TASK-AWARE LATENT-SPACE COUPLING FOR GENERATIVE SOLAR FORECASTING

Yann Fabel, Bijan Nouri, Milon Miah, Niklas Blum, Luis F. Zarzalejo, Julia Kowalski, Robert Pitz-Paal

EuroSun 2026

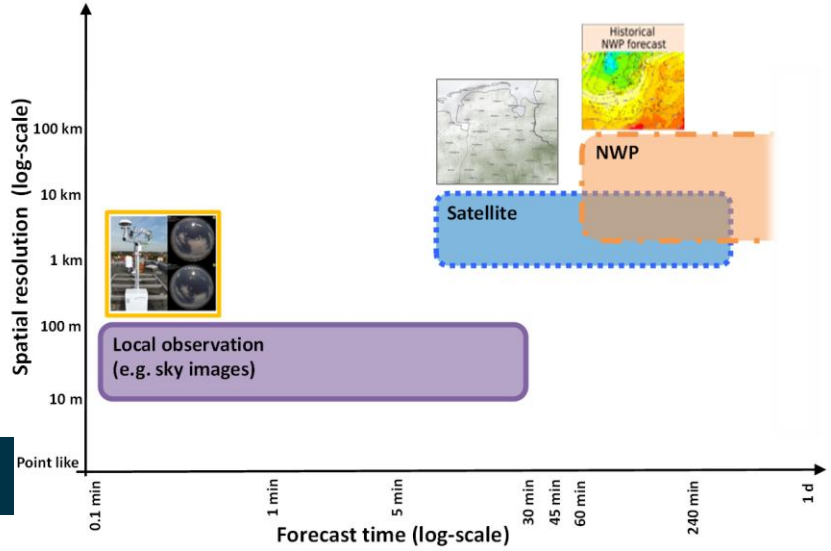
17/09/2026, Freiburg

DLR

Challenges of Intra-hour Solar Forecasting

- **Motivation**
  - Solar irradiance exhibits strong minute-scale variability due to local cloud dynamics
  - Rapid irradiance changes (ramps) can lead to sudden drops of PV power
    → Potential violation of grid codes
  - Intra-hour forecasting can increase awareness, improving
    - Grid balancing & reserves
    - Storage dispatch
    - Market decisions
- **Challenges**
  - NWP and satellite-based models often still too coarse
    → Local observations with sky imagery
  - Cloud evolution is highly nonlinear and chaotic, limiting predictability
    → Data-driven approaches
  - Deterministic models without uncertainty representation
    → Generative models
  - Error-minimizing objectives favor smooth, averaged forecasts
    → Modelling of cloud dynamics in image space





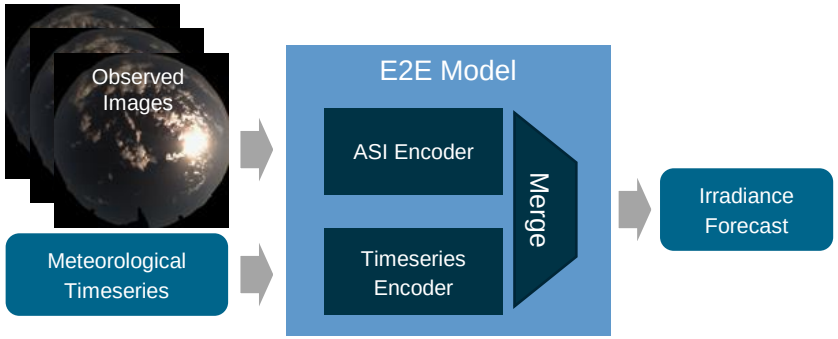
Goal: Multi-step GHI forecast up to 30 minutes ahead in 1 minute resolution

Overview of ASI-based Forecasting Approaches

End-to-End Forecasting

Learns mapping of input sequences to output via forecast error minimization

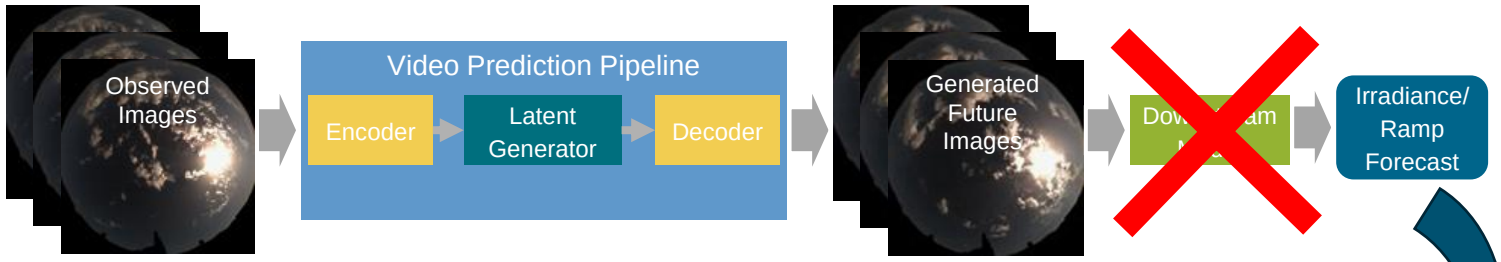
- ✗ Probabilistic (via post-processing)
- ✗ Ramp events
- ✓ High forecast skill



Decoupled Generative Forecasting

Separation of cloud modelling and irradiance prediction via independent VP and irradiance models

- ✓ Probabilistic
- ✓ Ramp events
- ✗ High forecast skill

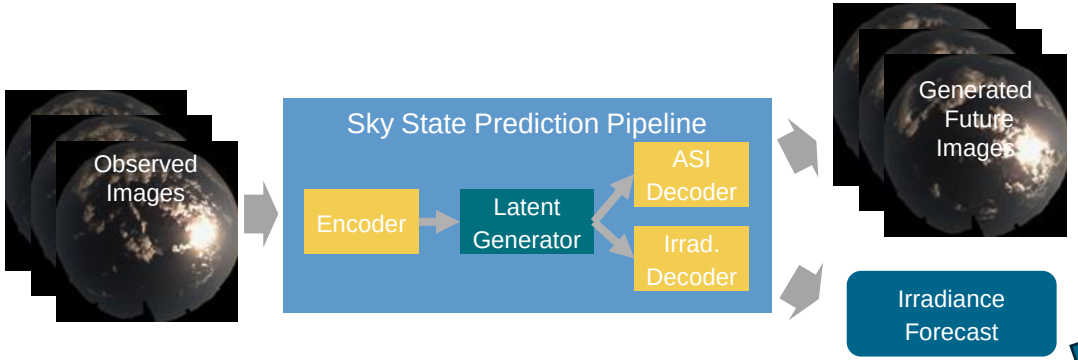


New Approach

Latent-Coupling Generative Forecasting

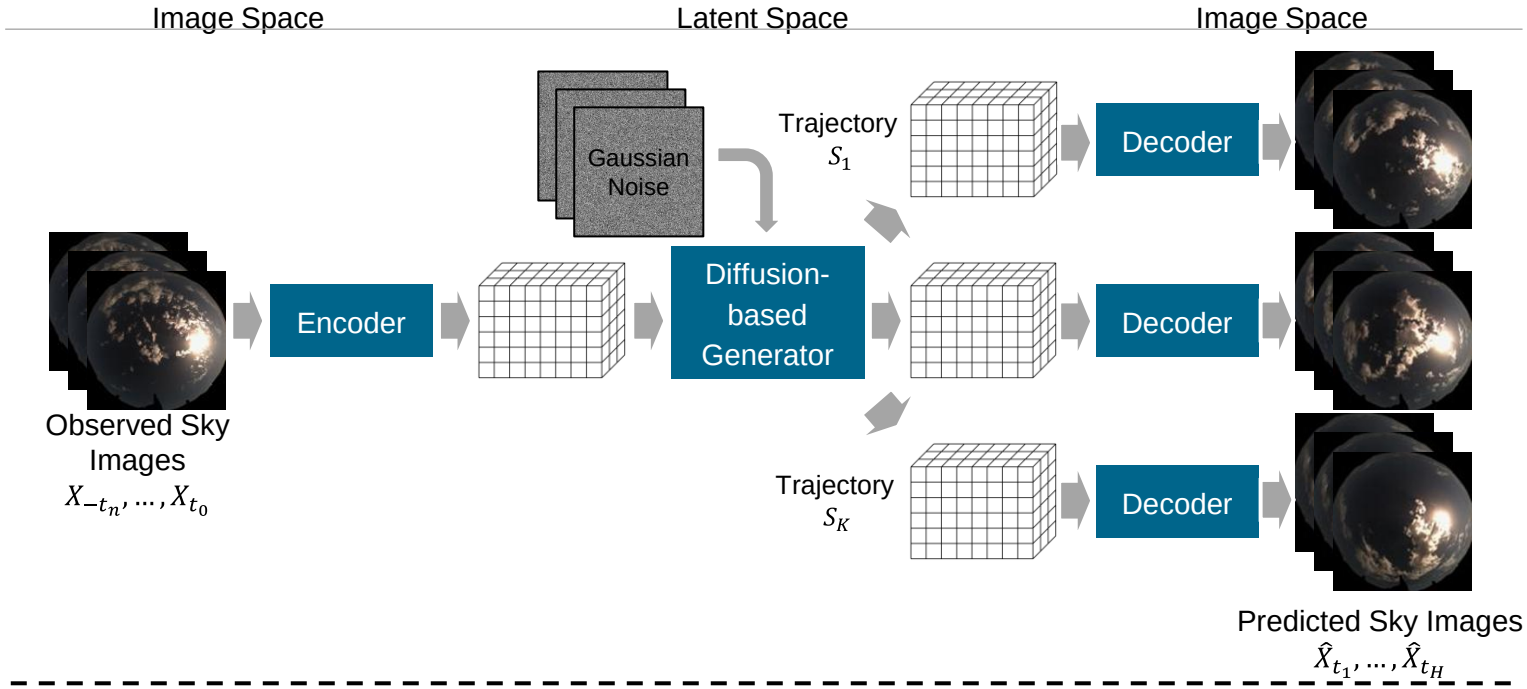
Integrate irradiance representations into ASI autoencoding

- ✓ Probabilistic
- ✓ Ramp events
- ? High forecast skill

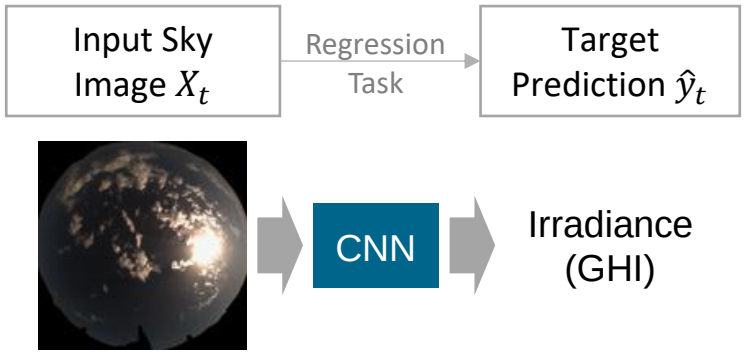


Decoupled Generative Solar Forecasting

- **Model 1: Video Prediction (VP)**
  - Image latents learned via autoencoding:
    Deep Compression AutoEncoder (DCAE [9])
  - Diffusion-based video prediction in latent space:
    Frame AutoRegressive architecture (FAR [1])
  - Multiple trajectories by sampling from gaussian noise



- **Model 2: Irradiance Estimation**
  - Predict GHI for a given sky image using a Convolutional Neural Network (CNN)
  - Trained separately on real ASI data
  - Applied on synthetic ASI data from VP model in inference



→ Domain shift between real and synthetic images
→ Error propagation from VP decoder
→ Strong bias in irradiance predictions

Latent-Coupling Generative Forecasting – FarSky

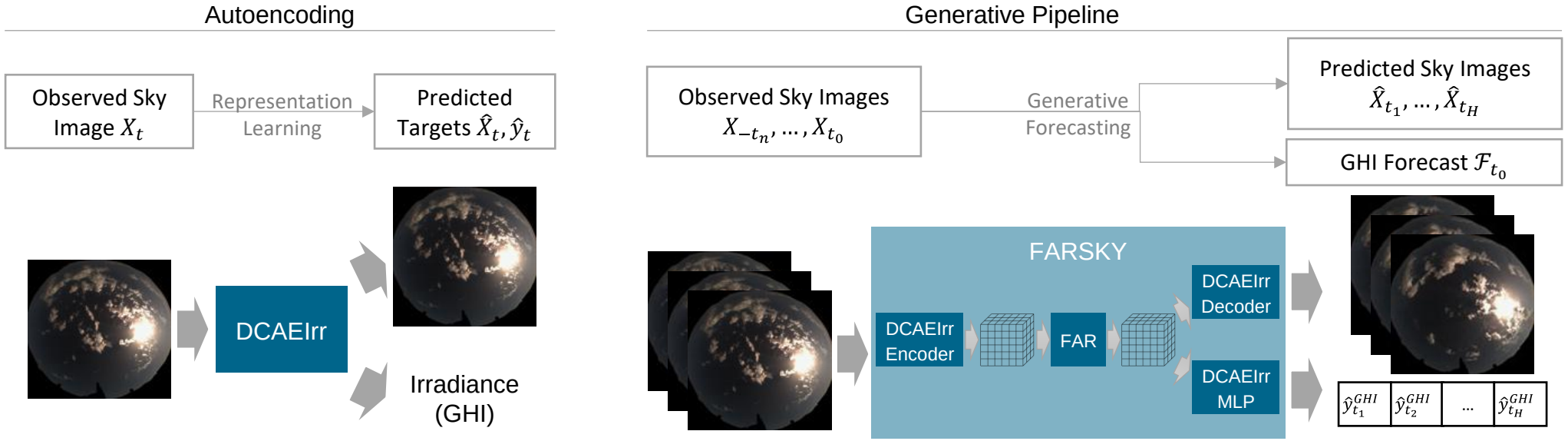
FarSky Framework

- Integration of irradiance representations in autoencoding by additional irradiance decoder
- Same generative diffusion pipeline (FAR)
- Irradiance information is maintained in latent space throughout generative pipeline
- Irradiance forecast is obtained directly from latent space via irradiance decoder

→ No error propagation from image decoding

→ No domain shift problems

→ Increased robustness and generalization due to combined optimization on image and irradiance errors



Datasets for Model Development and Testing

Measurement Site



- Plataforma Solar de Almería (PSA)[1], Spain
  - Cold desert climate (BWk)
  - Challenging cloud conditions due to four surrounding mountain ranges
- Four meteo stations each providing
  - 1-min All-Sky Imagery (fixed exposure of 160µs)
  - 1-min averages of solar irradiance (GHI, DNI, DHI)

Data Selection

- Excluded days dominated by clear-sky conditions
- Full calendar days used as indivisible samples for splitting
- No overlap of days between training, validation, and test sets

Datasets

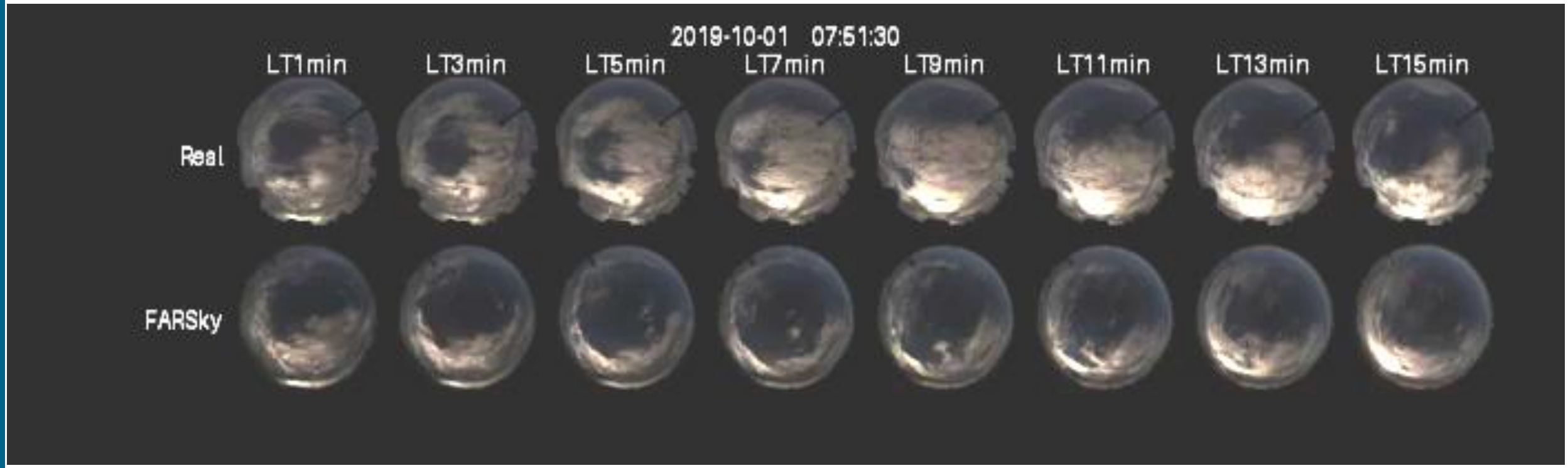
Purpose	Data Source	Station(s)	Period	Days
Training	Multi-station PSA data	HP, KONTAS, METAS, PVOT	2019-2024	1219
Validation	SolarVision Almería [7] [2]	PVOT	2023	12
Test 1	ASI Nowcasting Benchmark 2019 [2] [3]	METAS	2019	28
Test 2	SolarVision Almería [7] [2]	PVOT	2022-2024	60

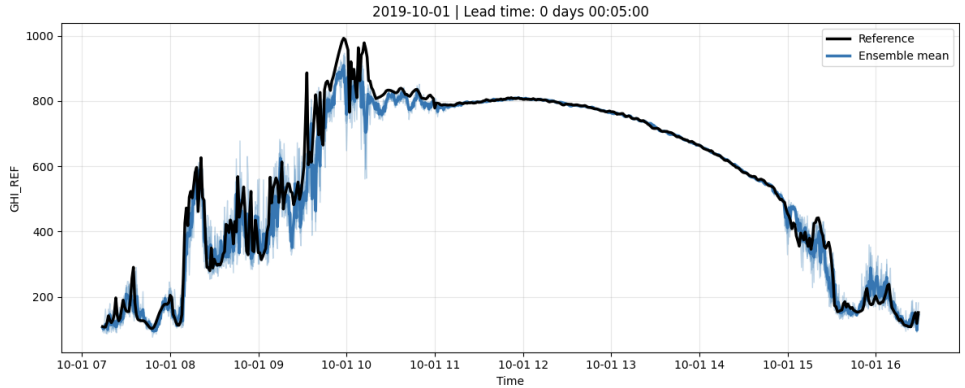
Results presented here

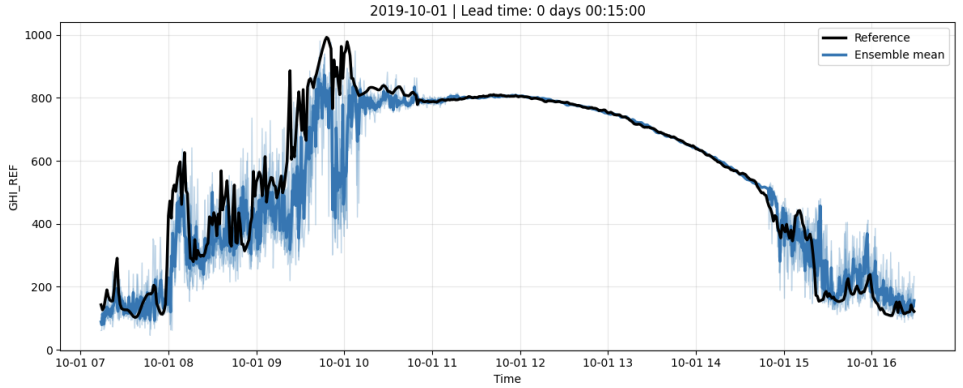
Results reported in the preprint [8]

[1] Owned and operated by CIEMAT
[2] Data available at: https://doi.org/10.1594/PANGAEA.980067
[3] Data available at: https://doi.org/10.5281/zenodo.21682387

FarSky – Qualitative Results







Effect of Latent-Coupling on Generative Solar Forecasting FarSky vs Decoupled baseline (FAR + CNN)

- **Ensemble forecast**: 4 trajectories per forecast issue
- **Deterministic forecast**: ensemble mean
- **Probabilistic forecast**: student's t-distribution fitted to the ensemble
- **Ramp event prediction**: event predicted if detected in ≥25% of ensemble trajectories (ramp event definition according to [3])
- **Evaluation**: metrics averaged over all lead times of the 30-min forecast horizon

	Deterministic			Probabilistic			Ramp Events		
Model	**RMSE ↓ [W/m²]**	**MAE ↓ [W/m²]**	**MBE [W/m²]**	**CRPS ↓ [W/m²]**	**IS (α=0.32) [W/m²] ↓**	**Calib. Error ↓**	**Precision ↑**	**Recall ↑**	**F1 ↑**
FAR + CNN	75.4	51.3	-31.0	44.3	247.0	0.33	0.42	**0.76**	0.54
FarSky	**62.7**	**33.4**	**5.0**	**26.7**	**141.8**	**0.17**	**0.46**	0.75	**0.57**

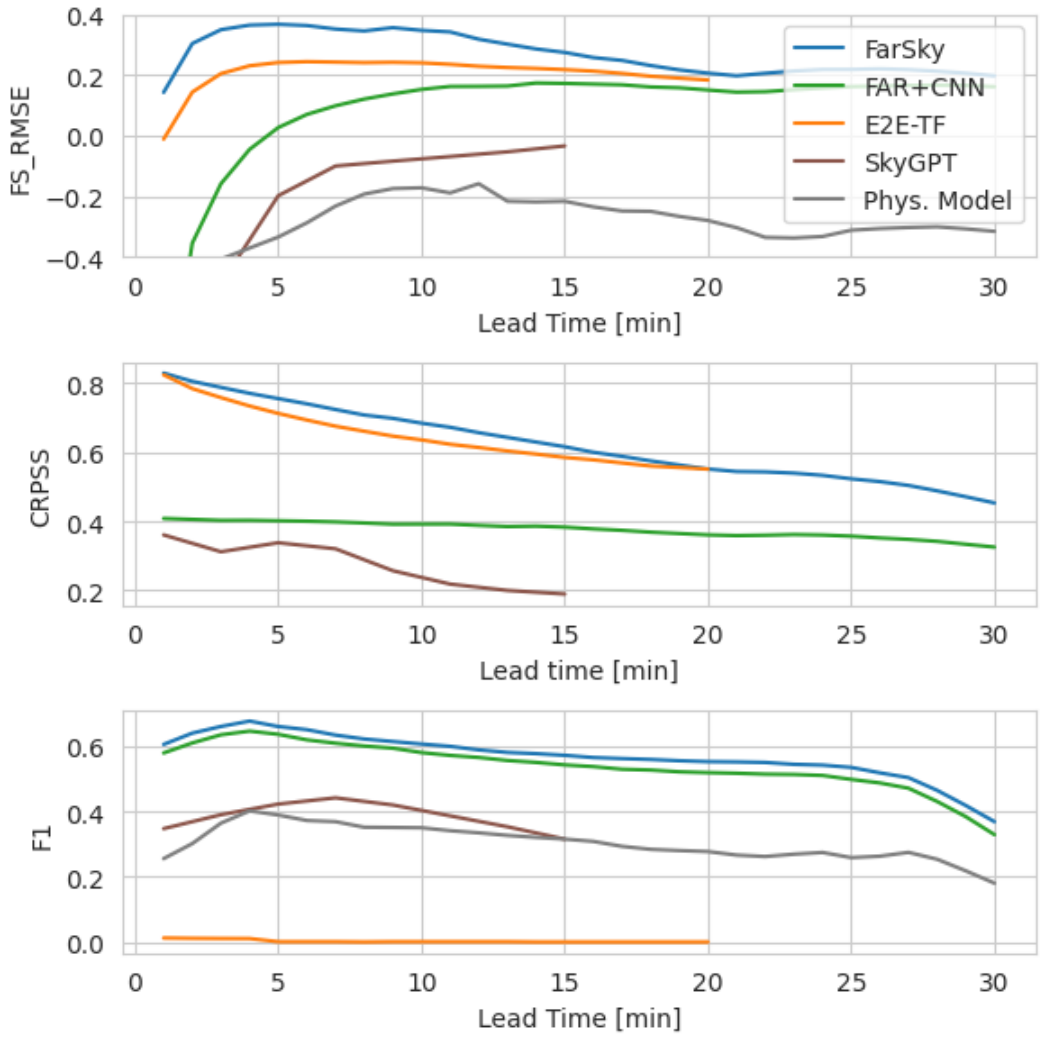
Calib. Error $= \frac{1}{K}\sum_k |PICP_k - PI_k|$

→Significant improvements forecast accuracy, primarily through bias reduction
→Improved probabilistic performance in accuracy (CRPS), reliability (calib. error) and sharpness (IS)
→ Ramp-event prediction largely unaffected, with a slight improvement in F1-score

Comparison to State-of-the-Art Models

- State-of-the-art models:
  - Decoupled diffusion model (FAR+CNN) [10]
  - End-to-end transformer model (E2E-TF) [4]
    - 20 min forecast horizon
    - Probabilistic forecasts of via post-processing using quantile-based error distributions [6]
  - Physics-based model [5]
    - Optimized on ramp events
  - SkyGPT [11]
    - Adapted version trained on GHI with PSA data

→FarSky model outperforms all models in the evaluated deterministic, probabilistic and ramp event metrics



Note: no error distributions of physics-based model available required for probabilistic post-processing

Summary & Conclusion

- **Problems with existing paradigms**
  - Deterministic end-to-end ML models
    - Strong average accuracy
    - Temporal smoothing → no information on variability within forecast
  - Decoupled generative approaches
    - Better capturing of variability
    - Domain shift → Degraded irradiance estimates on synthetic image data
- **Core contribution: Latent-coupled generative framework FarSky**
  - Joint representation of sky imagery and irradiance
  - Direct irradiance prediction from latent space (without prior image decoding) → no domain shift problems
  - Implicit regularization → increased robustness
- **Outcomes**
  - Increased forecast skill (>30% on benchmark dataset)
  - Significant improvement over state-of-the-art in ramp event prediction
- **Outlook**
  - Improve generalization (multi-site datasets, multi-modal/multi-task training)
  - Explore potential and direction of ASI foundation models

Thank you for your attention!
Questions?
yann.fabel@dlr.de

Preprint:
https://arxiv.org/abs/2608.11254

Forecast quality depends on the modeling paradigm — latent coupling resolves the trade-off between accuracy and variability.

References

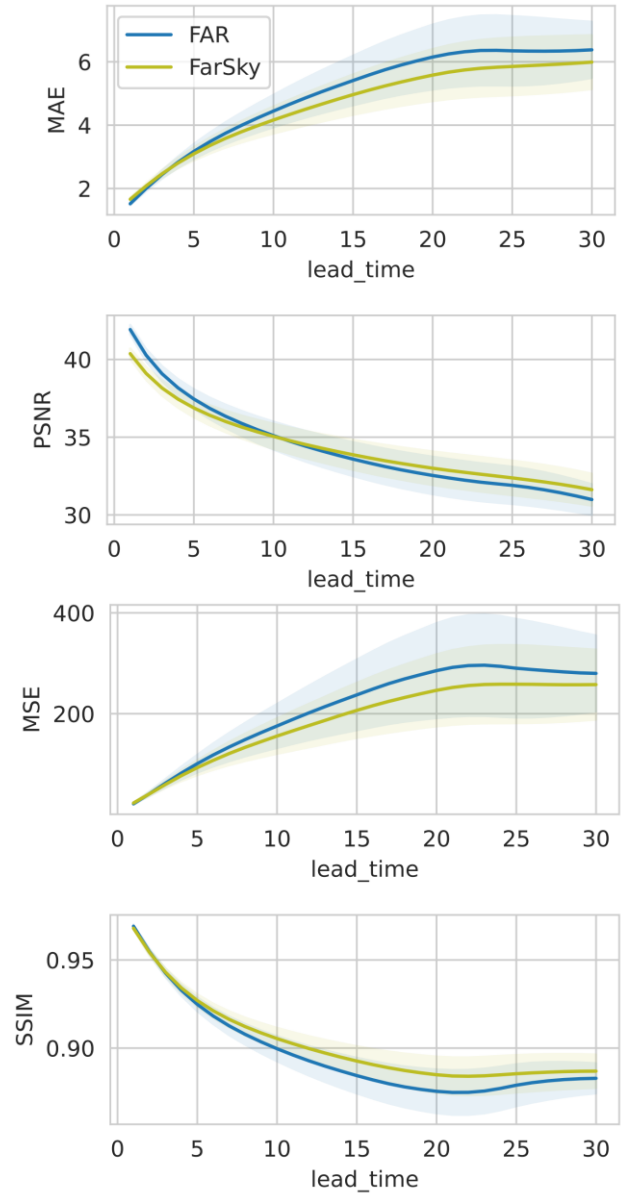
1. Gu et al. 2025, **Long-Context Autoregressive Video Modeling with Next-Frame Prediction,** DOI: 10.48550/ARXIV.2503.19325
2. Logotetis et al. 2022, **Benchmarking of solar irradiance nowcast performance derived from all-sky imagers**, DOI: 10.1016/j.renene.2022.08.127
3. Nouri et al. 2024, **Ramp Rate Metric Suitable for Solar Forecasting**, DOI: 10.1002/solr.202400468
4. Fabel et al. 2023, **Combining deep learning and physical models: a benchmark study on all-sky imagerbased solar nowcasting systems**, DOI: 10.1002/solr.202300808
5. Nouri 2020, **Solarstrahlungs-Kürzestfrist-Vorhersagesystem für die Ertragsoptimierung eines Parabolrinnenkraftwerks,** DOI: 10.18154/RWTH-2020-03451
6. Nouri et al. 2023, **Probabilistic solar nowcasting based on all-sky imagers**, DOI: 10.1016/j.solener.2023.01.060
7. Fabel et al. 2026, **A comprehensive study of the all-sky image and solar irradiance dataset SolarVision Almería**, DOI: 10.1016/j.solener.2026.114622
8. Fabel et al. 2026, **FarSky: Task-Aware Latent-Space Coupling for Generative Intra-Hour Solar Forecasting,** DOI: 10.48550/ARXIV.2608.11254
9. Chen et al. 2024, **Deep Compression Autoencoder for Efficient High-Resolution Diffusion Models**, DOI: 10.48550/ARXIV.2410.10733
10. Fabel et al. 2025, **Cutting-Edge Generative AI For Intra-Hour Solar Forecasting,** URL: https://elib.dlr.de/217688/
11. Nie et al 2024, **SkyGPT: Probabilistic ultra-short-term solar forecasting using synthetic sky images from physics-constrained VideoGPT**, DOI: 10.1016/j.adapen.2024.100172

BACKUP SLIDES

Comparison of Decoupled and Latent-Coupling Approaches Video Prediction

- Lead time-specific evaluation on 8-bit RGB images
- Metrics
  - MAE: mean absolute pixel error per image
  - MSE: mean squared pixel error per image
  - PSNR: Peak-signal-to-noise-ratio
    - Measures quality of reconstructed image by ratio of signal power to noise power
  - SSIM: Structural-similarity-index
    - Measures perceived image quality by comparing luminance, contrast and structure between original and reconstructed image
- Mean (solid line) and standard deviation (shaded area) over all predicted samples per lead time

→Very similar performance of video prediction with slightly better results of FarSky at higher LTs, indicating better representations from latent-coupling



Effect of Latent-Coupling on Generative Solar Forecasting

- **Deterministic Forecasts**
  - Lead time specific mean of GHI Ensemble Forecast
  - Metrics: RMSE, MAE, MBE, Forecast Skill (based on persistence)

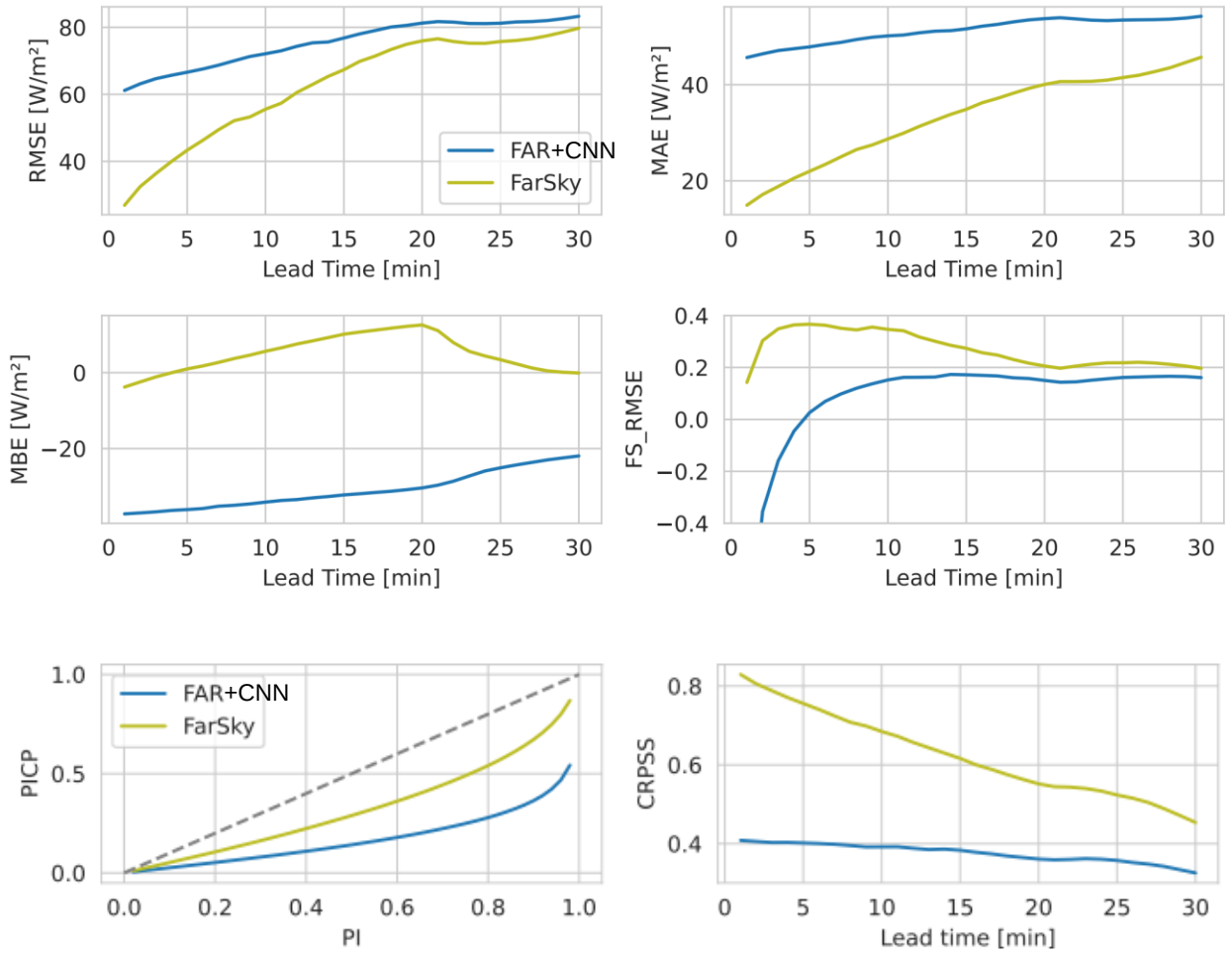
→Significant improvements in all metrics, particularly due to bias reduction

→FS almost consistently over 20%

- **Probabilistic Forecasts**
  - Lead time specific students t-distribution from Ensemble (4 samples)
  - PICP over PI: Indicator for reliability (how often an observed value falls in PI)
  - CRPSS: Skill of difference between forecasted cumulative distribution and observed value

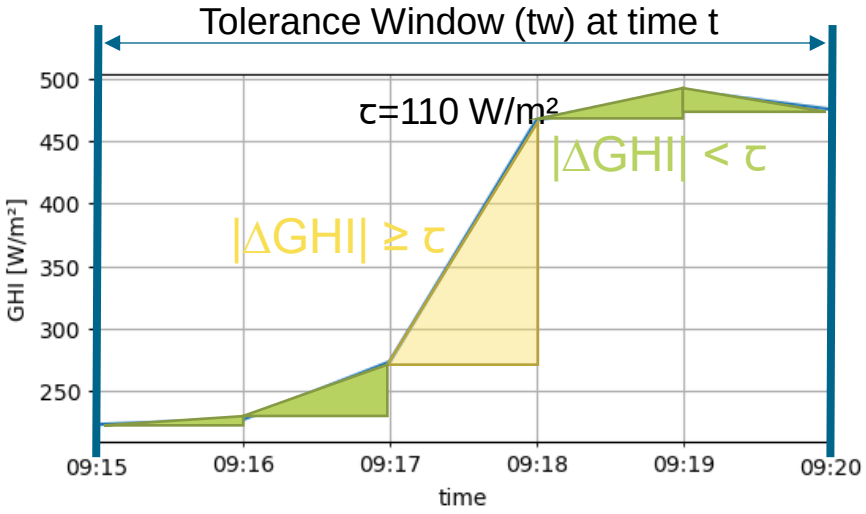
→Increased reliability and better accuracy

→Ensemble size limited (further analysis required)

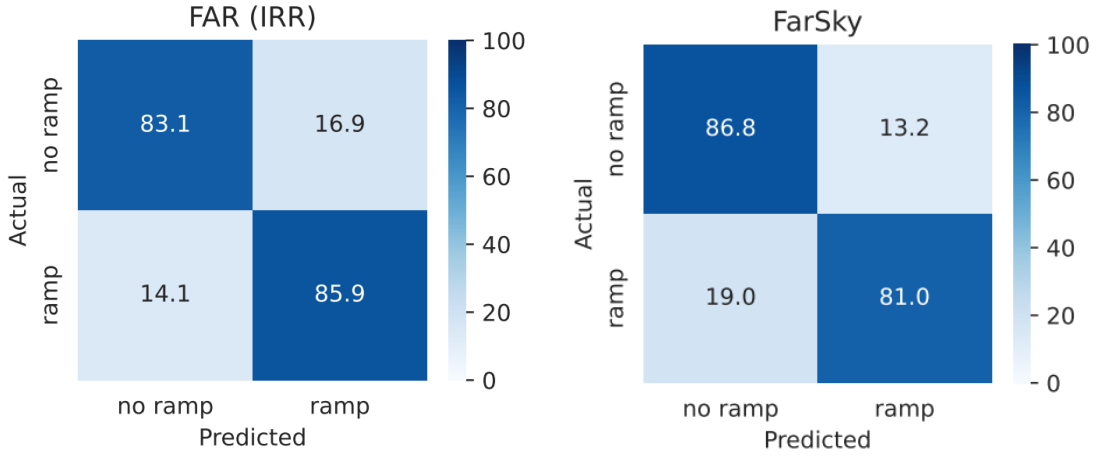


Comparison of Decoupled and Latent-Coupling Approaches
Ramp Event Prediction

- Ramp Event [3]:
  - Change in irradiance surpassing a predefined threshold within a short time window
- Evaluation via
  - Confusion Matrix per lead time
  - Classification scores (accuracy, precision, recall, f1-score)



Confusion matrix for lead time 15min



Average ramp scores over lead times

Model	Acurracy	Precision	Recall	F1
FAR (IRR)	0.86	0.42	**0.76**	0.54
FarSky	**0.88**	**0.46**	0.75	**0.57**

→ Both approaches are capable of predicting ramp events (no negative impact of bias)
→ FarSky shows slightly better performance