Task-aware Latent-Space Coupling for Generative Solar Forecasting

Summary

Cloud passages cause rapid irradiance fluctuations (ramp events) in solar power plants, making accurate short-term forecasting challenging. Ground-based all-sky imagers (ASI) combined with irradiance data enable high-resolution cloud monitoring. Machine learning outperforms physical models, but MSE-optimized forecasts often over-smooth and miss variability. Recent generative video prediction models improve ramp detection by learning cloud dynamics, but separating video generation from irradiance estimation introduces domain shift, biases, poor calibration, and artifacts. The proposed FarSky framework addresses this by coupling a diffusion-based video predictor (FAR) with a multi-task autoencoder in a shared latent space, jointly learning image reconstruction and irradiance decoding. This preserves radiative information, eliminates synthetic image decoding, and reduces domain shift. Evaluations on a ASI benchmark dataset show significant gains in deterministic/probabilistic forecasts and more balanced ramp event prediction across lead times.

Keywords: All-sky imagers, Generative forecasting, Ramp events, Latent-space coupling

Field of application for the findings: Solar irradiance forecasting, Ramp-rate control, Grid integration and stability, Optimization of power plant operations, Energy storage and hybrid renewable systems

Addressed audience: Researchers in solar energy and machine learning, Meteorologists and solar forecasters, Power plant operators and owners, Grid and transmission system operators, Project developers in renewables, Energy market traders and participants

Introduction

The aim of this work is to develop a generative intra-hour solar forecasting framework that addresses key limitations of existing data-driven, all-sky imager-based approaches. Previous studies have demonstrated that separating the learning of cloud dynamics from irradiance prediction can mitigate the smoothing effects inherent to direct end-to-end regression optimization on irradiance errors (Nie et al. 2024). Hence, this improves the representation of short-term variability and enhances the detection of ramp events. Generative video prediction models constitute a suitable approach in this context, as they explicitly learn the spatio-temporal evolution of clouds in image space. In addition, generative models are intrinsically probabilistic and enable the generation of multiple future trajectories, providing a direct mechanism for uncertainty estimation. Among recent advances, diffusion-based models have emerged as the state-of-the-art technique for generative image and video modeling, offering highly realistic outputs and reduced blurring effects, particularly for video prediction tasks. A recent architecture is the frame-autoregressive (FAR) model (Gu et al. 2025), which has demonstrated superior performance and faster convergence compared to other diffusion-based video prediction approaches.

While prior work has typically decoupled video prediction and irradiance estimation by applying a separate downstream model to synthetically generated images, this work proposes a latent-coupled generative forecasting framework that integrates irradiance-relevant information directly into the generative pipeline. As diffusion-based video prediction models often operate in latent space, context image sequences must first be encoded by an autoencoder. To this end, a multi-task autoencoder is developed that jointly learns image reconstruction and irradiance decoding, thereby embedding information on irradiance into the latent representations. The implementation of the multi-task autoencoder is based on an extended Deep Compression Auto-Encoder (DCAE) (Chen et al. 2024) by incorporating an additional multi-layer perceptron to decode GHI from latent representations of sky images. Following the training of the multi-task autoencoder, the FAR model is trained on the shared latent representations, preserving irradiance-relevant features throughout the generative process. The combination of both models forms the novel framework, FarSky, as shown in Fig. 1. During inference, FarSky encodes context images first into latent space, generates future latent trajectories, and decodes both irradiance and optionally sky images directly from the generated latents.

The proposed framework is evaluated on diverse all-sky imager datasets and benchmarked against previous generative forecasting pipelines and state-of-the-art reference models. In addition to standard deterministic error metrics, a comprehensive probabilistic evaluation is conducted, and the ability to capture short-term variability is assessed using dedicated ramp event metrics.

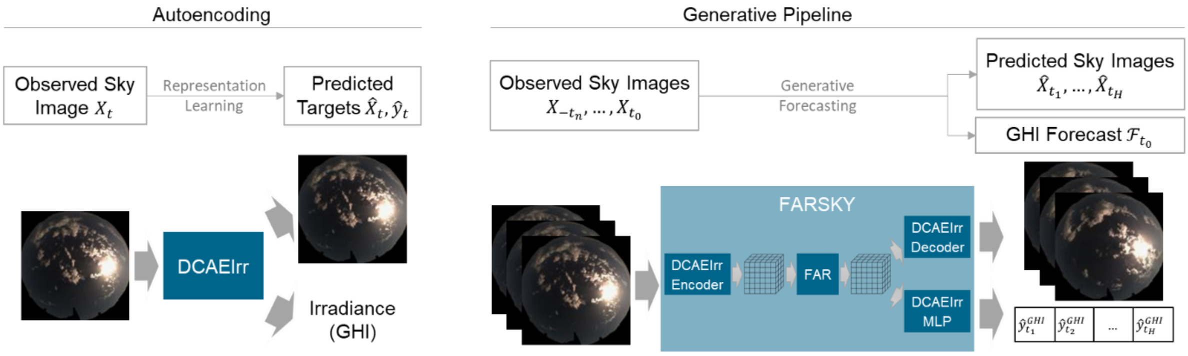

Fig. 1: Graphical representation of the proposed generative intra-hour solar forecasting framework FarSky.

Scientific Novelty

Previous generative approaches for ASI-based solar forecasting have demonstrated improved capability in capturing short-term variability and detecting ramp events. However, they often struggled to achieve competitive performance with respect to standard deterministic metrics such as forecast skill when compared to direct irradiance regression models. This trade-off has limited the practical applicability of generative forecasting pipelines so far.

FarSky presents a novel generative forecasting framework that overcomes these limitations by integrating task-aware latent-space coupling into the generative pipeline. Through the multi-task autoencoder, irradiance forecasts can be derived directly from latent space, eliminating the need for a separate downstream irradiance estimation model. This avoids the intermediate decoding of latent representations back into image space and therefore reduces error accumulation. Moreover, this design effectively mitigates the domain shift between real and generated images that has been identified as a major source of systematic error in previous generative forecasting pipelines relying on downstream models trained exclusively on real image data.

Beyond mitigating domain shift, the joint representation learning strategy improves the robustness of irradiance estimation compared to regressors trained solely on the irradiance prediction task. By constraining the latent space through both visual and radiative objectives, FarSky reduces systematic biases and exhibits improved generalization when applied across different datasets.

Overall, FarSky represents a significant methodological advancement in solar forecasting. The proposed framework achieves improvements in standard deterministic metrics such as forecast skill while simultaneously preserving the ability to represent short-term variability and ramp events. Moreover, by leveraging a generative formulation, the approach remains intrinsically probabilistic, providing uncertainty estimates that are essential for reliable intra-hour solar forecasting.

Results and conclusion

This section presents first results of the FarSky framework. All shown results are based on a separate test dataset, first introduced in an ASI Forecasting benchmark (Logothetis et al. 2019), that was not used for training.

Tab. 1 shows a direct comparison between our multi-task autoencoder and a CNN-based irradiance regressor which estimates GHI from sky imagery. Note that this evaluation compares solar estimation performance based on observed sky images only. Although both models are trained on the same training dataset, showing similar convergence, the results on the benchmark dataset differ significantly. The CNN regressor shows pronounced negative bias, which consequently effect the other metrics. In contrast, our multi-task autoencoder exhibit much smaller bias and much lower RMSE and MAE. This indicates an improved robustness when optimizing simultaneously on image reconstruction and irradiance estimation compared to a single-task model.

Tab. 1: Error metrics on irradiance estimation between mutli-task autoencoder and CNN regressor on real sky images.

Model	RMSE [W/m²]	MAE [W/m²]	MBE [W/m²]
Irradiance CNN (ResNet34)	52.0	37.5	-24.8
Multi-task Autoencoder (DCAE)	25.9	14.8	-5.7

Next, the video prediction model is trained with latent representations encoded by the multi-task autoencoder. After training the video prediction model, the full generative pipeline is executed on the benchmark dataset, creating minute-resolution forecasts with a horizon of 30 minutes. In this analysis, for each forecast issue, four ensemble members are generated to obtain probabilistic forecasts. As reference model, a decoupled generative pipeline is implemented which separates video prediction and downstream irradiance estimation. It is based on the same video prediction architecture FAR, but using the default DCAE for image encoding without latent-space coupling. Training configuration is kept the same for comparability. To obtain irradiance forecasts, the presented CNN irradiance regressor is subsequently applied on the decoded sky images generated by FAR.

Ramp events are analyzed based on changes in the clear-sky index within the forecast following the definition presented in (Nouri et al. 2024). Analogous to the deterministic evaluation, ramp events are analyzed for each lead time. A tolerance window centered around the predicted value for a respective lead time is considered to detect changes in the clear-sky index that surpass a predefined threshold.

Beyond the direct comparison of the two generative forecasting approaches, we additionally consider a non-generative deep learning solar forecasting model (Fabel et al. 2023). This model is based on state-of-the-art transformer architectures, trained end-to-end on an irradiance forecasting objective for forecast horizon of 20 minutes and is therefore referred to here as **E2E-TF.**

Fig. 2 shows an overview of deterministic, probabilistic and ramp event evaluation. The plots depict three forecast scores as a function of lead time for the three models, higher values indicating better performance. The scores are: forecast skill (FS) based on RMSE with smart persistence as baseline; the continuous rank probability skill score (CRPSS) with CSD-Clim (Salle et al. 2021) as baseline; and the F1-Score of ramp event classification performance. FarSky exhibits the strongest FS, with the largest gains over the decoupled FAR approach and superior performance to E2E-TF in particular at short lead times. While CRPSS differences between FarSky and E2E-TF are smaller, FarSky still attains the highest overall scores. The most pronounced gap occurs for F1-scores: E2E-TF remains near zero across all lead times, indicating an inability to anticipate ramp events, whereas the generative models exceed F1-scores of 0.5 up to 25 minutes.

The results demonstrate that the generative approach itself is the key factor to anticipate short-term variability, representing a significant improvement over state-of-the-art. However, only the latent-coupling approach also achieves superior performance in deterministic and probabilistic evaluations. In the full paper publication, more detailed evaluations comparing the proposed framework with generative and non-generative state-of-the-art solar forecasting models will be conducted, discussing strengths and limitations of the respective modeling paradigms.

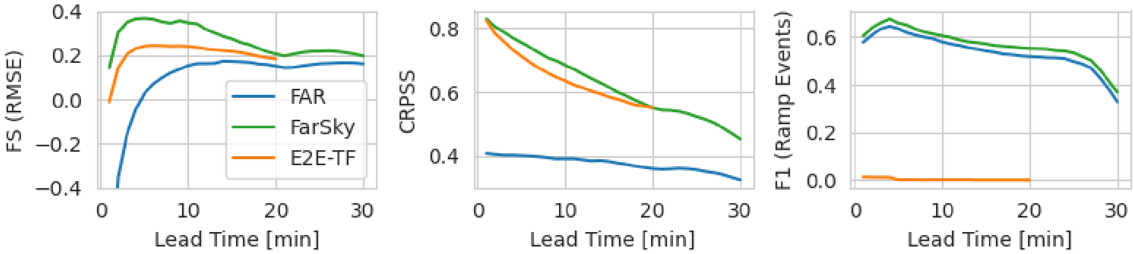

Fig. 2: Performance metrics comparing three different solar forecasting approaches.

References

- Chen, J., Cai, H., Chen, J., Xie, E., Yang, S., Tang, H., ... & Han, S. 2024. Deep compression autoencoder for efficient high-resolution diffusion models. arXiv preprint arXiv:2410.10733.
- Fabel, Y., Nouri, B., Wilbert, S., Blum, N., Schnaus, D., Triebel, R., ... & Pitz-Paal, R. 2024. Combining deep learning and physical models: a benchmark study on All-Sky Imager-Based solar nowcasting systems. Solar RRL, 8(4), 2300808.
- Gu, Y., Mao, W., & Shou, M. Z. 2025. Long-context autoregressive video modeling with next-frame prediction. arXiv preprint arXiv:2503.19325.
- La Salle, J. L. G., David, M., & Lauret, P. 2021. A new climatology reference model to benchmark probabilistic solar forecasts. Solar Energy, 223, 398-414.
- Nie, Y., Zelikman, E., Scott, A., Paletta, Q., & Brandt, A. 2024. Skygpt: Probabilistic ultra-short-term solar forecasting using synthetic sky images from physics-constrained videogpt. Advances in Applied Energy, 14, 100172.
- Nouri, B., Fabel, Y., Blum, N., Schnaus, D., Zarzalejo, L. F., Kazantzidis, A., & Wilbert, S. 2024. Ramp rate metric suitable for solar forecasting. Solar RRL, 8(24), 2400468.