies and broad academic research areas, including, for example, energy meteorology, atmospheric and climate sciences. To ensure a long-term impact, SkyImageNet will be well documented and will result in several publications. In addition, guidelines to contribute to the work via new datasets, code for relevant functionalities, or model additions will be included. We thereby build the foundation for a dataset and code base that can be used and extended by the whole community and can result in larger versions of SkyImangeNet with more added data points from other research groups.

ACKNOWLEDGMENTS

The authors would also like to thank the support of the Michael Hammer Postdoctoral Fellowship from Institute for Data, Systems, and Society (IDSS) at Massachusetts Institute of Technology (MIT), ESA Φ-Lab and the Climate Office of the European Space Agency.

REFERENCES

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Thomas Carriere and George Kariniotakis. An Integrated Approach for Value-Oriented Energy Forecasting and Data-Driven Decision-Making Application to Renewable Energy Trading. *IEEE Transactions on Smart Grid*, 10(6):6933–6944, November 2019. ISSN 1949-3061. doi: 10.1109/TSG.2019.2914379.

Chi Wai Chow, Bryan Urquhart, Matthew Lave, Anthony Dominguez, Jan Kleissl, Janet Shields, and Byron Washom. Intra-hour forecasting with a total sky imager at the uc san diego solar energy testbed. *Solar Energy*, 85(11):2881–2893, 2011.

Yinghao Chu, Hugo T. C. Pedro, and Carlos F. M. Coimbra. Hybrid intra-hour DNI forecasts with sky image processing enhanced by stochastic learning. *Solar Energy*, 98:592–603, December 2013. ISSN 0038-092X. doi: 10.1016/j.solener.2013.10.020.

Yinghao Chu, Mengying Li, Hugo T C Pedro, and Carlos F M Coimbra. Real-time prediction intervals for intra-hour DNI forecasts. *Renewable Energy*, 83:234–244, 2015a. ISSN 18790682. doi: 10.1016/j.renene.2015.04.022. URL `http://dx.doi.org/10.1016/j.renene.2015.04.022`.

Yinghao Chu, Bryan Urquhart, Seyyed M.I. Gohari, Hugo T.C. Pedro, Jan Kleissl, and Carlos F.M. Coimbra. Short-term reforecasting of power output from a 48 MWe solar PV plant. *Solar Energy*, 112:68–77, feb 2015b. ISSN 0038-092X. doi: 10.1016/J.SOLENER.2014.11.017. URL `https://www.sciencedirect.com/science/article/pii/S0038092X14005611`.

Ben Dixon, María Pérez-Ortiz, and Jacob Bieker. Comparing the carbon costs and benefits of low-resource solar nowcasting, October 2022.

Cong Feng and Jie Zhang. SolarNet: A sky image-based deep convolutional neural network for intra-hour solar forecasting. *Solar Energy*, 204(April):71–78, 2020. doi: 10.1016/j.solener.2020.03.083. URL `https://doi.org/10.1016/j.solener.2020.03.083`.

Cong Feng, Jie Zhang, Wenqi Zhang, and Bri Mathias Hodge. Convolutional neural networks for intra-hour solar forecasting based on sky image sequences. *Applied Energy*, 310:118438, mar 2022. doi: 10.1016/J.APENERGY.2021.118438.

Chia-Lin Fu and Hsu-Yung Cheng. Predicting solar irradiance with all-sky image features via regression. *Solar Energy*, 97:537–550, 2013.

Leron Julian and Aswin C. Sankaranarayanan. Precise Forecasting of Sky Images Using Spatial Warping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1136–1144, 2021.

Edward W. Law, Merlinde Kay, and Robert A. Taylor. Evaluating the benefits of using short-term direct normal irradiance forecasts to operate a concentrated solar thermal plant. *Solar Energy*, 140:93–108, 2016a. ISSN 0038-092X. doi: https://doi.org/10.1016/j.solener.2016.10.037. URL https://www.sciencedirect.com/science/article/pii/S0038092X16305023.

Edward W. Law, Merlinde Kay, and Robert A. Taylor. Evaluating the benefits of using short-term direct normal irradiance forecasts to operate a concentrated solar thermal plant. *Solar Energy*, 140:93–108, December 2016b. ISSN 0038-092X. doi: 10.1016/j.solener.2016.10.037.

Ricardo Marquez and Carlos FM Coimbra. Intra-hour dni forecasting based on cloud tracking image analysis. *Solar Energy*, 91:327–336, 2013.

Yuhao Nie, 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(May):341–354, 2021. doi: 10.1016/j.solener.2021.05.095. URL https://doi.org/10.1016/j.solener.2021.05.095.

Yuhao Nie, Xiatong Li, Andea Scott, Yuchi Sun, Vignesh Venugopal, and Adam Brandt. SKIPP'D: a SKy Images and Photovoltaic Power Generation Dataset for Short-term Solar Forecasting. *arXiv preprint arXiv:2207.00913*, 2022a.

Yuhao Nie, Quentin Paletta, Andea Scott, Luis Martin Pomares, Guillaume Arbod, Sgouris Sgouridis, Joan Lasenby, and Adam Brandt. Sky-image-based solar forecasting using deep learning with multi-location data: training models locally, globally or via transfer learning?, 2022b. URL https://arxiv.org/abs/2211.02108.

Yuhao Nie, Xiatong Li, Quentin Paletta, Max Aragon, Andea Scott, and Adam Brandt. Open-source sky image datasets for solar forecasting with deep learning: A comprehensive survey. *Renewable and Sustainable Energy Reviews*, 189:113977, January 2024. ISSN 1364-0321. doi: 10.1016/j.rser.2023.113977.

Quentin Paletta and Joan Lasenby. Convolutional Neural Networks Applied to Sky Images for Short-Term Solar Irradiance Forecasting. In *EU PVSEC*, pp. 1834 – 1837, 2020a. ISBN 3-936338-73-6. doi: 10.4229/EUPVSEC20202020-6BV.5.15. URL https://www.eupvsec-proceedings.com/proceedings?paper=49346.

Quentin Paletta and Joan Lasenby. A temporally consistent image-based sun tracking algorithm for solar energy forecasting applications. In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*, pp. 10, 2020b. URL https://www.climatechange.ai/papers/neurips2020/8.

Quentin Paletta, Guillaume Arbod, and Joan Lasenby. Benchmarking of deep learning irradiance forecasting models from sky images – An in-depth analysis. *Solar Energy*, 224:855–867, August 2021. ISSN 0038-092X. doi: 10.1016/j.solener.2021.05.056. URL https://linkinghub.elsevier.com/retrieve/pii/S0038092X21004266.

Quentin Paletta, Anthony Hu, Guillaume Arbod, Philippe Blanc, and Joan Lasenby. SPIN: Simplifying Polar Invariance for Neural networks Application to vision-based irradiance forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 5182–5191, 2022a. URL https://openaccess.thecvf.com/content/CVPR2022W/OmniCV/html/Paletta_SPIN_Simplifying_Polar_Invariance_for_Neural_Networks_Application_to_Vision-Based_CVPRW_2022_paper.html.

Quentin Paletta, Anthony Hu, Guillaume Arbod, and Joan Lasenby. ECLIPSE: Envisioning CLoud Induced Perturbations in Solar Energy. *Applied Energy*, 326:119924, November 2022b. ISSN 0306-2619. doi: 10.1016/j.apenergy.2022.119924.

Quentin Paletta, Guillermo Terrén-Serrano, Yuhao Nie, Binghui Li, Jacob Bieker, Wenqi Zhang, Laurent Dubus, Soumyabrata Dev, and Cong Feng. Advances in solar forecasting: Computer vision with deep learning. *Advances in Applied Energy*, 11:100150, September 2023. ISSN 2666-7924. doi: 10.1016/j.adapen.2023.100150.

Quentin Paletta, Yuhao Nie, Yves-Marie Saint-Drenan, and Bertrand Le Saux. Improving cross-site generalizability of vision-based solar forecasting models with physics-informed transfer learning. *under review*, 2024.

Hugo T. C. Pedro, Carlos F. M. Coimbra, and Philippe Lauret. Adaptive image features for intra-hour solar forecasts. *Journal of Renewable and Sustainable Energy*, 11(3):036101, May 2019. doi: 10.1063/1.5091952.

S. Quesada-Ruiz, Y. Chu, J. Tovar-Pescador, H. T. C. Pedro, and C. F. M. Coimbra. Cloud-tracking methodology for intra-hour DNI forecasting. *Solar Energy*, 102:267–275, April 2014. ISSN 0038-092X. doi: 10.1016/j.solener.2014.01.030.

Gordon Reikard. Predicting solar radiation at high resolutions: A comparison of time series forecasts. *Solar energy*, 83(3):342–349, 2009.

Yuchi Sun, Vignesh Venugopal, and Adam R Brandt. Short-term solar power forecast with deep learning: Exploring optimal input and output configuration. *Solar Energy*, 188:730–741, aug 2019. doi: 10.1016/j.solener.2019.06.041. URL `https://linkinghub.elsevier.com/retrieve/pii/S0038092X19306164`.

Guillermo Terrén-Serrano and Manel Martínez-Ramón. Geospatial perspective reprojections for ground-based sky imaging systems. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–7, 2022.

Vignesh Venugopal, Yuchi Sun, and Adam R. Brandt. Short-term solar PV forecasting using computer vision: The search for optimal CNN architectures for incorporating sky images and PV generation history. *Journal of Renewable and Sustainable Energy*, 11(6):066102, nov 2019. ISSN 1941-7012. doi: 10.1063/1.5122796. URL `http://aip.scitation.org/doi/10.1063/1.5122796`.

Dazhi Yang. A guideline to solar forecasting research practice: Reproducible, operational, probabilistic or physically-based, ensemble, and skill (ROPES). *Journal of Renewable and Sustainable Energy*, 11(2), 2019. ISSN 19417012. doi: 10.1063/1.5087462. URL `http://dx.doi.org/10.1063/1.5087462`.

Dazhi Yang, Jan Kleissl, Christian A Gueymard, Hugo TC Pedro, and Carlos FM Coimbra. History and trends in solar irradiance and pv power forecasting: A preliminary assessment and review using text mining. *Solar Energy*, 168:60–101, 2018.

Jinsong Zhang, Rodrigo Verschae, Shohei Nobuhara, and Jean François Lalonde. Deep photovoltaic nowcasting. *Solar Energy*, 176(September):267–276, 2018. doi: 10.1016/j.solener.2018.10.024. URL `https://doi.org/10.1016/j.solener.2018.10.024`.