Scenario-Based Evaluation of Probabilistic Time Series Forecasting for Solar Energy

Ruohan Li[1], Yiqun Xie[1*], Xiaowei Jia[2], Gengchen Mai[3], Sophia Hou[1], Zhihao Wang[1], Zhili Li[1]
[1]University of Maryland, [2]University of Pittsburgh, [3]University of Texas at Austin
{r526li,xie,zhwang1,lizhili}@umd.edu,xiaowei@pitt.edu,gengchen.mai@austin.utexas.edu,sophiahou0805@gmail.com

Abstract

Probabilistic time-series forecasting plays a vital role in decision-making under uncertainty, especially in applications like solar energy, where forecast reliability directly impacts energy planning and grid stability. While recent models have improved in generating predictive distributions rather than single-point estimates, existing evaluations often focus on average performance and overlook how model quality varies across different real-world scenarios. In solar energy monitoring, for example, the difficulty of forecasting can change significantly due to atmospheric variability, sensor types, and climate conditions. This work addresses the need for scenario-aware evaluation of probabilistic models by benchmarking state-of-the-art forecasting methods using SolarCube–a large-scale solar radiation dataset spanning diverse regions, cloud regimes, and environmental conditions. We define structured "easy" and "hard" cases across four scenarios and examine how different probabilistic model families (e.g., diffusion, VAE, flow-based) capture uncertainty under these conditions. Our goal is to move beyond overall metrics and reveal how model reliability changes across scenarios that are critical for downstream applications.

CCS Concepts

• **Applied computing** → *Environmental sciences*; • **Computing methodologies** → **Neural networks**.

Keywords

Probabilistic, Time Series Forecasting, Diffusion, Solar Energy

ACM Reference Format:
Ruohan Li[1], Yiqun Xie[1*], Xiaowei Jia[2], Gengchen Mai[3], Sophia Hou[1], Zhihao Wang[1], Zhili Li[1]. 2025. Scenario-Based Evaluation of Probabilistic Time Series Forecasting for Solar Energy. In *The 33rd ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '25), November 3–6, 2025, Minneapolis, MN, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3748636.3762777

1 Introduction

While recent advances in deep generative models—such as diffusion-based, variational autoencoder (VAE)-based, and flow-based approaches—have significantly improved probabilistic forecasting capabilities, current evaluation practices remain limited. Most studies report overall metrics (e.g., RMSE, CRPS, log-likelihood) averaged across test samples, without considering how model performance varies across different conditions. Understanding scenario-based uncertainty is critical for real-world applications, as the value of uncertainty estimates is most evident in challenging situations where errors are likely to be high. In other words, whether an estimated confidence internal (e.g., 95%) can cover the ground truth is often much more important when the prediction errors tend to be larger, and less so if the prediction is already very close to the truth. Yet, few studies evaluate how predictive distributions behave under such diverse conditions.

In this study, we use large-scale solar energy estimation as an example, where (1) probabilistic forecasting is important from the application perspective and (2) the uncertainty is highly dependent on the scenarios. Application-wise, accurate probabilistic forecasts support a range of operational and planning decisions, from real-time grid balancing to day-ahead energy market bidding. For short-term forecasting, a recent multistage stochastic bidding study showed that the use of probabilistic forecasts for intraday markets increased revenue by 22% and halved the imbalance [9]. In day-ahead forecasting, improvements in probabilistic accuracy have been estimated to yield economic benefits of $5–10/MWh by reducing imbalance costs and improving efficiency [10].

Moreover, solar forecasting spans diverse conditions with varying difficulty. Solar irradiance is relatively easy to predict under clear-sky scenarios, but becomes significantly more difficult under dynamic atmospheric variability, such as fast-moving clouds. Horizons create additional scenarios. Short-term (nowcasting) depends on recent dynamics, whereas day-ahead benefits from regular diurnal patterns yet can expose autoregressive models to error accumulation and reduced long-range skill. As a result, overall performance metrics—averaged across all cases—can mask poor model performance in exactly the situations where uncertainty quantification is most needed. As accurate and well-calibrated uncertainty estimates during such challenging conditions are far more valuable than performance under ideal circumstances, evaluating models across distinct scenarios is essential for understanding and improving their reliability in operational decision-making.

We propose a scenario-based evaluation framework, and we demonstrate the importance of using the large-scale benchmark dataset SolarCube for solar energy forecasting tasks. This dataset spans multiple continents and incorporates data from various sources, including satellite observations and in-situ radiation measurements. Our contributions are summarized as follows:

- We present a scenario-based evaluation framework to assess probabilistic forecasting models using the SolarCube benchmarking dataset for various solar energy forecasting tasks and scenarios.
- We define structured "easy" and "hard" forecasting scenarios based on atmospheric variability, regional or domain shifts, and

[*]Corresponding author: Yiqun Xie.

SIGSPATIAL '25, Minneapolis, MN, USA

ACM ISBN 979-8-4007-2086-4/2025/11
https://doi.org/10.1145/3748636.3762777

climate zone consistency, spanning both short-term and long-term forecasting tasks.

- We benchmark various families of probabilistic models under these scenarios and analyze how their uncertainty estimates respond to varying conditions.
- We show that traditional performance metrics can obscure substantial variability in model behavior, highlighting the need for scenario-aware model development and validation in fair and risk-sensitive applications.

2 Related Work

Probabilistic Time Series Forecasting. Deep learning has advanced probabilistic forecasting by addressing limitations of classical models. Loss-based methods estimate parametric distributions through maximum likelihood training but are constrained by fixed distributional assumptions. To improve flexibility, VAE-based models introduce latent variables to capture complex uncertainty [5]. Normalizing flow-based approaches further enhance expressiveness by learning invertible transformations conditioned on historical context [7]. Diffusion-based models offer an alternative generative framework, modeling uncertainty through iterative denoising processes [8, 6]. These categories progressively address the trade-offs between distributional flexibility, temporal dependency modeling, and sampling efficiency in multivariate time series forecasting.

SolarCube Dataset. SolarCube [3] is a benchmark dataset for solar forecasting, designed to support both short-term and long-term prediction tasks. It provides temporally aligned ground-measured solar radiation, cloud masks, and three bands of geostationary satellite observations at a 15-minute resolution. The dataset spans 19 study regions globally and includes evaluation metrics tailored specifically for solar forecasting. We further update SolarCube by incorporating hourly meteorological variables from ERA5, including cloud cover, temperature, and wind components (u and v). These additions improve the dataset's utility for extended-horizon forecasting and enable climate-aware evaluation scenarios.

3 Scenario-Based Evaluation

To comprehensively evaluate the robustness and reliability of probabilistic solar forecasting models, we design scenario-based evaluations in both day-ahead and 3-hour-ahead forecasting settings. All scenarios are categorized into *Easy* and *Hard* cases to reflect different levels of prediction difficulty.

3.1 Day-Ahead Forecasting

For day-ahead forecasting, we followed the setting of point-based long-term task in [3] with the incorporation of four additional meteorological variables from reanalysis data. We evaluate performance under three different scenarios:

Inter-Day Variability Scenarios. These scenarios follow the variability-based scenario design proposed in [2]. The *Easy* scenario corresponds to days with minimal changes in average solar radiation compared to the previous day, whereas the *Hard* scenario involves substantial changes in solar radiation from one day to the next.

Climate Zone Transferability Scenarios. SolarCube covers sites across diverse global climate zones. We design the evaluation scenarios using Köppen climate zones [1], with a focus on how consistent or mismatched the training and testing climates are. We test the model on two sites located in the *Cfa* zone. The *Easy* scenario uses training data from the same *Cfa* zone, while the *Hard* scenario uses training data from *Dfa* and *Dfb* zones, representing a climate mismatch.

Cross-Region Generalization Scenarios. SolarCube spans multiple continents, in-situ networks, and satellite sensors, introducing variability that impacts model generalization. To test cross-regional transfer, we test on two U.S. sites. The *Easy* scenario trains on other U.S. sites with matching sensors and ground measurement networks, while the *Hard* scenario trains on East Asia data, introducing both regional and sensor-domain shifts.

3.2 3-Hour-Ahead Forecasting

For 3-hour-ahead forecasting, we followed the experimental setup of the point-based short-term task in [3], but aggregated 15-minute data into hourly intervals. We evaluate its performance under the following scenarios:

Cloud-Based Scenarios. In short-term solar forecasting, cloud variability is key. We define scenarios based on cloud mask changes: *Easy*—no change in cloud mask, indicating stable atmospheric conditions; *Hard*—cloud mask changes, indicating dynamic conditions.

3.3 Models

Informer_MLE (I_MLE):. A variant of the Informer architecture [11] trained using maximum likelihood estimation (MLE) to produce probabilistic forecasts via distributional outputs instead of point predictions.

TLAE:. A probabilistic forecasting model that learns a latent representation of multivariate time series and generates future trajectories via a variational autoencoder framework [5].

CSDI:. A non-autoregressive denoising diffusion model adapted for probabilistic forecasting, generating future samples by reversing a noise process conditioned on observed historical context [8].

TimeGrad: An autoregressive denoising diffusion model for probabilistic time series forecasting, where a recurrent neural network (RNN) encodes the historical context into a hidden state. The diffusion model then conditions on this state to sequentially generate future time steps [6].

TempFlow: An autoregressive probabilistic model that combines an RNN with conditional normalizing flows. At each future time step, an RNN encodes the context and previously generated outputs to condition a flow-based transformation that generates the next prediction [7].

3.4 Evaluation Metrics

Following [3], we adopt the coefficient of determination (R^2) and relative root mean squared error (rRMSE) to evaluate deterministic forecasts, accounting for the strong diurnal and seasonal variability in solar radiation. To assess the quality of probabilistic forecasts, we use the Continuous Ranked Probability Score (CRPS) [4], which measures the difference between the predicted cumulative distribution function (CDF) and the empirical CDF of the observation. Since I_MLE exhibits significantly high uncertainty in the evening—when the true value is certainly zero—we exclude evening hours when calculating CRPS for all models.

4 Experiment Results

4.1 Day-ahead Forecasting Results

Table 1 shows the model performance in the inter-day variability scenarios. Models generally perform well under easy conditions, capturing the temporal dynamics with high confidence and narrow uncertainty bands as shown in Figure 1. Under hard conditions, performance drops across most models due to increased irradiance variability. All models showed significant drops in CRPS as well. Comparatively, diffusion-based models, particularly TimeGrad and CSDI, demonstrate relatively more stable performance and better-calibrated uncertainty intervals. Flow-based and VAE-based methods show wider prediction intervals and greater variance. Additionally, while all methods can represent uncertainty, diffusion-based models show more desired uncertainty changes moving from easy to hard scenarios (i.e., enlarged intervals), and the other methods tend to have a similar-sized buffer around mean predictions.

For the climate zone transferability scenarios, there is no significant change in overall performance between easy and hard cases for most models, as shown in Table 2 and Figure 2. TLAE is the only model that exhibits notable performance degradation under climate shifts, potentially due to its sensitivity to domain-specific latent representations. The cross-region generalization scenarios show greater variation in model performance in Table 3. While TempFlow and TimeGrad maintain consistent rRMSE between easy and hard cases, their CRPS increases, indicating degraded uncertainty calibration. CSDI and I_MLE maintain consistency in both rRMSE and CRPS. TLAE shows larger performance degradation under hard conditions.

Overall, for day-ahead forecasting, CSDI is the most stable model, maintaining higher accuracy and better-calibrated uncertainty estimates across easy and hard cases under different scenarios. I_MLE often produces much wider prediction intervals compared to other models, showing the instability of uncertainty representation with only the MLE loss. TLAE shows more forecasting quality degradation from easy to hard, but its uncertainty interval remains similarly small and does not well reflect the deviation from true values.

Table 1: Performance metrics for inter-day variability

Model	Easy			Hard		
	R^2	rRMSE	CRPS	R^2	rRMSE	CRPS
CSDI	**0.78**	**30.3**	**57.8**	0.43	58.4	100.6
TLAE	0.51	45.5	113.8	0.03	77.3	155.5
I_MLE	0.74	33.3	100.4	0.28	65.8	152.1
TimeGrad	0.70	35.4	74.5	**0.46**	**57.1**	**97.8**
TempFlow	0.75	32.5	63.8	0.37	61.2	106.6

4.2 3-Hour-Ahead Forecasting Results

In the 3-hour-ahead forecasting scenarios with cloud-based variability, CSDI consistently outperforms all other models, achieving the lowest rRMSE and CRPS in both easy and hard cases, as shown in Table 4. Unlike previous scenarios, TLAE exhibits competitive accuracy but suffers from significantly higher CRPS, indicating unreliable uncertainty estimates. Figure 4 presents an example, where red shading highlights time steps categorized as hard due to changes in the cloud conditions relative to the last input step. In this instance, a sharp drop in solar radiation caused by cloud presence deviates from the previous increasing trend. Among the models,

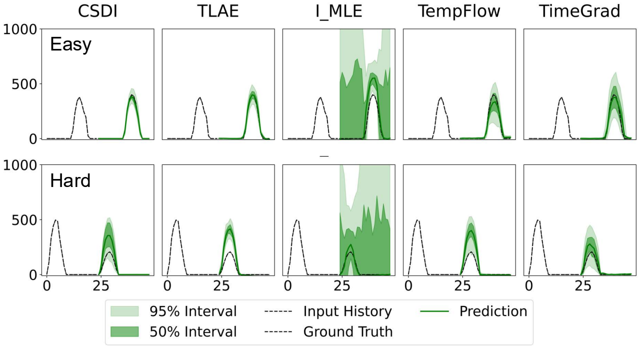

Figure 1: Validation results for easy and hard cases under the inter-day variability scenarios.

Table 2: Performance metrics for climate zone transferability

Model	Easy			Hard		
	R^2	rRMSE	CRPS	R^2	rRMSE	CRPS
CSDI	**0.63**	44.6	**83.3**	0.63	44.8	85.2
TLAE	0.35	59.4	139.7	0.11	69.4	174.0
I_MLE	0.59	46.8	123.1	0.58	47.6	123.3
TimeGrad	**0.63**	**44.5**	84.3	**0.64**	**44.0**	**84.0**
TempFlow	0.60	46.3	90.3	0.62	45.3	87.2

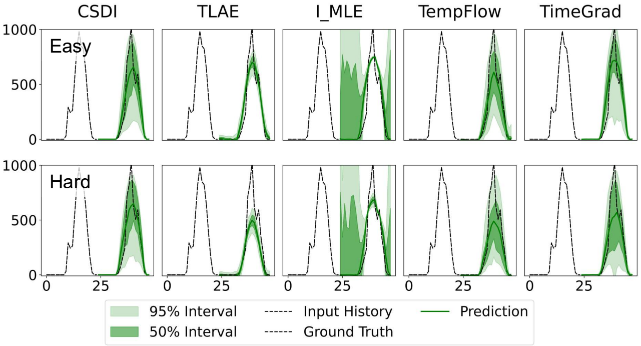

Figure 2: Validation results for easy and hard cases under the climate zone transferability scenarios.

Table 3: Performance metrics for cross-region generalization

Model	Easy			Hard		
	R^2	rRMSE	CRPS	R^2	rRMSE	CRPS
CSDI	**0.70**	**37.2**	**65.2**	**0.70**	**37.4**	**67.3**
TLAE	0.32	56.3	137.8	0.19	61.6	157.7
I_MLE	0.64	40.9	110.0	0.64	41.2	113.3
TempFlow	0.69	38.1	68.4	0.64	40.7	77.8
TimeGrad	0.70	37.7	68.5	0.67	38.9	74.2

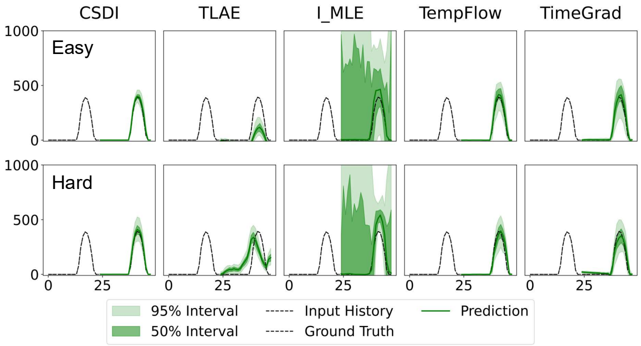

Figure 3: Validation results for easy and hard cases under the cross-region generalization scenarios.

TempFlow and CSDI were able to capture the transition in this example, and TempFlow shows tighter coverage with the intervals (CSDI shows better numbers averaging over all examples). TLAE does not reflect the drop and yields narrow prediction intervals that do not cover the ground truth. In contrast, I_MLE and TimeGrad do not capture the trend either but produce wider intervals that contain the true values.

Table 4: Performance metrics for 3-hour forecast

Model	Easy			Hard		
	R^2	rRMSE	CRPS	R^2	rRMSE	CRPS
CSDI	**0.82**	**26.2**	**50.4**	**0.73**	**20.2**	**55.3**
TLAE	0.79	27.7	109.4	0.66	22.8	113.5
I_MLE	0.62	37.6	83.2	0.43	29.5	89.2
TempFlow	0.71	32.7	72.5	0.70	21.5	61.4
TimeGrad	0.74	30.9	77.5	0.64	23.5	85.8

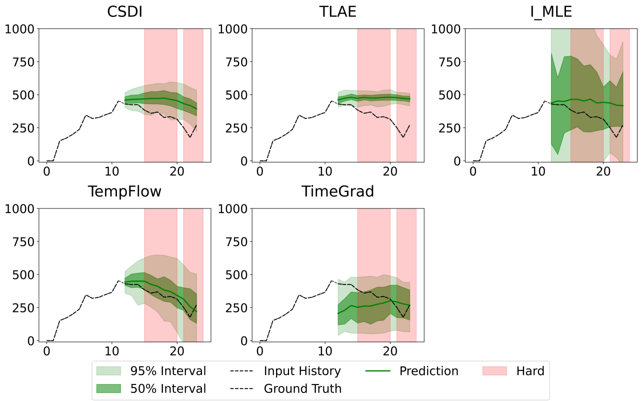

Figure 4: Validation results for easy and hard cases under the cloud-based scenarios.

5 Conclusion

This study presented a scenario-based evaluation framework for probabilistic solar forecasting, emphasizing the need to move beyond average metrics to assess model reliability under diverse and realistic conditions. Our results showed that while many models perform well under stable atmospheric settings and generalize across regions and sensor modalities, they often struggle to provide accurate uncertainty estimates under highly dynamic atmospheric variability—in both short- and long-term forecasts. Among the models, CSDI provides comparatively more robust performances across the scenarios, while models like TempGrad and TempFlow offer varied strengths over different scenarios. Overall, current probablistic forecasters still require improvements to represent uncertainty when it matters most over the challenging cases and be more reflective of the potential prediction errors. We recommend scenario-based evaluations to better understand and address these gaps, helping develop more robust and context-aware forecasting systems. In future work, we will extend the analysis to additional benchmark datasets for broader coverage and leverage the evaluation to develop new models to enhance uncertainty representation.

Acknowledgments

Ruohan Li, Yiqun Xie, Zhihao Wang, and Zhili Li are supported in part by the US NSF under Grant No. 2126474, 2147195, 2425844, and 2530610; NASA under grant 80NSSC25K0013 and 80NSSC25K7221; Google's AI for Social Good Impact Scholars program; NSF NCAR's Derecho HPC system; and the Zaratan cluster at the University of Maryland; Xiaowei Jia is supported in part by the US NSF under Grant No. 2239175, 2147195, 2425845, 2530609; NASA under Grant No. 80NSSC21K0314, and 80NSSC24K1061; and Pitt Momentum Funds and CRC at the University of Pittsburgh.

References

[1] Hylke E Beck et al. 2023. High-resolution (1 km) köppen-geiger maps for 1901–2099 based on constrained cmip6 projections. *Scientific data*, 10, 1, 724.

[2] Oussama Boussif, Ghait Boukachab, Dan Assouline, Stefano Massaroli, Tianle Yuan, Loubna Benabbou, and Yoshua Bengio. 2024. Improving* day-ahead* solar irradiance time series forecasting by leveraging spatio-temporal context. *Advances in Neural Information Processing Systems*, 36.

[3] Ruohan Li, Yiqun Xie, Xiaowei Jia, Dongdong Wang, Yanhua Li, Yingxue Zhang, Zhihao Wang, and Zhili Li. 2024. Solarcube: an integrative benchmark dataset harnessing satellite and in-situ observations for large-scale solar energy forecasting. *Advances in Neural Information Processing Systems*, 37, 3499–3513.

[4] James E Matheson and Robert L Winkler. 1976. Scoring rules for continuous probability distributions. *Management science*, 22, 10, 1087–1096.

[5] Nam Nguyen and Brian Quanz. 2021. Temporal latent auto-encoder: a method for probabilistic multivariate time series forecasting. In *Proceedings of the AAAI conference on artificial intelligence* number 10. Vol. 35, 9117–9125.

[6] Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. 2021. Autoregressive Denoising Diffusion Models for Multivariate Probabilistic Time Series Forecasting. In *Proceedings of the 38th International Conference on Machine Learning* (Proceedings of Machine Learning Research). Marina Meila and Tong Zhang, (Eds.) Vol. 139. PMLR, (18–24 Jul 2021), 8857–8868. http://proceedings.mlr.press/v139/rasul21a.html.

[7] Kashif Rasul, Abdul-Saboor Sheikh, Ingmar Schuster, Urs Bergmann, and Roland Vollgraf. 2021. Multivariate Probabilistic Time Series Forecasting via Conditioned Normalizing Flows. In *International Conference on Learning Representations 2021*. https://openreview.net/forum?id=WiGQBFuVRv.

[8] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. 2021. Csdi: conditional score-based diffusion models for probabilistic time series imputation. *Advances in neural information processing systems*, 34, 24804–24816.

[9] LR Visser, TA AlSkaif, Adil Khurram, Jan Kleissl, and WGHJM van Sark. 2024. Probabilistic solar power forecasting: an economic and technical evaluation of an optimal market bidding strategy. *Applied Energy*, 370, 123573.

[10] Jie Zhang, Bri-Mathias Hodge, Anthony Florita, Siyuan Lu, Hendrik F Hamann, and Venkat Banunarayanan. 2013. Metrics for evaluating the accuracy of solar power forecasting. Tech. rep. National Renewable Energy Lab.(NREL), Golden, CO (United States).

[11] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence* number 12. Vol. 35, 11106–11115.