Energy and AI 25 (2026) 100786

Contents lists available at ScienceDirect

Energy and AI

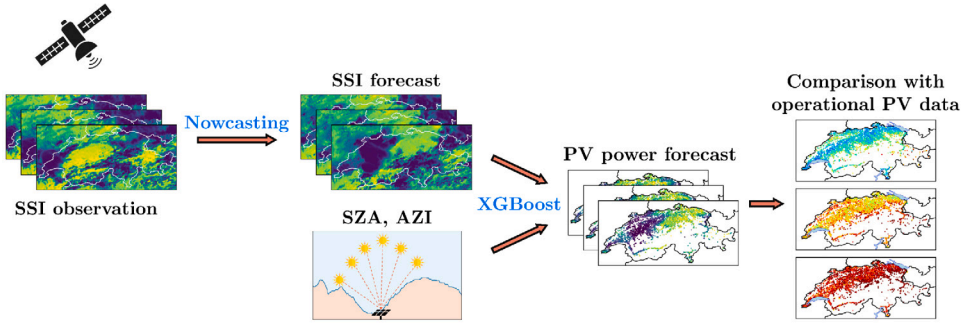
journal homepage: www.elsevier.com/locate/egyai

Intraday spatiotemporal PV power prediction at national scale using satellite-based solar forecast models

Luca Lanzilao [a,*], Angela Meyer [a,b]

[a] School of Engineering and Computer Science, Bern University of Applied Sciences, Quellgasse 21, Biel, 2501, Bern, Switzerland
[b] Department of Geosciences and Remote Sensing, TU Delft, Stevinweg 1, Delft, 2628 CN, South-Holland, The Netherlands

GRAPHICAL ABSTRACT



HIGHLIGHTS

- Satellite-based PV forecasting on a countrywide fleet of more than 6400 PV systems.
- Performance comparison of six satellite- and physics-based PV forecast models.
- Satellite-based models more accurate than IFS-ENS even after bias correction.
- Satellite-based models forecast daily PV generation with relative errors below 10%.
- Visualization of mesoscale cloud impacts on national PV power production.

ARTICLE INFO

Keywords:
Solar-energy nowcasting
Spatiotemporal forecasting
Probabilistic forecasts
Numerical weather prediction models

ABSTRACT

We present a novel framework for spatiotemporal photovoltaic (PV) power forecasting and use it to evaluate the reliability, sharpness, and overall performance of six intraday PV power nowcasting models. The model suite includes satellite-based deep learning and optical-flow approaches and physics-based numerical weather prediction models, covering both deterministic and probabilistic formulations. Forecasts are first validated against satellite-derived surface solar irradiance (SSI). Irradiance fields are then converted into PV power using station-specific machine learning models, enabling comparison with production data from 6434 PV stations across Switzerland. To our knowledge, this is the first study to investigate spatiotemporal PV forecasting at a national scale. We additionally provide the first visualizations of how mesoscale cloud systems shape national PV power production on hourly and sub-hourly timescales. Our results show that satellite-based approaches outperform the Integrated Forecast System (IFS-ENS), particularly at short lead times. Among them, SolarSTEPS and SHADECast deliver the most accurate SSI and PV power predictions, with SHADECast providing the most reliable ensemble spread. The deterministic model IrradianceNet achieves the lowest root mean square error,

* Corresponding author.
E-mail address: luca.lanzilao@bfh.ch (L. Lanzilao).

https://doi.org/10.1016/j.egyai.2026.100786
Received 11 March 2026; Received in revised form 13 May 2026; Accepted 24 May 2026
Available online 30 May 2026

while probabilistic forecasts of SolarSTEPS and SHADECast provide better-calibrated uncertainty. Forecast accuracy generally decreases with elevation. At a national scale, satellite-based models forecast the daily total PV generation with relative errors below 10% for 82% of the days in 2019–2020, demonstrating robustness and their potential for operational use.

1. Introduction

Solar energy plays a central role in the decarbonization of power systems and the mitigation of climate change. As its share in electricity generation continues to grow [1], accurate forecasts of surface solar irradiance (SSI) and photovoltaic (PV) power generation are becoming increasingly critical for grid operators, energy traders, and system planners to ensure grid stability and maintain the balance between supply and demand.

Although PV power forecasting has advanced considerably, most existing approaches remain limited in their ability to support operational decision-making at the scale required by modern power systems. In particular, many studies focus on single-site or small regional deployments [2]. For example, a large fraction of existing methods rely on deterministic approaches [3,4], often based solely on PV production time series [5–7]. Moreover, most existing studies rely on limited datasets, typically comprising only a few dozen PV systems [4,8–11]. Such approaches fail to capture the spatial variability of cloud dynamics and PV production across large geographic areas. Even the largest datasets used to date, such as [12] with 316 PV systems, remain insufficient to represent the variability encountered in national-scale PV fleets. However, grid-level decisions, such as reserve allocation, congestion management, and cross-border energy trading, require accurate and spatially consistent forecasts at regional to national scales. As a result, models developed and validated on limited datasets may not generalize to large and heterogeneous PV fleets.

To address this limitation, spatiotemporal forecasting approaches based on satellite-derived SSI and numerical weather prediction (NWP) models have gained increasing attention. Two main paradigms can be distinguished: satellite-based nowcasting and NWP-based forecasting, which must subsequently be coupled with an irradiance-to-power conversion model to obtain PV power forecasts [13]. NWP models have been widely applied and validated for SSI forecasting, with lead times extending up to 15 days [14–18]. These models rely on data assimilation and the numerical integration of atmospheric equations [19,20], enabling physically consistent forecasts at global scale [21]. Consequently, operational NWP systems such as the high-resolution Integrated Forecasting System ensemble (IFS-ENS) developed by the European Centre for Medium-Range Weather Forecasts (ECMWF) are typically initialized only a limited number of times per day. In the case of IFS-ENS, four runs are performed daily at 6-hour intervals, each producing forecasts that extend several days ahead [22].

In contrast, satellite-based approaches provide high-frequency observations of SSI, enabling accurate nowcasting at short lead times. By tracking cloud motion using techniques such as optical flow [14,23], these methods can generate forecasts with high spatial and temporal resolution. Extensions incorporating stochastic perturbations allow for probabilistic forecasting and uncertainty quantification [24]. Nevertheless, classical advection-based approaches struggle to capture cloud evolution [25–27]. Recent advances in machine learning, combined with the availability of multi-annual satellite data records, have enabled the development of data-driven spatiotemporal forecasting models that address this limitation. Early approaches employed deep learning architectures such as long short-term memory (LSTM) networks, convolutional neural networks (CNNs), and graph neural networks (GNNs), trained on combinations of ground-based and satellite data [28,29]. A notable advancement was introduced by [30], who proposed a ConvLSTM-based model for probabilistic SSI forecasting. More recently, [31] developed a latent diffusion model that improves forecast accuracy and increases extreme event prediction performance by 60% compared to the model developed by [30].

Despite these advances, it remains unclear how different forecasting paradigms, i.e. satellite-based, machine learning-based, and NWP-based, compare when evaluated consistently across large and heterogeneous PV networks. In addition, existing studies often rely on different datasets, metrics, and experimental setups, making direct comparisons difficult. A systematic benchmarking framework that evaluates multiple forecasting approaches under identical conditions is therefore required to quantify their relative strengths and limitations, particularly in operational settings.

In this work, we present a comprehensive framework for spatiotemporal PV power forecasting and evaluation at national scale. We first benchmark several state-of-the-art approaches for SSI forecasting, including the probabilistic optical flow model SolarSTEPS [27], the deep generative diffusion model SHADECast [31], the deterministic deep learning model IrradianceNet [30], and the IFS-ENS model, a widely used NWP model [32–34]. The resulting irradiance forecasts are then converted into PV power using station-specific machine learning models, enabling direct comparison with production data from 6434 PV systems across Switzerland. This study addresses three key gaps in the literature. First, it provides a large-scale and systematic comparison of satellite-based and NWP-based forecasting models using operational PV data at national scale. Second, it demonstrates the applicability of probabilistic spatiotemporal forecasting methods to large PV fleets, enabling a comprehensive assessment of forecast uncertainty, reliability, and sharpness. Third, it introduces a novel visualization of mesoscale cloud dynamics and their impact on aggregated PV production, offering new insights into the link between atmospheric processes and power system variability.

The remainder of this paper is structured as follows. Section 2 describes the datasets and evaluation metrics. Section 3 presents the forecasting models and the irradiance-to-power conversion approach. Section 4 details the forecasting framework. Section 5 discusses the results, and Section 6 concludes the paper.

2. Datasets and metrics

2.1. Surface solar irradiance

SSI represents the combination of direct and diffuse solar radiation incident on a horizontal plane at the Earth's surface. It is also called global horizontal irradiance and is typically measured by ground-based pyranometers or retrieved from satellite observations. In this work, we use satellite-derived SSI fields provided by EUMETSAT Climate Monitoring Satellite Application Facility (CM SAF), notably the High-Resolution European Surface Solar Radiation Data Record (HANNA) — see Appendix B. The SSI values are derived from data collected by the Spinning Enhanced Visible and Infrared Imager (SEVIRI) onboard the Meteosat Second Generation (MSG) satellite positioned in geostationary orbit at 0° longitude [35], which provides Earth scans with a sampling distance at nadir of 3 km. HANNA builds on the HelioMont algorithm [36,37] to estimate SSI from the SEVIRI channel intensities. The resulting satellite data are projected onto a grid with a spatial resolution of 0.01° for both longitude and latitude, covering a domain of [-17°W, 35°E] × [30°N, 70°N]. The MSG SSI fields are available for the years 2019–2020 at a temporal resolution of 15 min [38].

The HANNA SSI has undergone initial validation using ground-based measurements from several networks, including the Baseline Surface Radiation Network (BSRN), the Global Energy Balance Archive

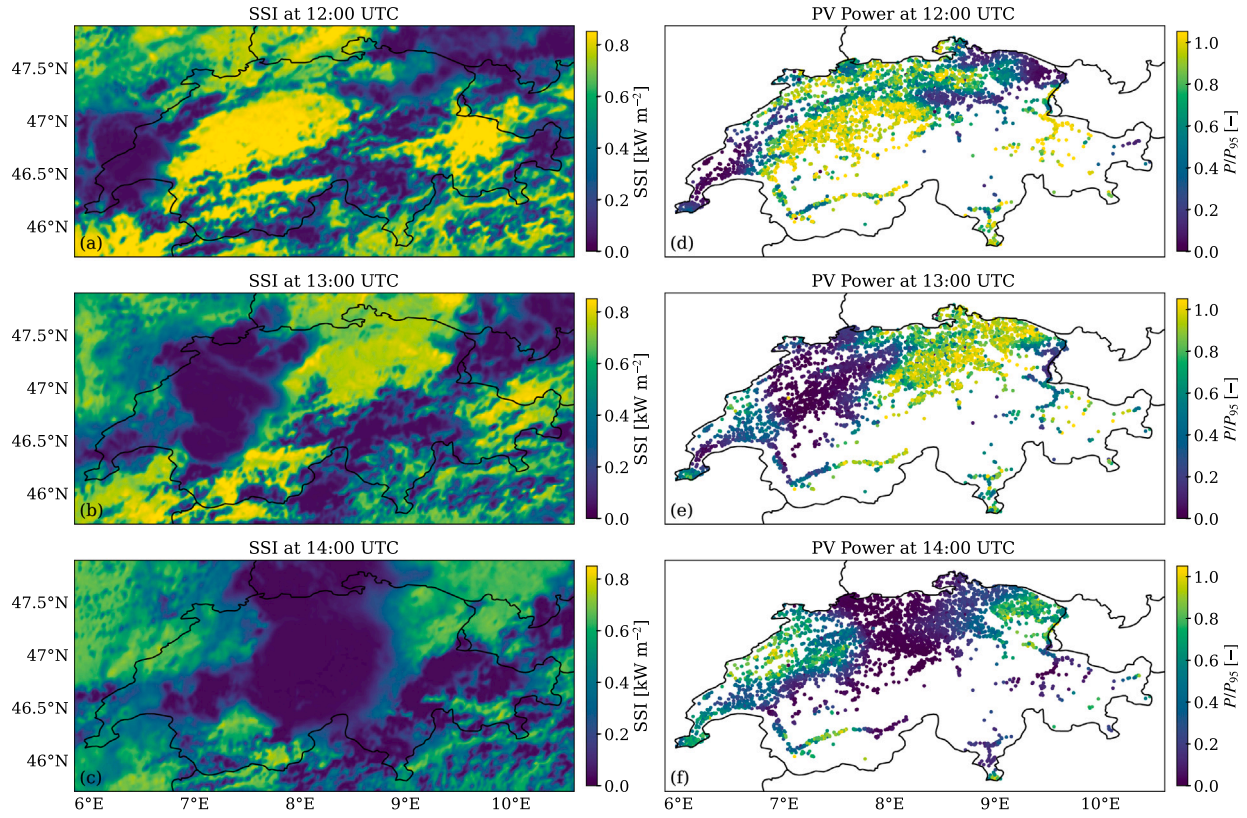


Fig. 1. (a–c) Satellite-based SSI fields and (d–f) PV power output of the 6434 PV stations, observed on 6 August 2019 at three time stamps: 12:00 UTC, 13:00 UTC and 14:00 UTC. The black lines denote national borders. Note that the PV power output is normalized using the station-specific 95th percentile of the power time series.

(GEBA), Deutscher Wetterdienst (DWD), and SwissMetNet. In total, 406 stations were used for this validation. Results indicate that, for stations located in the 0–500 meter elevation band, the station- and monthly-averaged mean absolute difference measures 5.3 W m^{-2} while the same error measures 6.9 W m^{-2} and 10.9 W m^{-2} for stations lying in the elevation range 500–1500 and above 1500 m, respectively [38]. It is worth noting that 5 out of the 731 days in the dataset were excluded from our analysis due to the presence of missing values for some of the time steps over our region of interest. This issue is also acknowledged in [38].

Fig. 1(a–c) displays HANNA SSI fields obtained over the study area on the 6th of August 2019 at three time stamps: 12:00, 13:00, and 14:00 UTC. These panels illustrate how the dataset enables a sharp representation of cloud structures (panels (a–c)), identifiable by their radiative effect on the surface as areas with lower SSI values. During this period, we also observe convective cloud growth over central Switzerland, along with a gradual decrease in average SSI consistent with the progressing diurnal cycle. In Appendix A, the full two-hour period at 15-minute temporal resolution is shown.

2.2. PV power production

Our study focuses on the years 2019–2020, which coincide with the HANNA SSI record. This resulted in 7633 usable PV stations from across Switzerland with 15-minute resolution measurements, whose proprietary production data were obtained for this study under a data usage agreement. Given the large multi-annual record of thousands of PV stations, we developed and applied automated cleaning procedures to identify and filter anomalous data to ensure data quality. Specifically, we split the PV production time series of each PV station into two segments, one for 2019 and one for 2020, and computed statistical metrics such as mean, standard deviation, and skewness for each year. Given the seasonal stability in the statistical properties of solar generation, we expected these distributions to be broadly consistent. To identify outliers, we applied a tolerance threshold. This threshold is determined by analyzing SSI measurements from the Baseline Surface Radiation Network (BSRN) station in Payerne, which provides 1-minute resolution data from 1992 onwards. We computed the annual total SSI and the relative difference between consecutive years, finding a mean interannual difference of 6.3% and a maximum difference of 17.7%. In addition, [39] reported interannual variability in PV power output ranging from 2.45% to 12.07% for the 90 PV stations located in Portugal. Based on these observations, we chose a 10% threshold as a pragmatic criterion to filter anomalous PV stations. Stations for which any of the computed metrics differed by more than this threshold between the two years were flagged as outliers and excluded from the dataset. This filtering strategy reduced the number of stations considered in this study from 7633 to 6434. The majority of the excluded stations exhibited null or persistently low power generation over extended periods, likely due to temporary shading, sensor malfunctions, data transmission issues, or other operational anomalies. Assuming an average of 12 h of daylight per day over the year, our study relies on roughly 225 million PV production data points that can be used to train or validate forecast models. In the remainder of the article, the PV power generated from each station is normalized with P_{95}, i.e. the station-specific 95th percentile of the power time series. This normalization provides a robust basis for comparing stations with different capacities, mitigates the impact of rare extreme values while still preserving the upper range of realistic operating conditions.

Fig. 1(d–f) shows the location and power output of the 6434 PV stations considered in this study, measured on the 6th of August 2019 at three time stamps: 12:00, 13:00, and 14:00 UTC. The cloud system over Switzerland, visible in the SSI fields shown in Fig. 1(a–c), is clearly reflected in the PV power output. Areas of low PV power generation

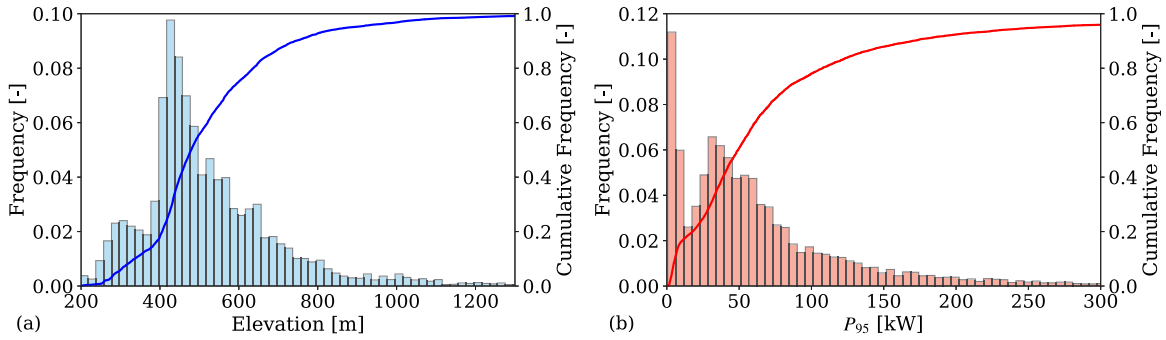


Fig. 2. Empirical probability density function and cumulative density function (CDF) of the PV stations (a) elevation and (b) P_{95} value.

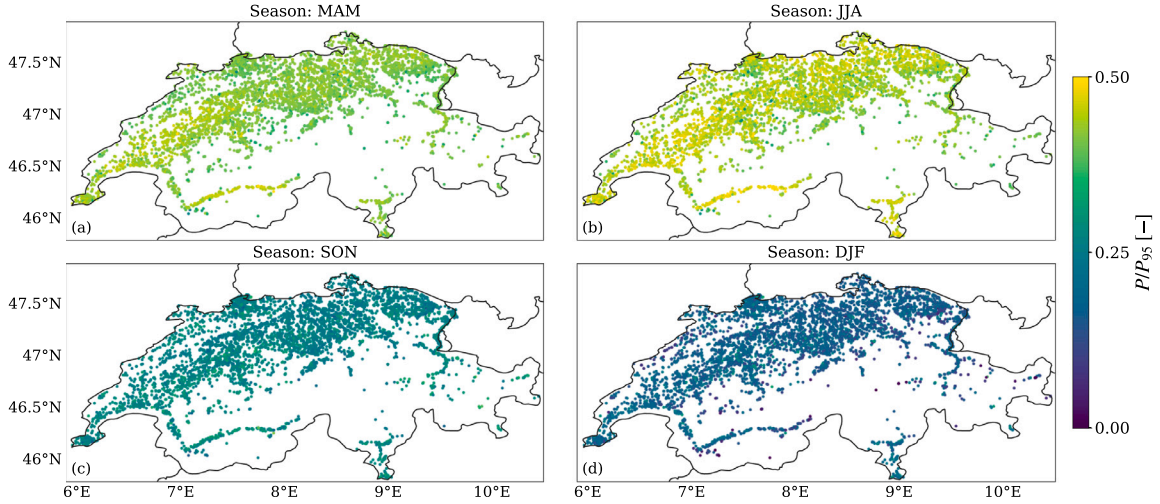


Fig. 3. Two-year average of PV power output normalized with the station-specific P_{95} value during (a) MAM, (b) JJA, (c) SON, and (d) DJF. The averaging is performed over timestamps between local sunrise and sunset times.

align remarkably well with regions of low SSI, demonstrating a strong correlation between these variables. We note that the full two-hour period at 15-minute temporal resolution is shown in Appendix A. Fig. 1 highlights the critical importance of spatiotemporal information over purely time-series-based methods unaware of the surroundings SSI field for forecasting solar energy. To our knowledge, it is the first visualization of how mesoscale cloud systems affect the production of a large country-wide fleet of PV systems. It illustrates how, by relying on geostationary satellites, we can infer the footprint of cloud cover on the country-wide PV production with high accuracy and low latency.

Fig. 2(a) shows the distribution of the site elevation of the 6434 PV systems analyzed in this study. Approximately 56% of the stations are situated between 0–500 m above sea level (a.s.l.), a range where the HANNA dataset exhibits the highest accuracy. Fewer than 1% of the stations are located above 1300 m a.s.l., with the highest reaching an elevation of approximately 2450 m a.s.l. Fig. 2(b) illustrates the distribution of the 95th percentile of the power time series of each station and its corresponding cumulative density function (CDF). About 16% of the stations have a P_{95} value below 10 kW, typically corresponding to residential installations. The majority of the stations (about 76%) fall within the 10–200 kW range while only 4% have a P_{95} value exceeding 300 kW.

Finally, Fig. 3 shows the seasonally averaged normalized power output calculated during daytime hours (from sunrise to sunset) over the 2019–2020 period. While the intraseasonal variability is relatively low, interseasonal differences are significant as expected. For instance, the station-averaged P/P_{95} ratio reaches 0.43 in summer (JJA), but drops to 0.18 in winter (DJF).

2.3. Metrics

To evaluate the performance of the forecast models, we employ different metrics for deterministic and probabilistic predictions. For deterministic forecasts, accuracy is assessed using the mean absolute error (MAE), root mean square error (RMSE), and mean bias error (MBE). The MAE provides an intuitive measure of the average magnitude of the forecast errors, while the RMSE assigns greater weight to large deviations and is therefore more sensitive to extreme errors. The MBE, on the other hand, captures the average signed error, thereby indicating whether a model tends to systematically overestimate or underestimate the observations.

For probabilistic forecasts, deterministic scores are computed using the ensemble mean, while additional metrics evaluate both the reliability and sharpness of the predictive distributions. The prediction interval coverage probability (PICP) measures the proportion of observations that fall within the forecast intervals and is thus an indicator of calibration. Complementary to this, the mean prediction interval width (MPIW) and the prediction interval normalized average width (PINAW) quantify the average width of the prediction intervals and reflect the forecast sharpness, with narrower intervals being preferable provided that coverage remains adequate. The continuous ranked probability score (CRPS) offers a more comprehensive evaluation by integrating information about both reliability and sharpness into a single metric. For deterministic models, the CRPS reduces to the MAE [40]. Finally, rank histograms provide a graphical diagnostic of ensemble forecasts, revealing whether the ensemble spread appropriately represents the observations and highlighting potential issues such as bias, underdispersion, or overdispersion. We note that, when the metric acronym

is preceded by the letter 'n', this indicates that the metric has been normalized.

Taken together, these metrics allow for an in-depth comparison of deterministic and probabilistic models, capturing not only point prediction accuracy but also the quality of uncertainty quantification. For more details on the definition and computation of these metrics as well as the normalization factors, we refer the reader to Appendix C.

3. Models

To enable a comprehensive comparison between satellite-based spatiotemporal nowcasting models and a physics-based NWP model, we consider six distinct approaches, covering both deterministic and probabilistic methods. Section 3.1 provides a brief overview of each model and a summary of its underlying assumptions and design choices. Finally, Section 3.2 describes the methodology adopted to convert SSI into PV power.

3.1. SSI forecast models

3.1.1. SolarSTEPS

SolarSTEPS is a probabilistic optical-flow-based model developed for forecasting satellite-derived clear-sky index (CSI) fields [27]. First, the model computes cloud motion vectors (CMVs) from a sequence of CSI fields using the Lucas–Kanade optical-flow algorithm [23]. Subsequently, a Fast Fourier Transform (FFT) is applied to decompose each input CSI field into distinct spatial scales. Each spatial scale is then forecast using a separate linear autoregressive (AR) model. These AR models are designed to simulate the temporal evolution of cloud patterns in a Lagrangian frame and include spatially correlated noise to generate an ensemble of forecasts. Each scale is modeled with its own set of AR coefficients. The forecast components from all spatial scales are then summed and advected using the initially estimated CMVs to produce the final forecast.

SolarSTEPS stands out for its ability to capture both cloud advection, using an optical-flow algorithm, and cloud evolution, through scale decomposition combined with AR models. Additionally, it supports the generation of ensemble forecasts, allowing for uncertainty quantification. However, as linear AR models assume data stationarity, the model is unable to predict distribution shift, which occurs in weather conditions in which cloudiness is growing or decreasing over the course of the forecast. Additionally, we also evaluate a simplified version of SolarSTEPS, referred to as SolarSTEPS-pa. This pure advection (PA) variant does not model the temporal variability of the CSI. Instead, it perturbs the CMVs derived from the input image sequence and uses them to advect the CSI. As a result, this version captures only the effect of cloud advection, omitting both the scale decomposition and AR modeling components. The model was originally calibrated on the HelioMont dataset [37], and the AR models coefficients were estimated using the Yule–Walker equations [41]. Moreover, the ensemble forecast consists of 10 members. In this work, we adopt the same parameter settings as reported in [27] to which we refer for more details.

3.1.2. IrradianceNet

IrradianceNet is a deterministic deep learning model introduced by [30], designed to forecast satellite-derived CSI fields using only satellite-based CSI inputs. The model employs a spatiotemporal autoencoder architecture composed of three ConvLSTM layers in both the encoder and decoder. The encoder processes sequences of CSI fields to capture spatiotemporal dependencies. The final hidden state of the encoder serves as a compressed representation of the input dynamics, which is then passed to the decoder. The decoder performs upsampling to generate future CSI fields, effectively achieving spatiotemporal forecasting.

IrradianceNet was one of the first deep learning approaches for solar irradiance forecasting and has shown significant agreement with ground-based pyranometer observations [30]. In addition to forecast blurring, a key limitation of the model is its deterministic nature, which prevents ensemble forecasting and therefore does not allow for uncertainty quantification. However, we include it in our study since it provides a reliable deterministic baseline.

The model was originally trained on the SARAH-2.1 dataset provided by EUMETSAT [42]. However, [27] later retrained it using HelioMont data [37], extending the forecast horizon from 2 to 8 time steps via an autoregressive strategy. In this work, we adopt the same model architecture and pre-trained weights as presented in [27].

3.1.3. SHADECast

SHADECast is the first deep generative diffusion model developed for intraday solar energy forecasting. The model takes as input a sequence of CSI fields and combines a variational autoencoder (VAE), a latent deterministic nowcaster, and a latent diffusion model to generate probabilistic SSI forecasts. The VAE, built with a symmetric architecture of two downsampling 3D residual blocks, compresses the input CSI sequence into a latent representation. This latent representation is then passed to a deterministic nowcaster operating in latent space, which comprises four Adaptive Fourier Neural Operator (AFNO) blocks, followed by a temporal transformer and an additional four AFNO blocks. To generate the probabilistic forecast, a latent diffusion model is employed, which maps Gaussian noise to future CSI representations. The diffusion process is guided by a latent denoiser with a symmetric U-Net architecture, conditioned on the output of the nowcaster. Finally, the VAE decoder transforms the forecast latent representations back into the physical space, generating the predicted CSI fields.

As a state-of-the-art model for solar energy nowcasting, SHADECast is included in our study. The model was trained on satellite-derived CSI fields from the HelioMont dataset and generates ensemble forecasts with 10 members. In this work, we adopt the same architecture and pre-trained weights as described in [31], and refer to that source for detailed information on the model architecture and training procedure.

3.1.4. IFS-ENS

We incorporate forecasts from the IFS-ENS model developed by ECMWF [20] to also include a physics-based probabilistic forecast model. The model provides SSI forecasts four times a day at 00:00, 06:00, 12:00, and 18:00 UTC, with outputs available at hourly intervals. To maintain consistency with the probabilistic satellite-based models, we randomly select 10 ensemble members from the original set of 50. The IFS-ENS forecasts have an effective horizontal resolution of approximately 9 km, defined by the model's spectral truncation and corresponding to a quasi-uniform spacing on the native reduced Gaussian grid. This ensures a nearly uniform resolution in physical space across the globe. When expressed in latitude–longitude coordinates, the equivalent spacing depends on latitude. A 9 km distance corresponds to approximately 0.08° in latitude, while over Switzerland it corresponds to about 0.11° in longitude due to the convergence of meridians. In this work, we interpolate the model forecasts onto a regular latitude–longitude grid with a spacing of 0.08° in both latitude and longitude for consistency.

To investigate whether bias correction could improve performance, we applied an ML-based correction to the IFS-ENS SSI forecasts. The correction network is a symmetric U-Net–based convolutional neural network [43]. Further details on the training procedure and network design are provided in Appendix D. Throughout this work, we refer to the bias-corrected forecasts as IFS-ENS-corrected.

3.2. Irradiance-to-power conversion model

To enable validation on the PV power production dataset presented in Section 2.2, the satellite-based and NWP models SSI forecasts have to be converted to corresponding PV power output. The relationship between SSI and PV power is often referred to as the solar power curve [13,44]. Several physical and semi-empirical models have been developed for this purpose. They typically comprise a chain of sub-models, including separation [45], transposition [46], temperature [47], inverter [48], shading [49], and loss models [50], and require detailed technical specifications of each PV installation, such as panel orientation, tilt angle and module electrical characteristics [51]. However, obtaining such detailed information is highly challenging when working with thousands of PV installations. Furthermore, since only SSI observations are available, estimating the direct and diffuse components would require additional modeling steps, thereby introducing further uncertainties and potential biases into the model chain. For these reasons, we adopt a data-driven regression approach that maps the SSI forecasted values to PV power production in a site-specific manner. The main limitation of a data-driven approach lies in its dependence on historical time-series data, making it unsuitable for PV systems without any prior production records. However, given the extensive availability of operational PV power data in our dataset, an ML-based conversion approach is well supported.

The power output of a PV system is governed by both meteorological and geographical factors. Among the meteorological variables, SSI is the dominant predictor, exhibiting a very strong correlation with PV power output. For example, [52] reported an R^2 value of 0.988, while [53] obtained 0.984. In contrast, air temperature represents a second-order effect. For instance, [52] reported a correlation coefficient with PV power of 0.377. A key limitation is that air temperature is only an indirect proxy for module temperature, which is the physically relevant quantity governing PV efficiency and additionally depends on factors such as mounting configuration and panel material. This is supported by [54], who reported a correlation of 0.71 between PV power and module temperature, compared to only 0.13 for ambient air temperature. For these reasons, air temperature was not included as a predictor in this study. Other meteorological variables, such as cloud type, dew point, and relative humidity, exhibit progressively weaker correlations with PV power output [55,56] and were therefore also excluded. As a result, satellite-observed SSI was selected as the sole meteorological predictor. Additionally, we incorporated the solar zenith angle (SZA) and the solar azimuth angle (AZI) to account for geographical effects, such as shading by mountains. SZA and AZI were computed using HORAYZON, a ray-tracing algorithm for computing the local horizon and sky view factor [57]. This is particularly important in regions where topographical features can cause substantial variation in SSI. Fig. 4(a, b) illustrates the importance of accounting for local SZA and AZI angles by showing the sun position over an entire year at 30-minute intervals for two PV installations in the vicinity of Geneva and Biasca, two locations in Switzerland, which we will denote with S-GE and S-TI, respectively. Notably, the S-TI site receives considerably less direct sunlight during the winter months. Without accounting for the local SZA and horizon, this reduced irradiance could mistakenly be attributed to cloud cover.

To account for the periodic nature of time, we also included both the day of the year (DoY) and the hour of the day (HoD) as predictors. Instead of using their raw values, we applied sine and cosine transformations to preserve their cyclical nature, therefore introducing four additional predictors. In summary, each PV system is characterized by seven predictors, i.e. satellite-observed SSI, SZA, AZI and the four time-based features derived from the DoY and HoD, and one target variable, i.e. its measured PV power output. We emphasize that the choice to use satellite-observed SSI, rather than forecasted SSI, as a predictor is intentional. In fact, this choice ensures that the irradiance-to-power model learns the underlying physical relationship between irradiance and PV power, independent of forecast-specific errors. Training the model directly on forecasted SSI would entangle the irradiance-to-power relationship with model-specific biases and uncertainties, potentially reducing generalization and introducing dependence on a particular SSI forecasting model. In contrast, training on observed SSI decouples the conversion step from forecast-specific errors, allowing the same model to be consistently applied across all six SSI forecasting approaches. This ensures a fair, modular, and comparable evaluation of the different forecasting models. Moreover, to ensure consistency with the forecasting setup, the SSI HANNA fields are downsampled to a spatial resolution of 0.02° before being interpolated to the station locations during the training of the irradiance-to-power conversion models (see Section 4). This guarantees that the SSI fields used for the irradiance-to-power conversion are fully consistent with the spatial resolution of the SSI forecasts generated by the satellite-based models.

We perform the irradiance-to-power conversion with the extreme gradient boosting (XGBoost) algorithm [58] due to its good performance and compatibility with GPU acceleration. We note that alternative ML models, such as decision trees [59], support vector machines [60], light gradient boosting [61], histogram-based gradient boosting [62], and long short-term memory networks [63], were also explored, but they did not yield performance improvements over XGBoost for this task. To account for differences in nominal capacity and elevation among the PV systems considered in this study, we trained a separate XGBoost model for each station. Each model uses the HANNA SSI fields interpolated to the station location together with the corresponding operational PV power measurements, using the predictors and target variable described above. The two-year dataset was partitioned into non-overlapping blocks of 12 days. Within each block, ten days were assigned to the training set, one day to the validation set, and the remaining day to the test set. In total, the training set comprises 609 days, while the validation and test sets contain 61 days each. Only daylight periods, i.e. between sunrise and sunset, were included. All input and target variables were scaled to the [0,1] range to improve numerical stability and facilitate model convergence. The hyperparameters of each model, such as the learning rate, maximum tree depth, number of estimators, and the L1 and L2 regularization terms, were tuned independently for each station using Optuna [64]. In total, we trained 6434 station-specific irradiance-to-power conversion models. Their performance on the test set is presented in Section 5.1.

4. Spatiotemporal PV power forecast workflow

The satellite-based nowcasting models, i.e. IrradianceNet, SolarSTEPS, and SHADECast, were originally trained and calibrated on seven years (2007–2013) of satellite-derived CSI fields from the HelioMont dataset, which provides SSI fields at a spatial resolution of 0.02°. During training, forecasts were generated at a temporal resolution of 15 min with a prediction horizon of 2 h, a setup that is also adopted in this study. For the present evaluation, the HANNA product is used, offering SSI fields at a finer spatial resolution of 0.01° for the period 2019–2020. This dataset was selected to ensure temporal overlap with the PV power measurements. Therefore, to maintain consistency with the original training configuration, the HANNA fields, which are used exclusively for inference and evaluation, are downsampled to 0.02° using a 2×2 pixel averaging kernel. The SSI fields are then converted to CSI, a dimensionless ratio of observed SSI to clear-sky SSI (SSI_{cs}) that typically ranges from 0 to approximately 1.2. SSI_{cs} is computed using the Ineichen model [65] based on precomputed look-up tables from the SOLIS model [66]. This method was chosen for its independence from external data and is implemented in the pvlib Python package [67]. Each nowcast takes four consecutive CSI fields (covering one hour) as input and predicts the next eight fields, corresponding to a two-hour forecast horizon, which are subsequently converted back to SSI. Given HANNA's 15-minute temporal resolution, this input–output sequence spans 3 h in total.

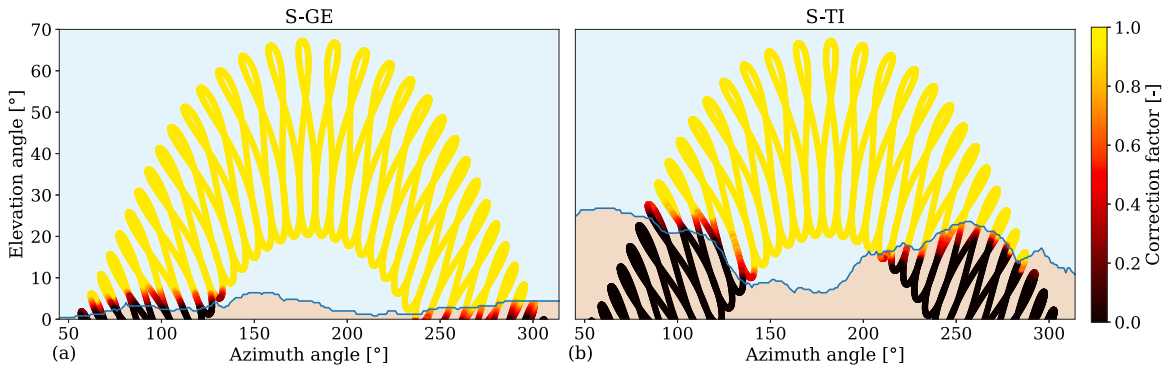


Fig. 4. Solar zenith and elevation angles computed over a full year at 30-minute resolution for two PV systems located in the vicinity of (a) Geneva and (b) Biasca. The correction factor represents the fraction of time since the previous timestamp during which the sun remains above the local horizon, ranging from 0 (entirely below the horizon) to 1 (entirely above the horizon). This figure was generated using the HORAYZON library [57].

SHADECast and IrradianceNet operate on upscaled CSI fields from HANNA over the region [4.41°E, 12.09°E] × [42.99°N, 50.67°N], corresponding to images of size 384 × 384 pixels. This domain covers a region centered around Switzerland, which consists of our area of interest. Given this domain size, SHADECast requires 6.5 s to generate an ensemble forecast with 10 members on a single NVIDIA GH200 GPU. Due to architectural constraints, IrradianceNet cannot process arbitrarily large inputs. Therefore, each CSI field is divided into nine patches of size 128 x 128 pixels, with linear interpolation applied along the borders of each patch to reconstruct the full field. This patching strategy, previously adopted by [30,31], allows IrradianceNet to be applied to arbitrarily large spatial domains. Consequently, a single SHADECast inference corresponds to nine separate inferences with IrradianceNet. The latter requires approximately 1.9 s to produce a deterministic forecast over a 128 × 128 pixel area on the same GPU model, resulting in a total inference time of about 17 s for the full domain.

SolarSTEPS employs optical-flow techniques to generate CSI forecasts, which can result in regions with missing values. This is an inherent limitation of optical-flow models, as they advect CSI values using estimated CMVs but do not extrapolate new information beyond the spatial coverage of the input data. As a result, areas advected from regions lacking upstream data remain undefined in the forecast outputs [27]. To mitigate this issue, SolarSTEPS and its simplified variant SolarSTEPS-pa use slightly larger input domains, defined over the region [3.13°E, 13.37°E] × [42.99°N, 50.67°N], corresponding to 512 × 384 pixels. On this input size, generating a forecast ensemble with 10 members with SolarSTEPS and SolarSTEPS-pa requires 18.0 and 6.5 s, respectively, on a single AMD EPYC 7742 CPU. The longer computational time attained by SolarSTEPS results from the additional calculations required to model cloud evolution.

The comparison between the satellite-based models and the NWP model IFS-ENS is performed by adopting a user-centric evaluation approach that reflects real-world forecasting needs. For example, suppose a user requests a forecast at 11:55 for the period starting at 12:00. In this case, the satellite-based models generate predictions for 12:00 to 13:45 using SSI observations from 11:00, 11:15, 11:30, and 11:45. For IFS-ENS, the most recent forecast available that covers the requested time window is selected based on its operational dissemination schedule [22,68]. In this context, the forecast initialized at 00:00 with a 12-hour lead time is compared with the satellite-based forecast for 12:00 (i.e., a 15-minute lead time), while the forecast initialized at 06:00 with a 7-hour lead time is compared with the satellite-based forecast for 13:00 (i.e., a 75-minute lead time). Consequently, although we will refer to the forecast as having a 15- and 75-minute lead time, this will correspond to a 7- to 12-hour lead time for the IFS-ENS model. This distinction also highlights the fundamental difference in forecast latency between the two approaches.

For each day, forecasts are generated for full-hour timestamps between 1 h after sunrise and 3 h before sunset. This process is repeated for the full two-year period, giving a total of 6158 inferences per model. We acknowledge that evaluating forecasts at an hourly resolution may slightly disadvantage satellite-based models, which can provide updates at a higher temporal frequency (e.g., every 15 min) in an operational setting. However, the choice to evaluate one forecast per hour was made to ensure a fair and consistent comparison with the IFS-ENS system, which operates at an hourly resolution. As each forecast is independent, we leverage high-throughput computing to run them in parallel, assigning one job per processor on a compute node. The computations were performed on the Swiss high-performance computing system Alps at the Swiss National Supercomputing Centre.

5. Results

We begin by evaluating the accuracy of the irradiance-to-power conversion models in Section 5.1. Next, Section 5.2 compares the SSI forecasts of all models using the HANNA SSI fields as reference. Finally, Section 5.3 presents the intraday PV power forecast evaluation based on operational data from the 6434 PV systems.

5.1. Irradiance-to-power conversion

An initial qualitative assessment of model performance is provided by comparing measured and predicted PV power output, together with the corresponding observed SSI, for stations S-GE and S-TI over three randomly selected test-set days. Fig. 5(a, d) shows results for a clear-sky day, characterized by a smooth, parabolic SSI profile. The S-GE station, situated in a relatively flat region (see Fig. 4(a)), exhibits no major impacts from the local horizon in its PV profile. In contrast, S-TI is located in a mountainous area where power output begins to ramp up only once the sun is visible above the local horizon. In both cases, the irradiance-to-power conversion model predicts the produced PV power remarkably well. Fig. 5(b, e) displays a day with moderate variability, where SSI ramps cause sudden drops in the power output of the two stations. The associated drops in PV power output are captured by the XGBoost model, although prediction errors are overall higher than in the clear-sky scenario. Finally, Fig. 5(c, f) shows a day with highly variable conditions characterized by intermittent cloud cover. Although the model captures the overall temporal evolution, the daily nMAE and nRMSE are considerably higher than in the low- and moderate variability conditions. In all cases, we find a strong positive correlation between SSI and PV power output. This highlights the importance of high-quality SSI observations and satellite retrievals for accurate irradiance-to-power conversions [69]. We remark that the conversion is applied only between sunrise and sunset, as PV power output is zero outside this period.

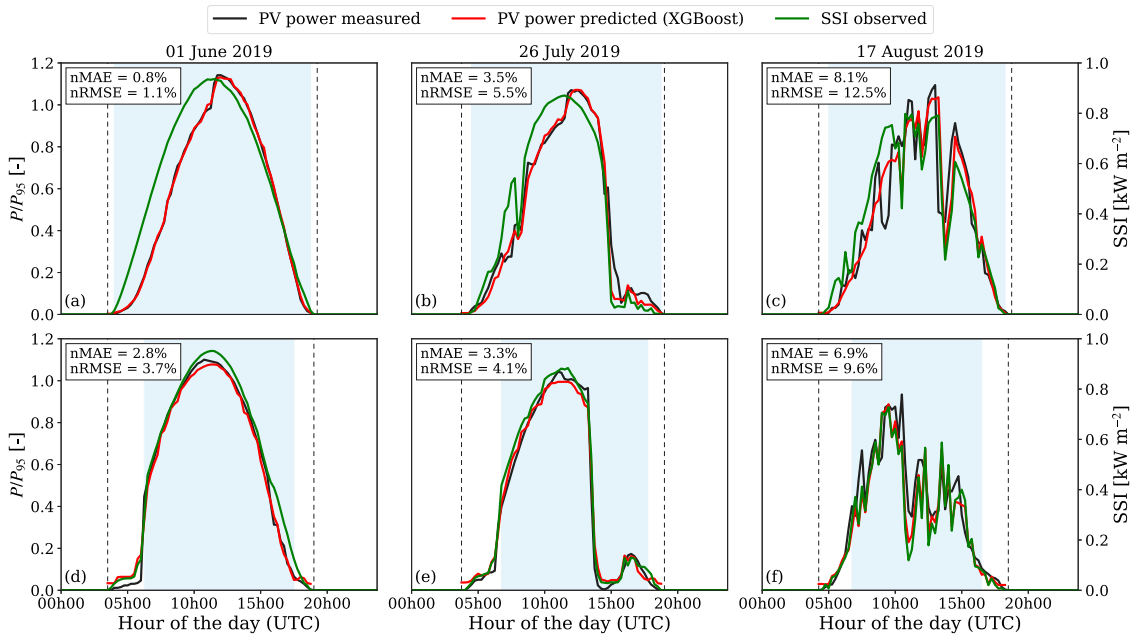


Fig. 5. PV power measurements, XGBoost predictions of PV production, and SSI observations at 15-minute resolution for stations S-GE (a–c) and S-TI (d–f) over three days that belong to the test set. SSI observations from the HANNA dataset are interpolated to the corresponding station locations. The PV power values are normalized with the station-specific P_{95} value. The light-blue shaded area highlights the period when the sun is above the horizon, determined from local SZA using HORAYZON. The vertical black dashed lines indicate local sunrise and sunset times. The nMAE and nRMSE values are computed for each individual day and normalized using the station-specific P_{95} value.

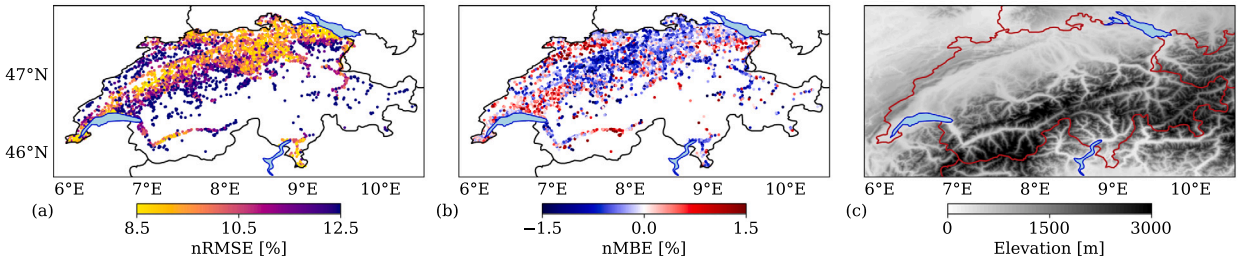


Fig. 6. (a) nRMSE and (b) nMBE of the predicted PV power, averaged over the test set. Both quantities are normalized with the station-specific P_{95} value. (c) Elevation map showing values in meters above sea level.

We trained a separate XGBoost model for each PV station. Fig. 6(a, b) shows the nRMSE and nMBE of the predicted PV power, averaged over the test set (i.e. 61 days) and normalized by the station-specific P_{95} value. Overall, nRMSE values range from 8%–10% in lowland regions, increasing to 11%–13% in mountainous areas, i.e., the Jura in the Northwest and the Alps in central and Southern Switzerland. Approximately 3.6% of stations exhibit nRMSE values above 15%. These stations are located at an average elevation of 873 m, compared to 514 m for the remaining 96.4% of stations. We suspect that one of the reasons for the reduced performance of the irradiance-to-power conversion at higher elevations is the lower accuracy of HANNA SSI estimates in mountainous terrain [38]. This highlights how a dense PV monitoring network can also serve as an effective means of evaluating SSI retrieval methods. Moreover, high-elevation PV systems typically experience more variable atmospheric conditions and frequent snow cover, both of which can introduce additional discrepancies between measured and predicted power output. Fig. 6(b) illustrates the nMBE. Most stations (65.7%) exhibit a negative nMBE, suggesting that the XGBoost models tend to underestimate PV production. This bias may be related to the use of RMSE-based loss minimization, which promotes regression toward the conditional mean of the target distribution and can therefore lead to a systematic underestimation of extreme PV production values [70,71]. In absolute terms, however, only 10.7% of stations show an nMBE larger than 1%.

Further analysis in Fig. 7 shows how nRMSE and nMBE relate to both P_{95} and elevation. As expected, the RMSE is strongly correlated with the P_{95} value, with a Pearson coefficient of $R = 0.99$. Likewise, nRMSE exhibits a positive correlation with elevation ($R = 0.55$). We expect that this correlation can be reduced through more accurate SSI observations in mountainous regions and complex terrain. A lower correlation between nMBE and elevation is observed, with a Pearson coefficient of $R = 0.26$.

Finally, Fig. 8 illustrates a 2D density histogram of predicted versus measured PV power over the test set, with both values scaled to the [0,1] range for consistency across stations. The test set covers 61 days, yielding roughly 17 million data points. The results indicate strong agreement between predictions and observations, as most points align closely with the 45° line and the linear regression fit shows minimal deviation. A slight overestimation appears at low power values (0–0.2), while at higher power levels (0.5–1) the models tend to underpredict. The symmetry and shape of the marginal histograms confirm a minor low bias and consistent spread between predicted and observed values. Overall, the accuracy of the predictions remains high, with average nMAE and nRMSE across all stations of 6.2% and 10.2%, respectively.

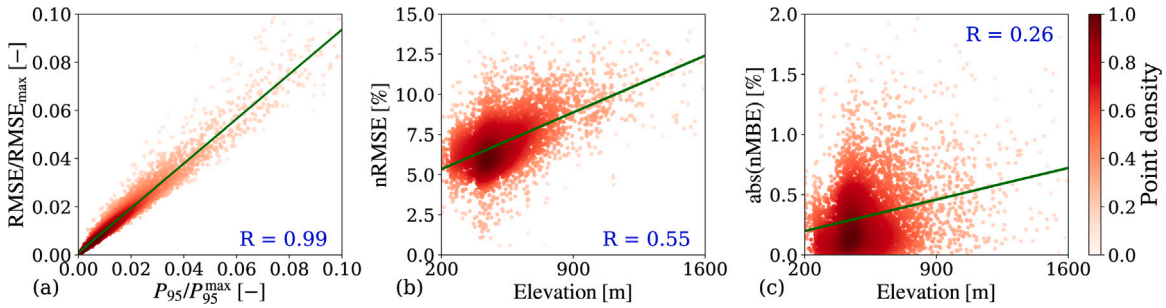


Fig. 7. Scatter plot of (a) RMSE as a function of the station-specific P_{95} value, with both variables normalized by their respective maximum values, and (b, c) nRMSE and nMBE as a function of elevation. The nRMSE and nMBE are averaged over the test set. The color indicates local data density estimated using a Gaussian kernel (nonlinear color normalization with $\gamma = 0.3$ is applied). The green dashed line indicates a linear fit, and the Pearson correlation coefficient R is reported in blue.

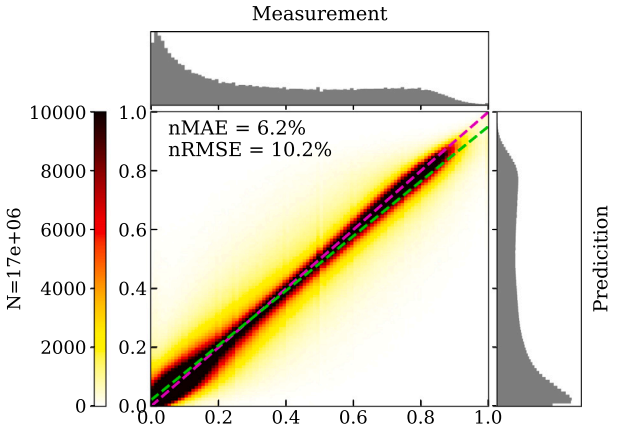


Fig. 8. Scatter plot of normalized PV power measurements (x-axis) versus XGBoost model predictions (y-axis) over the test set. For each station, the values are scaled to the [0, 1] range. The pink dashed line represents the 1:1 reference, while the light green dashed line shows the linear regression fit. The nMAE and nRMSE are computed as averages across the full test set and all stations, and are normalized with the station-specific P_{95} value. The marginal histograms above and at the side of the scatter plot display the distributions of measurements and predictions, respectively, using 100 bins. The plot includes approximately 17 million PV power measurements.

5.2. Performance evaluation of surface solar irradiance forecasts

The satellite-based models evaluated in this study operate at a spatial resolution four times finer than that of IFS-ENS. Therefore, to perform a grid benchmarking (i.e., comparing SSI forecasts at the pixel level over the study area), we first upscale HANNA and the satellite-based SSI inferences by aggregating them from 0.02° to 0.08°. Deterministic and probabilistic performance metrics are then computed at each grid point, using HANNA as ground truth, and averaged over all 6158 forecast instances. For probabilistic models, the RMSE, MAE and MBE are computed using the ensemble mean. For deterministic models, CRPS reduces to MAE, and thus the MAE is reported as the corresponding metric. Satellite-based forecasts are generated at the same temporal resolution as HANNA, allowing for a comparison based on instantaneous SSI fields. Hence, the forecasts have a temporal resolution of 15 min and a prediction horizon of 2 h. In contrast, IFS-ENS provides hourly SSI values averaged over the preceding hour. Therefore, comparisons with HANNA are performed using corresponding hourly averages. Further details of the metric definitions are provided in C.2.

The spatial comparison of forecast performance across the area of interest is presented in Fig. 9, with results averaged over all lead times and forecast instances. Among all models, IrradianceNet, which is the only deterministic model, achieves the lowest RMSE, corresponding to an average reduction of about 10.5% compared with the probabilistic ML and optical-flow models and 26.0% relative to IFS-ENS. However, IrradianceNet shows a comparatively high CRPS (note that CRPS reduces to MAE for deterministic models), while the probabilistic models tend to produce forecasts that are both more reliable and sharper. For instance, the spatially averaged CRPS of SHADECast is 17.7% lower than that of IrradianceNet. In other words, while the deterministic model excels in point accuracy, the probabilistic models provide more informative and better-calibrated uncertainty estimates.

Within the probabilistic models, SHADECast achieves the lowest CRPS, establishing it as the most reliable and sharp probabilistic SSI forecast model. While SolarSTEPS performs better in low-elevation regions, its performance deteriorates substantially over the Alpine arc, whereas SolarSTEPS-pa, which does not account for cloud evolution, exhibits a substantially weaker performance, with its domain-averaged CRPS being 34.7% higher than that of SolarSTEPS. All satellite-based models outperform IFS-ENS, which suffers from reduced performance in mountainous terrain. Notably, IFS-ENS-corrected achieves a 9.4% RMSE reduction and a 15.5% decrease in CRPS relative to IFS-ENS, with the largest improvements observed in high-elevation regions. This highlights the effectiveness of bias correction in enhancing the performance of physics-based probabilistic forecasts. Overall, most models exhibit a positive bias, indicating a general tendency to overestimate SSI, with MBE values reaching up to 45 W m^{-2}. This overestimation is especially pronounced at higher elevations, while biases at low elevations tend to be negative.

Fig. 10 provides a global overview of RMSE, CRPS and MBE spatially averaged over three different areas: the entire study area, low-elevation, and high-elevation areas. The distinction between low- and high-elevation regions is based on the median of the elevation distribution measured over all pixels, which is 790 m. We observe that SSI forecasts quality decreases in high-elevation areas. This is expected, as mountainous areas can be characterized by local meteorological processes and lower SSI satellite retrieval quality. Additionally, IFS-ENS operates at a limited spatial resolution, meaning that terrain features such as valleys, ridges, and slopes may be smaller than the model grid. As a result, important sub-grid effects, including terrain shading and slope orientation, are not fully captured, which can lead to reduced forecast accuracy in high-elevation regions.

To better understand how forecast performance varies throughout the day, Fig. 11 illustrates the diurnal patterns in forecast errors for satellite-based and NWP models, averaged over all forecasts and all pixels in the area of interest. The IFS-ENS forecasts exhibit an MAE that closely mirrors the diurnal SSI cycle. Forecast errors are small after sunrise, increase toward midday when irradiance is highest, and then decrease again as sunset approaches. The seasonal dependence is also significant, with MAE values being substantially higher in summer (JJA) than in winter (DJF). Satellite-based forecasts, available at hourly

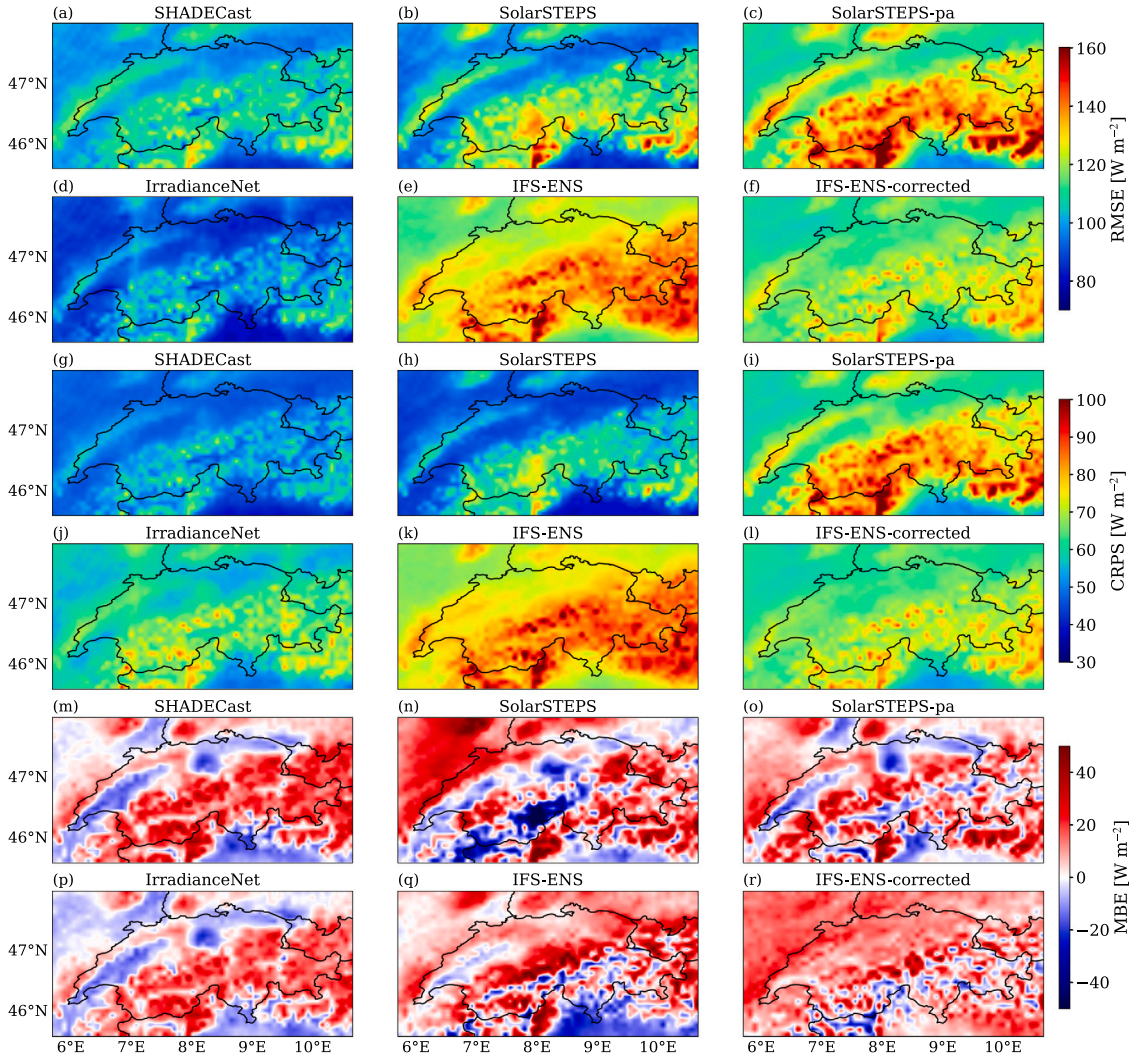


Fig. 9. (a–f) RMSE, (g–l) CRPS and (m–r) MBE computed at the pixel level over the study area for all models considered. Each metric is averaged over all inferences and lead times. The satellite-derived HANNA SSI serves as ground truth. Note that minor artifacts observed near the meridians at approximately 7° E and 9° E in the IrradianceNet fields arise from the patching procedure. The CRPS is replaced with MAE for IrradianceNet, the only deterministic model.

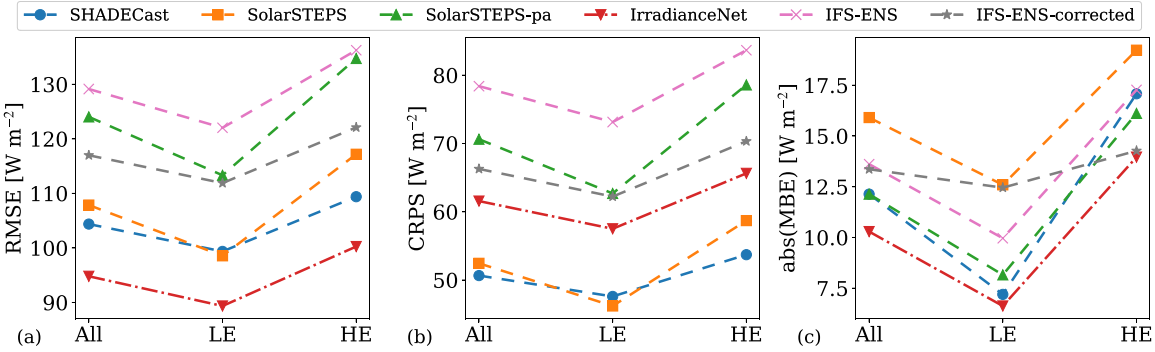


Fig. 10. (a) RMSE, (b) CRPS and (c) MBE averaged over all inferences and lead times. Results are additionally averaged over three regions of interest: the full area shown in Fig. 9 (All), pixels with mean elevation below 790 m (LE), and pixels with mean elevation above 790 m (HE). The elevation threshold was set to the median of the elevation distribution computed over all pixels. Dashed–dotted and dashed lines represent deterministic and probabilistic models, respectively. The CRPS is replaced with MAE for IrradianceNet, the only deterministic model.

intervals throughout the daylight period, show a different error pattern. At short lead times, SHADECast achieves significantly lower MAE than IFS-ENS because it relies directly on recent satellite observations and thus remains close to ground truth. However, its accuracy degrades as lead time increases, with MAE values often higher than those of IFS-ENS for lead times close to 2 h. In late-afternoon forecasts, the MAE naturally decreases for increasing lead times, as SSI approaches zero near sunset. Similar to the NWP model, the MAE exhibits a parabolic

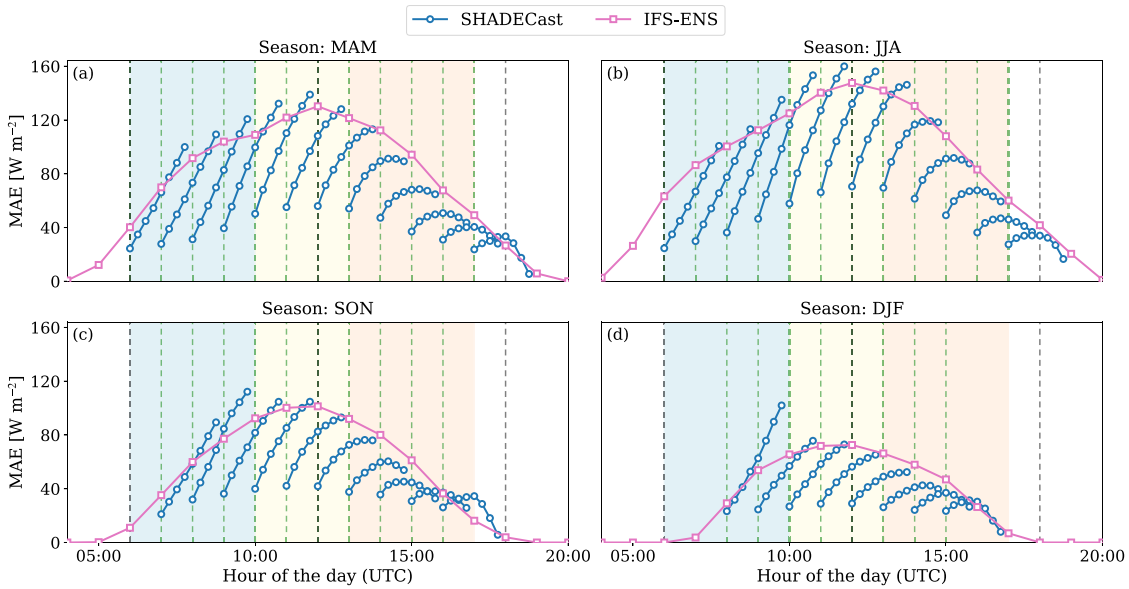


Fig. 11. Diurnal evolution of the MAE for SSI forecasts generated with SHADECast (satellite-based) and IFS-ENS (NWP), averaged across all pixels and forecast instances, during the months (a) MAM, (b) JJA, (c) SON and (d) DJF. The vertical dashed green and black lines mark the times at which satellite-based and NWP forecasts are issued, respectively. The light blue, yellow, and orange shaded regions mark the morning, midday, and afternoon periods, respectively.

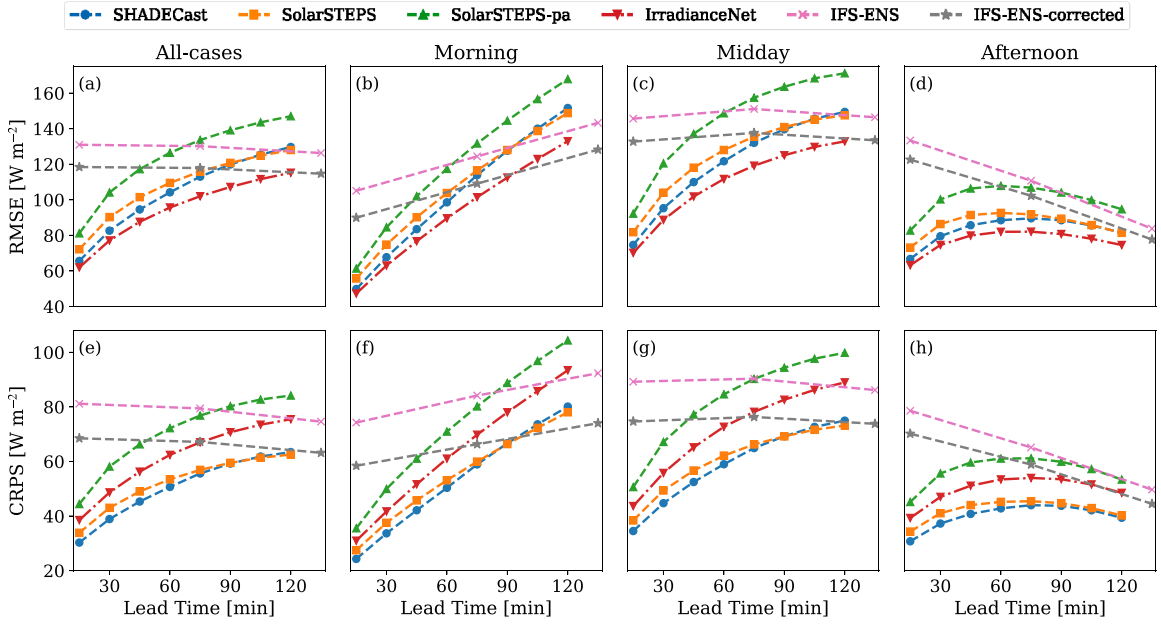


Fig. 12. (a–d) RMSE and (e–h) CRPS averaged across the study area and all forecasts for three different periods of the day: morning, midday and afternoon. Results averaged over all daylight hours (All-cases) are also included. The CRPS is replaced with MAE for IrradianceNet, the only deterministic model. Dashed–dotted and dashed lines represent deterministic and probabilistic models, respectively.

pattern over the course of the day, with the curvature becoming more pronounced as the lead time increases, and with larger errors observed during spring (MAM) and summer (JJA). For clarity, Fig. 11 reports only results for SHADECast among the satellite-based models. However, all satellite-driven approaches exhibit comparable behavior — see supplementary material.

Fig. 12 summarizes the evolution of RMSE and CRPS as a function of lead time, grouped into three periods of the day, i.e. morning, midday, and afternoon, as defined in Fig. 11. For satellite-based models, both RMSE and CRPS increase steadily during the morning and midday, reflecting the increase in SSI. In the afternoon, these metrics exhibit a parabolic-like behavior, due to the SSI approaching zero near sunset. The largest errors occur at midday, although the rate of error growth is steepest during the morning hours. IFS-ENS displays a different behavior, with RMSE and CRPS more closely following the diurnal cycle of SSI rather than showing a monotonic degradation with lead time. As a result, when averaging over all daylight hours, the resulting error profiles exhibit only a weak dependence on lead time. We note that the bias correction reduces the RMSE of about 9.4% in the all-cases scenario (i.e. averaged over all available forecasts), demonstrating its effectiveness in improving forecast accuracy. Overall, satellite-based models deliver superior accuracy at short lead times, but their performance degrades more rapidly compared with the NWP model as the forecast horizon extends.

To assess model performance under different weather regimes, we classify each day into four categories: cloudy, sunny, low-variability

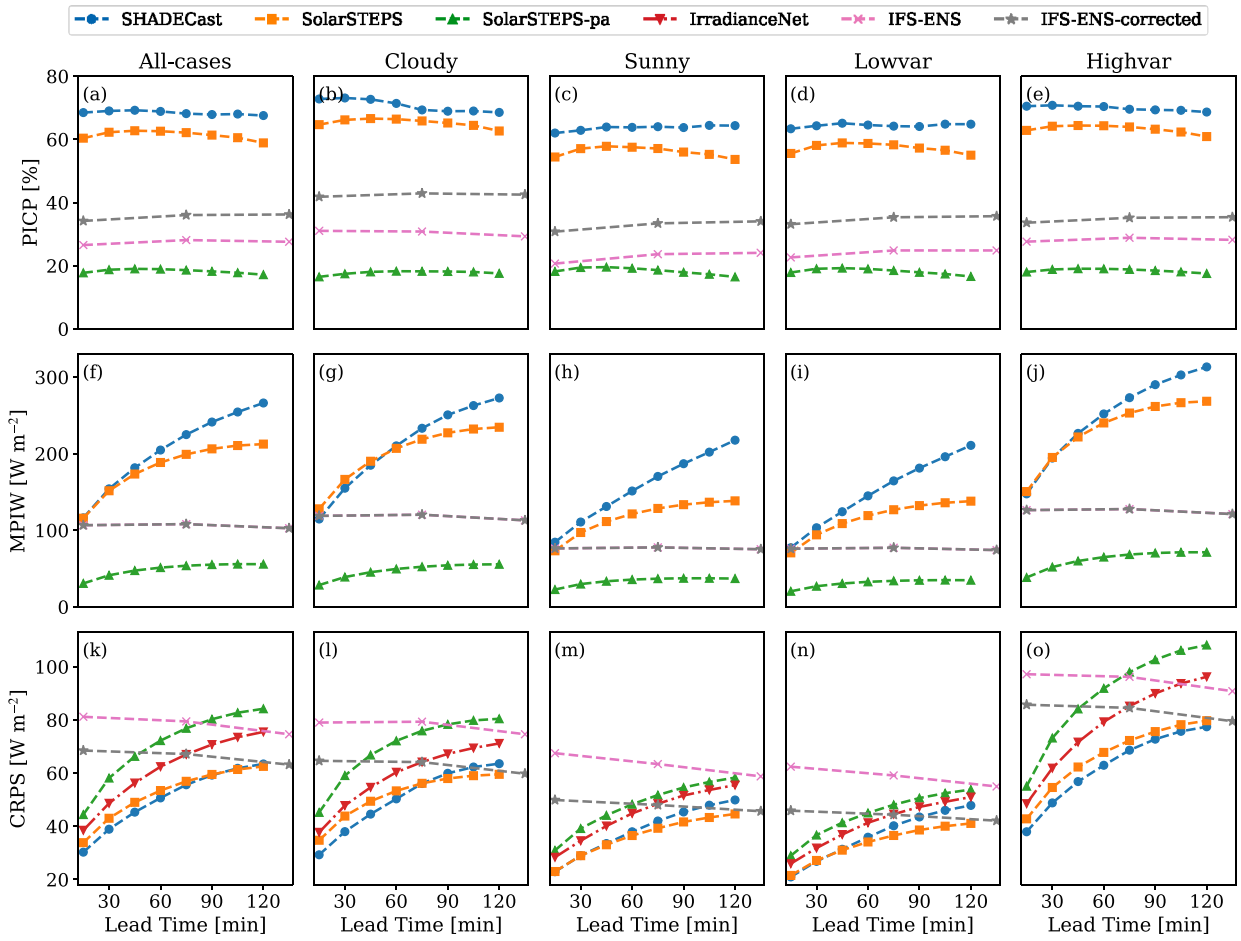


Fig. 13. (a–e) PICP, (f–j) MPIW and (k–o) CRPS averaged across the study area and all forecasts for five different weather scenarios: all-cases, cloudy, sunny, low-variability, and high-variability. The PICP and PINAW are only shown for probabilistic models. The CRPS is replaced with MAE for IrradianceNet, the only deterministic model. In panel (k–o), dashed–dotted and dashed lines represent deterministic and probabilistic models, respectively.

(lowvar), and high-variability (highvar). This classification is performed at the daily level, such that all inferences within a given day are assigned to the same category. Specifically, we compute the daily mean and standard deviation of the CSI fields used as input to the satellite-based models. From the resulting distributions across all days, we determine the 25th and 75th percentiles. Days with a mean CSI below the 25th percentile are classified as cloudy, while those above the 75th percentile are classified as sunny. Similarly, days with a standard deviation below the 25th percentile are classified as low-variability, whereas those above the 75th percentile are classified as high-variability. When performance metrics are averaged across all 6158 inferences, we refer to this case as all-cases.

Fig. 13(a, f, k) compares the reliability, sharpness, and overall probabilistic performance of all models as a function of lead time averaged over all available forecasts (all-cases). We observe that satellite-based models accounting for cloud evolution are the most reliable but less sharp, with PICP values reaching up to 70% for SHADECast and 60% for SolarSTEPS. Moreover, SHADECast exhibits the highest growth rate in the MPIW with increasing lead time, as observed by [31]. The absence of cloud evolution mechanisms in SolarSTEPS-pa results in an ensemble with limited spread, producing very low PICP and MPIW values at all lead times. These findings are consistent with [27,31]. Bias correction applied to the IFS-ENS forecasts increases the PICP by 29.4% compared to the non-corrected version, while the PINAW remains unchanged since the ensemble distribution is unaffected by the bias correction — see Appendix D. For lead times below 1 h, SHADECast achieves a 8.2% lower CRPS than SolarSTEPS, making it the most accurate model. At lead times near 2 h, SHADECast and SolarSTEPS show similar forecast accuracy.

Fig. 13 also illustrates results obtained across different weather scenarios. Although the absolute magnitude of the error metrics varies with atmospheric conditions, the relative ranking of the models remains consistent. The PICP exhibits only minor sensitivity to weather type, whereas both the mean prediction interval width and the CRPS vary significantly. For example, under cloudy conditions, the CRPS is about 50% higher than in sunny conditions. Similarly, high-variability conditions are more challenging than low-variability conditions, therefore showing less accurate forecasts. Notably, SHADECast demonstrates superior performance in high-variability conditions, whereas SolarSTEPS performs better in the low-variability regime. This difference may arise from SolarSTEPS inability to generate accurate forecasts in the presence of time-varying cloudiness distributions, a limitation of its underlying linear AR model [31]. A summary of the model comparison across metrics and conditions is presented in Tables 1 and 2.

5.3. Performance evaluation of PV power forecasts

First, SSI forecasts are converted into PV power using the procedure described in Section 5.1. For each forecast, lead time, and ensemble member, SSI values are interpolated to the locations of the PV stations and then used as inputs to the station-specific XGBoost models. Since we have a separate XGBoost model per PV station, this setup amounts to roughly 16×10^9 inferences, enabled by the computational efficiency of the XGBoost library and the use of multiprocessing. The resulting

Table 1
Comparison of forecasting models for SSI across six deterministic and probabilistic metrics. Values are averaged over all inferences and lead times. Results are additionally averaged over three regions of interest: the full area shown in Fig. 9 (All), pixels with mean elevation below 790 m (LE), and pixels with mean elevation above 790 m (HE). For each metric, the best-performing model is highlighted in bold while missing values are indicated by "–". Moreover, the CRPS is replaced with the MAE for IrradianceNet, the only deterministic model.

Model	MAE [W m^{-2}]			RMSE [W m^{-2}]			abs(MBE) [W m^{-2}]			PICP [%]			MPIW [W m^{-2}]			CRPS [W m^{-2}]		
	All	LE	HE	All	LE	HE	All	LE	HE	All	LE	HE	All	LE	HE	All	LE	HE
SHADECast	72.0	69.0	75.0	104.4	99.4	109.4	12.1	7.2	17.1	**68.4**	**66.3**	**70.5**	205.5	194.2	216.7	**50.7**	47.6	**53.7**
SolarSTEPS	70.7	62.7	78.8	107.9	98.6	117.2	15.9	12.6	19.2	61.4	56.4	66.3	182.3	149.6	214.9	52.4	**46.2**	58.7
SolarSTEPS-pa	78.7	69.5	87.9	124.0	113.3	134.7	12.1	8.2	16.1	18.3	16.7	19.9	**48.9**	**40.7**	**57.0**	70.6	62.7	78.6
IrradianceNet	**61.6**	**57.5**	**65.6**	**94.8**	**89.4**	**100.3**	**10.3**	**6.6**	**14.0**	–	–	–	–	–	–	61.6	57.5	65.6
IFS-ENS	94.3	88.6	100.0	129.1	122.0	136.3	13.6	10.0	17.3	27.4	27.9	26.9	105.8	103.6	107.9	78.4	73.1	83.7
IFS-ENS-corr	80.4	75.7	85.1	117.0	111.9	122.1	13.4	12.5	14.3	35.5	36.9	34.1	105.8	103.6	107.9	66.3	62.2	70.3

Table 2
Comparison of forecasting models for SSI using the CRPS, evaluated across different times of day and weather conditions. Results are averaged over all forecast instances and across the domain of interest. Performance is reported for the first (LT_F = 15 min) and last (LT_L = 120 min) lead times. Bold values indicate the best-performing model (i.e., lowest CRPS). Moreover, the CRPS is replaced with the MAE for IrradianceNet, the only deterministic model.

Model	All-cases		Morning		Midday		Afternoon		Cloudy		Sunny		Lowvar		Highvar	
	LT_F	LT_L	LT_F	LT_L	LT_F	LT_L	LT_F	LT_L	LT_F	LT_L	LT_F	LT_L	LT_F	LT_L	LT_F	LT_L
SHADECast	**30.3**	63.4	**24.3**	80.1	**34.5**	75.0	**30.7**	**39.4**	**29.3**	63.5	**22.9**	49.9	**21.0**	47.8	**37.9**	**77.4**
SolarSTEPS	33.8	**62.5**	27.5	78.0	38.4	**73.2**	34.3	40.2	34.8	**59.5**	23.0	**44.6**	21.5	**41.1**	42.8	79.7
SolarSTEPS-pa	44.4	84.2	35.5	104.4	50.6	99.8	45.2	53.4	45.2	80.4	30.9	58.3	28.9	53.8	55.1	108.1
IrradianceNet	38.5	75.5	30.9	93.5	43.7	89.0	39.3	48.5	37.7	71.2	28.4	55.5	25.9	51.0	48.5	96.2
IFS-ENS	81.2	74.6	74.3	92.3	89.2	86.2	78.6	49.7	79.0	74.6	67.5	58.8	62.4	55.0	97.2	90.8
IFS-ENS-corr	68.5	63.2	58.4	**74.0**	74.7	73.8	70.2	44.5	64.6	59.8	49.9	45.7	45.9	42.1	85.7	79.5

PV power forecasts are then validated against the operational dataset introduced in Section 2.2. Since the PV systems have significantly different nominal capacities, the error metrics are normalized with the station-specific P_{95} value. Further details on the evaluation metrics are provided in C.2. We remark that the satellite-based forecasts have a temporal resolution of 15 min and a prediction horizon of 2 h.

Fig. 14 shows the nRMSE, nCRPS, and nMBE of the power forecasts, evaluated against the operational dataset. Results are averaged over all weather conditions (All-cases scenario) and across all lead times. Since the accuracy of the PV forecast strongly depends on the quality of the underlying SSI forecasts, the spatial patterns of biases and uncertainties are similar to those in Fig. 9. In terms of nRMSE, IrradianceNet performs best, with a station-averaged nRMSE 6.1% lower than that of SHADECast. SHADECast and SolarSTEPS show similar performance overall, although SolarSTEPS achieves lower nRMSE in low-elevation regions. In contrast, the NWP model performs substantially worse. This shortcoming is also due to its spatial resolution, which is four times coarser than that of satellite-based models, underlining the importance of high-resolution data for site-specific comparisons.

A similar spatial pattern is found for the nCRPS, with SHADECast and SolarSTEPS providing the most accurate probabilistic forecasts. The absence of cloud evolution in SolarSTEPS-pa results in a station-averaged nCRPS that is 26.4% higher than that of SolarSTEPS. IFS-ENS performs worse overall, yielding a station-averaged nCRPS that is 31.0% higher than that of SHADECast. The pattern observed in the nMBE remains consistent with the one observed for the SSI forecasts, with PV power output underestimated mostly in the low-elevation regions (Fig. 9(m–r)).

Overall, Fig. 14 shows that PV forecast quality and accuracy decrease in high-elevation regions. To further examine this behavior, we divided the stations into two groups based on the elevation threshold defined in the previous section (790 m) and computed station-averaged results for each group. Fig. 15 shows the outcomes for three key metrics. Across all models, nRMSE and nCRPS are 11.8% and 13.6% higher in high-elevation regions compared to low-elevation regions, respectively. This decline in performance can be attributed to the lower accuracy of SSI forecasts in high-elevation areas together with the higher variability in the weather conditions, as discussed in Section 5.2. Additionally, the presence of snow can significantly alter PV power generation [72], an effect currently not captured by forecast models.

The PV power forecast errors divided into the three periods of the day are shown in Fig. 16. The generated PV power has a very strong positive correlation with SSI and therefore follows the same diurnal cycle. Hence, the pattern observed is very similar to the one shown in Fig. 12. Satellite-based models provide the most accurate forecasts at short lead times. This behavior is particularly pronounced during the morning and midday hours, when rapidly changing irradiance conditions cause RMSE and CRPS to increase sharply with lead time for satellite-based models. Next, Fig. 17 shows the station-averaged reliability, sharpness, and overall probabilistic performance of all models as a function of lead time across the different weather scenarios. In all scenarios, SHADECast remains the most reliable model, but it is also the least sharp, as its ensemble spread grows faster than that of the other models. In terms of nCRPS, SHADECast and SolarSTEPS perform very similarly for lead times below 1 h, while SolarSTEPS achieves the best performance at longer horizons, particularly in cloudy conditions. The NWP model shows a 50% higher nCRPS than SHADECast at the 15-minute lead time in the all-cases scenario, although this gap decreases with increasing lead time. When evaluating performance across different weather types, a similar pattern as in Fig. 13 emerges, that is, cloudy and high-variability conditions remain the most challenging for PV power forecasting. A summary of the model comparison across metrics and conditions is presented in Tables 3 and 4.

We also evaluated the ensemble forecast performance through rank histograms as shown in Fig. 18. These histograms illustrate how often the observations fall into each rank bin when compared to the sorted ensemble forecasts. For a perfectly reliable ensemble, the distribution should be uniform, indicating that observations are equally likely to fall anywhere within the ensemble spread. However, all models exhibit a U-shaped rank histogram. This indicates that the ensemble spread is too narrow relative to the actual variability of the observations, i.e., the forecasts are underdispersive or overconfident. Consistent with [31], SHADECast emerges as the least underdispersive model, followed by SolarSTEPS. By contrast, the NWP model is highly underdispersive, with around 75% of the measurements lying outside of the ensemble spread.

Finally, we evaluate daily total PV production by summing the predicted power across all installations at the shortest available lead time (15 min for satellite-based models) and comparing it with the

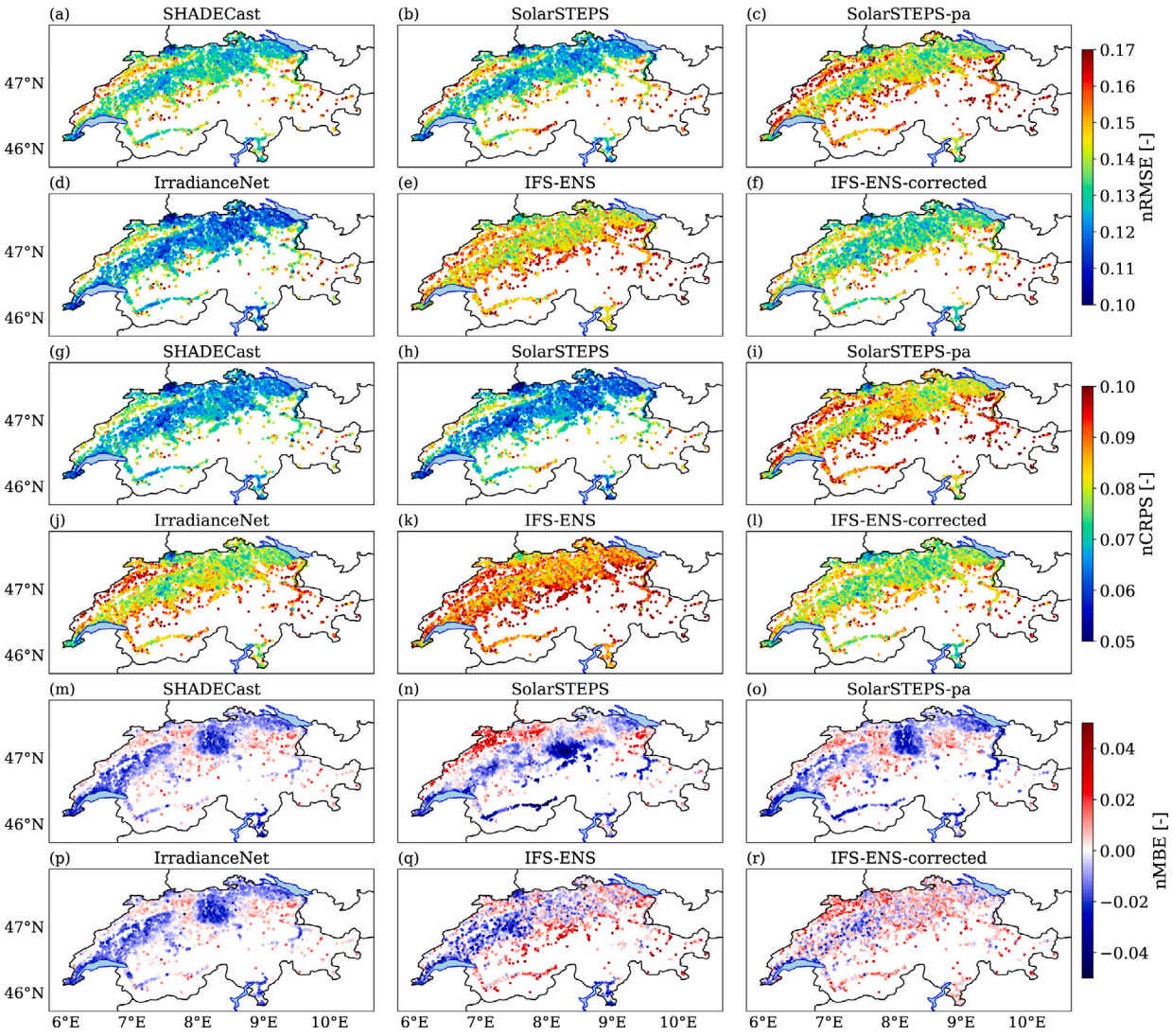


Fig. 14. (a–f) nRMSE, (g–l) nCRPS and (m–r) nMBE computed for each PV system with all models considered. Each metric is averaged over all inferences and lead times. Operational PV production data are used as ground truth. The nCRPS is replaced with nMAE for IrradianceNet, the only deterministic model.

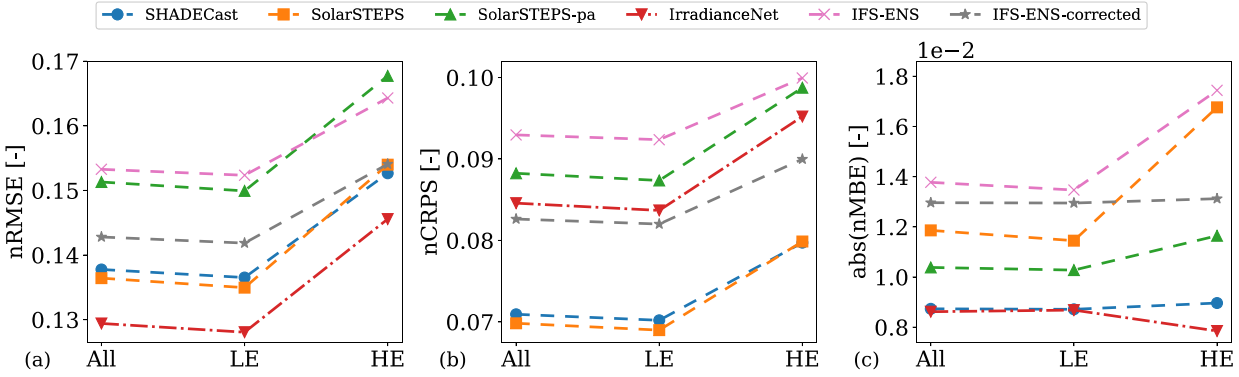


Fig. 15. (a) nRMSE, (b) nCRPS and (c) nMBE averaged across all inferences and lead times. Results are further averaged across three groups: all PV systems (All), stations below 790 m elevation (LE), and stations above 790 m elevation (HE). The elevation threshold was set to the median of the elevation distribution computed over all pixels. Dashed–dotted and dashed lines represent deterministic and probabilistic models, respectively. Moreover, the nCRPS is replaced with nMAE for IrradianceNet, the only deterministic model.

measured daily output. Applying this procedure to all days in the two-year period yields 726 samples, that is, one relative error value per day. The resulting distributions for each model are shown in Fig. 19. At this lead time, satellite-based models exhibit similar performance and clearly outperform IFS-ENS. The lowest errors occur in summer (JJA) while winter (DJF) shows the highest errors, likely driven by

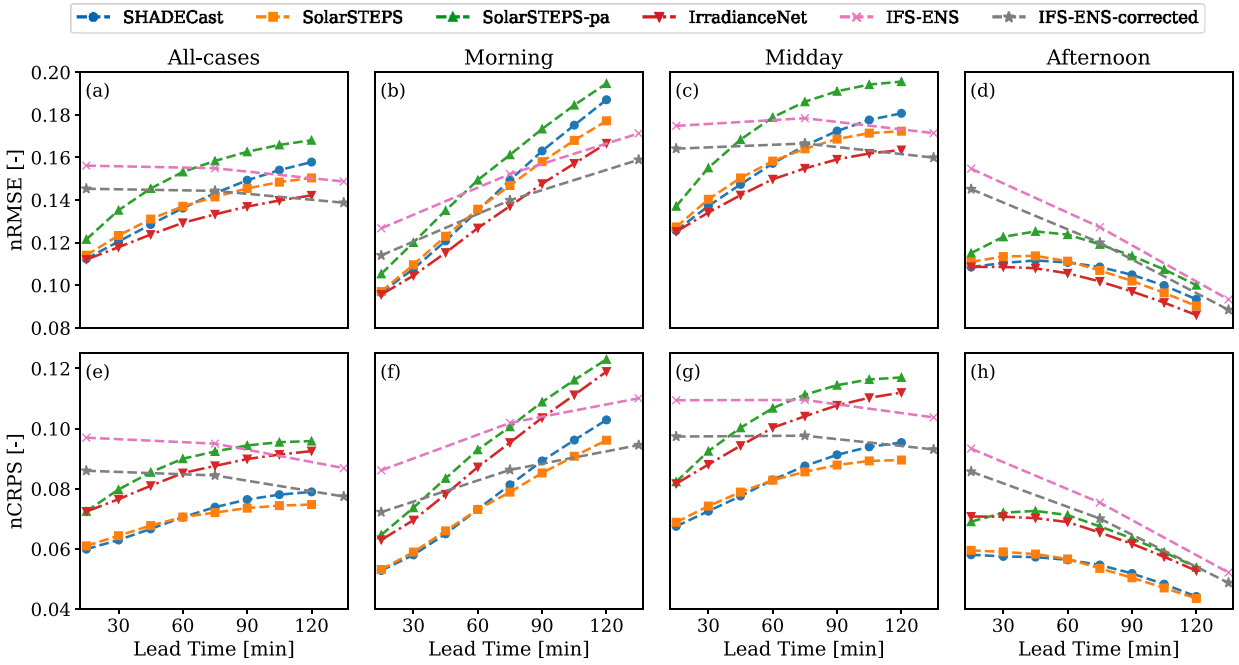


Fig. 16. (a–d) nRMSE and (e–h) nCRPS averaged across the study area and all forecasts for three different periods of the day: morning, midday and afternoon. Results averaged over all daylight hours (All-cases) are also included. The nCRPS is replaced with nMAE for IrradianceNet, the only deterministic model. Dashed–dotted and dashed lines represent deterministic and probabilistic models, respectively.

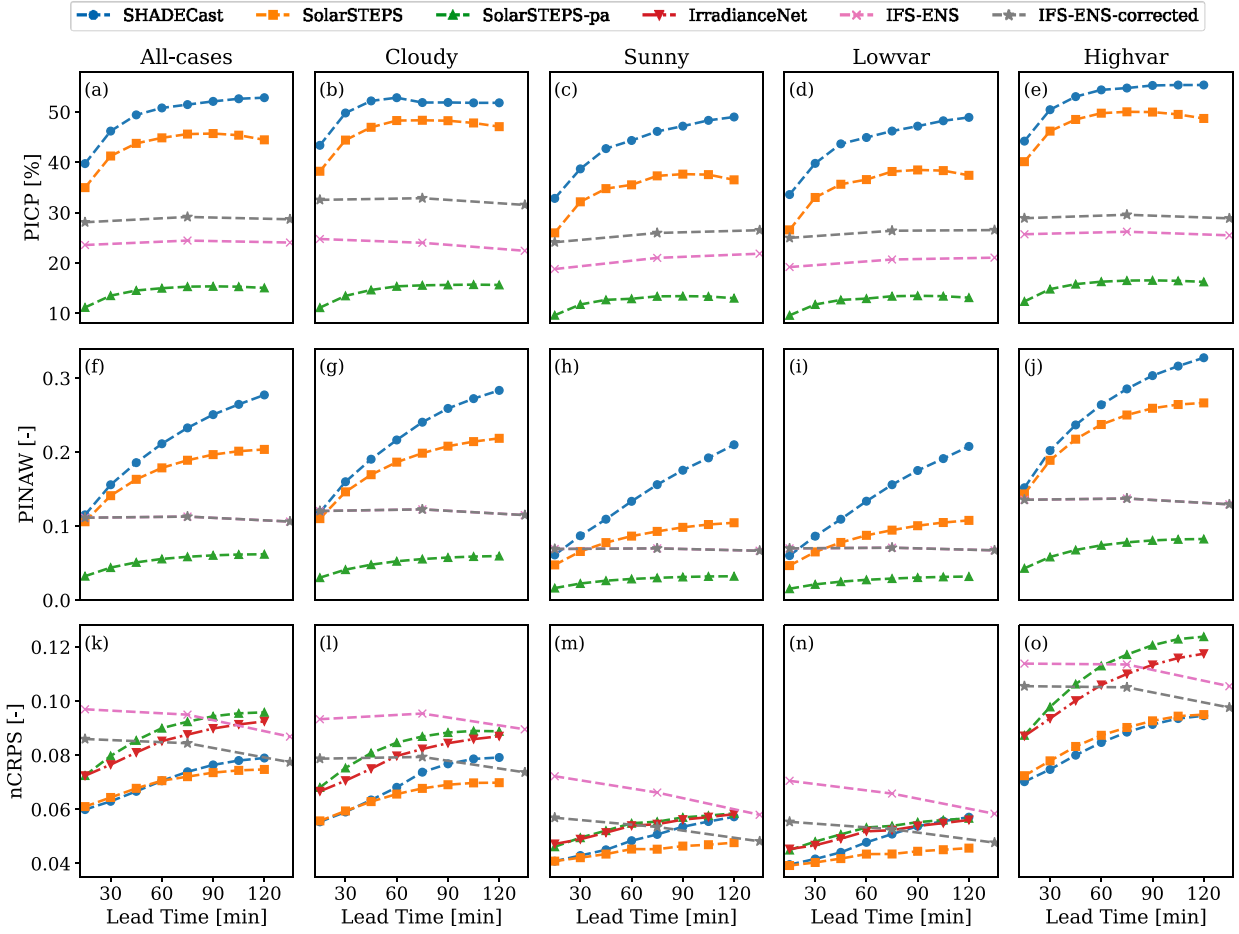


Fig. 17. (a–e) PICP, (f–j) PINAW and (k–o) nCRPS averaged across all PV systems and all forecasts for five different weather scenarios: all-cases, cloudy, sunny, low-variability, and high-variability. The PICP and PINAW are only shown for probabilistic models. The nCRPS is replaced with nMAE for IrradianceNet, the only deterministic model. In panel (k–o), dashed–dotted and dashed lines represent deterministic and probabilistic models, respectively.

Table 3
Comparison of forecasting models for PV power across six deterministic and probabilistic metrics. Values are averaged over all inferences and lead times. Results are further averaged across three groups: all PV systems (All), PV stations below 790 m elevation (LE), and PV stations above 790 m elevation (HE). For each metric, the best-performing model is highlighted in bold while missing values are indicated by "–". Moreover, the nCRPS is replaced with the nMAE for IrradianceNet, the only deterministic model. For clarity, metric values are scaled by powers of ten, as indicated in the column headers.

Model	nMAE [×10]			nRMSE [×10]			abs(nMBE) [×100]			PICP [%]			PINAW [×10]			nCRPS [×10]		
	All	LE	HE	All	LE	HE	All	LE	HE	All	LE	HE	All	LE	HE	All	LE	HE
SHADECast	0.96	0.96	1.07	1.38	1.37	1.53	0.87	**0.87**	0.90	**49.41**	**49.58**	**47.27**	2.12	2.11	2.19	0.71	0.70	**0.80**
SolarSTEPS	0.90	0.89	1.03	1.36	1.35	1.54	1.19	1.14	1.68	43.26	43.27	43.14	1.73	1.71	1.92	**0.70**	**0.69**	0.80
SolarSTEPS-pa	0.97	0.96	1.09	1.51	1.50	1.68	1.04	1.03	1.16	14.40	14.34	15.17	**0.53**	**0.53**	**0.60**	0.88	0.87	0.99
IrradianceNet	**0.85**	**0.84**	**0.95**	**1.29**	**1.28**	**1.46**	**0.86**	0.87	**0.78**	–	–	–	–	–	–	0.85	0.84	0.95
IFS-ENS	1.10	1.09	1.17	1.53	1.52	1.64	1.38	1.35	1.74	24.03	24.11	23.09	1.10	1.10	1.11	0.93	0.92	1.00
IFS-ENS-corr	0.98	0.97	1.06	1.43	1.42	1.54	1.30	1.29	1.31	28.64	28.80	26.76	1.08	1.08	1.07	0.83	0.82	0.90

Table 4
Comparison of forecasting models for PV power using the nCRPS, evaluated across different times of day and weather conditions. Results are averaged over all forecast instances and across all PV stations. Performance is reported for the first (LT_F = 15 min) and last (LT_L = 120 min) lead times. Bold values indicate the best-performing model (i.e., lowest nCRPS). Moreover, the nCRPS is replaced with the nMAE for IrradianceNet, the only deterministic model. For clarity, nCRPS values are multiplied by a factor of 10.

Model	All-cases		Morning		Midday		Afternoon		Cloudy		Sunny		Lowvar		Highvar	
	LT_F	LT_L	LT_F	LT_L	LT_F	LT_L	LT_F	LT_L	LT_F	LT_L	LT_F	LT_L	LT_F	LT_L	LT_F	LT_L
SHADECast	**0.60**	0.79	**0.53**	1.03	**0.68**	0.95	**0.58**	**0.44**	**0.55**	0.79	**0.41**	0.57	0.40	0.57	**0.70**	0.95
SolarSTEPS	0.61	**0.75**	0.53	**0.96**	0.69	**0.90**	0.60	0.44	0.56	**0.70**	0.41	**0.48**	**0.39**	**0.46**	0.72	**0.95**
SolarSTEPS-pa	0.73	0.96	0.65	1.23	0.82	1.17	0.69	0.54	0.68	0.89	0.46	0.59	0.45	0.57	0.87	1.24
IrradianceNet	0.72	0.93	0.63	1.19	0.82	1.12	0.71	0.53	0.67	0.87	0.47	0.58	0.45	0.56	0.87	1.18
IFS-ENS	0.97	0.87	0.86	1.10	1.09	1.04	0.93	0.52	0.93	0.90	0.72	0.58	0.71	0.58	1.14	1.06
IFS-ENS-corr	0.86	0.77	0.72	0.96	0.97	0.93	0.86	0.49	0.79	0.74	0.57	0.48	0.55	0.48	1.06	0.98

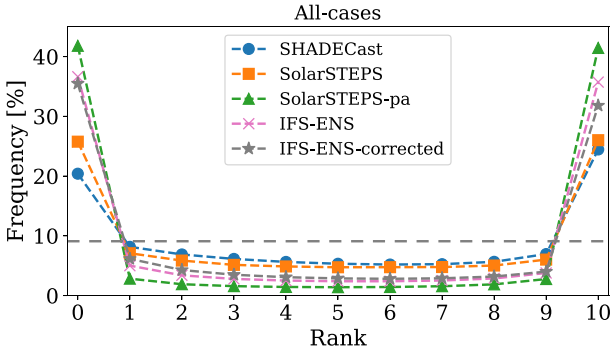


Fig. 18. Rank histogram of the probabilistic models, computed over all times in the test set (all-cases scenario) and across all lead times. The horizontal dashed gray line indicates perfect reliability, i.e. uniform distribution.

increased cloudiness, higher variability, and the presence of snow. It is important to note that the reported errors reflect both forecasting inaccuracies and uncertainties in the irradiance-to-power conversion process. Nevertheless, the median relative error for the satellite-based models remains around 2.2% during the MAM and JJA months, with only minor differences among the models. Moreover, the relative difference between measured and predicted national PV power remains below 1% for 18% of the 726 days analyzed, and below 10% for 82% of the days for all satellite-based models.

6. Conclusions

We presented a novel framework for spatiotemporal PV power forecasting and applied it to evaluate the reliability, sharpness, and overall performance of six PV forecast models. The models rely on a diverse set of SSI forecast approaches: SHADECast and IrradianceNet, two ML-based models, SolarSTEPS, an optical flow model, and IFS-ENS, a physics-based NWP model. While IFS-ENS is a classic NWP model, SHADECast, SolarSTEPS, and IrradianceNet solely rely on satellite observations. For each model, we first generated SSI forecasts for lead times of up to two hours, then used station-specific ML models to convert SSI into PV power, and finally assessed PV forecast performance against measurements from 6434 PV installations across Switzerland. To the best of our knowledge, this work introduces a novel spatiotemporal PV forecast framework and the first demonstration of PV forecasting on a large-scale countrywide PV network.

The decision to train a separate XGBoost model for each station was motivated by the partly complex topography of the study area. In addition, differences in panel orientation and inclination, as well as mounting configuration, across PV stations introduce site-specific patterns that an ML-based model can effectively capture. Given the strong correlation between SSI and PV production, we found that the accuracy of the irradiance-to-power conversion depends strongly on the quality of the underlying SSI measurements. Consequently, the availability of a dense PV network distributed over a wide area also provides a basis for evaluating the performance of SSI retrieval methods. For example, the HANNA SSI fields are known to have reduced accuracy at higher elevations. Accordingly, we observed a clear correlation between the nRMSE and the stations elevations, with a Pearson coefficient of R = 0.55. The irradiance-to-power conversion models achieved an average nMAE of 6.2% despite the limited set of input features.

The forecast accuracy of all models was evaluated against HANNA for SSI and against the operational dataset for PV power output. Satellite-based models outperformed the physics-based IFS-ENS model, particularly at short lead times. Two main factors contribute to this behavior. First, satellite-based models directly rely on recent satellite observations and thus remain close to ground truth, while the NWP model forecast error follows the diurnal cycle of SSI and PV power generation. Second, the satellite-based models operate at a spatial resolution four times finer than IFS-ENS, which proves particularly advantageous when evaluating site-specific performance. The application of bias correction improved the accuracy of IFS-ENS forecasts, highlighting its effectiveness in enhancing overall forecast performance.

We found that SHADECast outperforms SolarSTEPS in SSI forecasting at short lead times and in high-elevation regions, while SolarSTEPS achieves marginally better performance than SHADECast in low-elevation regions. For PV power generation, the two models show comparable performance, although SHADECast produces a more consistent ensemble spread and demonstrates higher reliability. Furthermore,

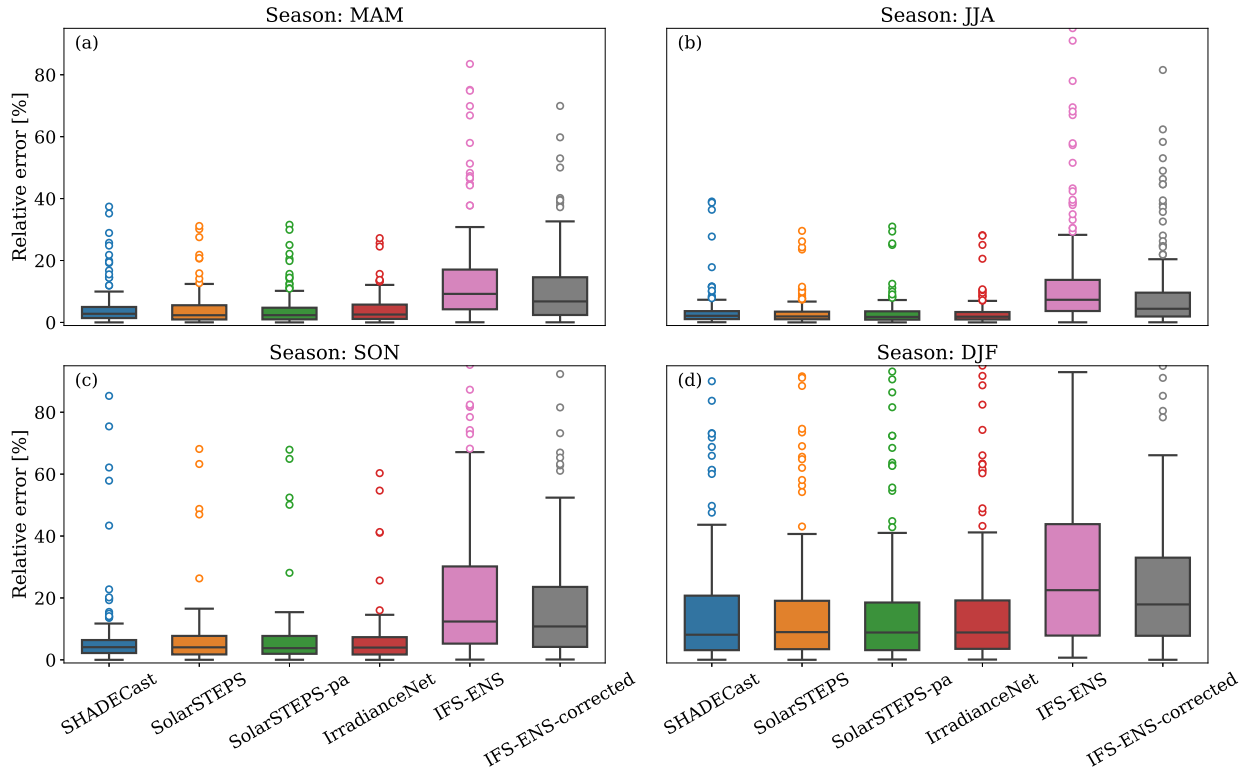


Fig. 19. Distribution of the relative error between measured and predicted total (i.e summed over all stations) PV power during the months (a) MAM, (b) JJA, (c) SON and (d) DJF, shown for all models at the 15-minute lead time. To improve readability, the y-axis range excludes a small number of outliers.

the absence of cloud evolution in SolarSTEPS-pa results in a significant drop in performance compared to SolarSTEPS. These findings emphasize the importance of models that account not only for cloud advection but also for cloud evolution. The deterministic model IrradianceNet consistently achieves the lowest RMSE. However, its MAE is higher than the CRPS of the probabilistic satellite-based models. In other words, the deterministic model excels in point accuracy, but the probabilistic models, despite having somewhat higher RMSE, offer better-calibrated uncertainty estimates.

For satellite-based models, forecast errors rise steadily during the morning and midday hours, reflecting the increasing SSI. In the afternoon, errors follow a parabolic pattern as SSI declines toward zero near sunset. The largest errors occur at midday, while the steepest rate of error growth is observed in the morning. In contrast, errors in NWP model forecasts more closely follow the diurnal SSI cycle. Among the four types of weather conditions examined, cloudy and high-variability conditions remain the most challenging for SSI and PV power forecasting, while sunny and low-variability conditions yield substantially better forecast performance. We also found that high-elevation regions exhibit lower forecast accuracy than low-elevation regions for both SSI and PV power. In addition to terrain effects and Alpine meteorological processes, snow cover can further complicate PV predictions in Alpine regions by modifying surface reflectivity, irradiance, and potentially partially covering PV installations.

At the country level, the total PV power predicted by satellite-based models closely aligns with observations, with relative errors below 10% for 82% of the 726 days considered. Relative errors are lowest during summer months, while winter conditions, characterized by higher cloud variability and snow cover, result in larger discrepancies.

Future research should investigate novel data aggregation strategies, for instance, at the level of power grid zones. Further, a comparison with regional higher-resolution model forecasts, such as from the Icosahedral Non-hydrostatic (ICON) model operated by MeteoSwiss for the area of this study, would also be of interest, yet forecasts and regional reanalysis were not made openly available. The growing share of solar energy in the power grid increases the need for more accurate forecasts to mitigate operational challenges such as grid instability and reserve power management. To address this, new probabilistic ML-based spatiotemporal PV forecast models can be investigated, with the goal of further reducing short-term grid imbalances and enhancing overall system reliability.

CRediT authorship contribution statement

Luca Lanzilao: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Angela Meyer:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used ChatGPT in order to help with writing style. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We acknowledge funding from the Swiss National Science Foundation, Switzerland (grant 200654). We thank Mathieu Schaer and Christian Steger of MeteoSwiss for the support to account for topographic shading using the HORAYZON library. We also thank Anke Tetzlaff and Uwe Pfeifroth of MeteoSwiss and the German Weather

Service, respectively, for compiling and providing the HANNA dataset. Finally, we would like to thank the anonymous reviewers for their valuable comments and suggestions, which helped improve the quality of this manuscript.

Appendix A. SSI and PV power forecasts

Fig. 20 compares the spatial patterns of SSI over the area of interest at lead times ranging from 15 to 120 min for the forecast models considered in this study. The forecasts were issued at 12:00 UTC on 6 August 2019. The satellite observations (including both SARAH and HANNA) show pronounced spatial variability, with significant convective cloud development over central Switzerland, localized high-irradiance regions, and sharp gradients associated with cloud structures. Data-driven models capture the observed spatial patterns and their temporal evolution considerably well for lead times up to 60 min, with performance declining at longer horizons. In contrast, the NWP-based forecasts (IFS-ENS and IFS-ENS-corrected) produce smoother, more homogeneous SSI fields, reflecting the coarser spatial resolution of the model. Overall, the figure illustrates a gradual loss of spatial detail as the forecast horizon increases, highlighting the advantage of high-resolution and data-driven nowcasting for short-term SSI prediction.

Fig. 21 shows the corresponding PV power forecasts. The top row presents observed PV power, revealing strong spatial heterogeneity and a well-defined cloud front advected from west to east, resulting in coherent regions of low and high production across Switzerland. All forecast models generally reproduce the observed SSI patterns, with low-irradiance areas corresponding to low PV power generation and vice versa. We note that the selected day is classified as high-variability weather, a scenario in which models show the lowest accuracy. In the supplementary material, we show additional cases.

Appendix B. Surface solar irradiance datasets

The accuracy of SSI observations is crucial for training irradiance-to-power conversion models and for providing ground truth in model benchmarking. In this section, we present a comparison between a widely used open-access dataset, the Surface Radiation Dataset Heliosat (SARAH-3) [73], and a recently released dataset, the High-Resolution European Surface Radiation Data Record (HANNA) [38]. This comparison also serves to justify our choice of HANNA as the reference SSI dataset.

The algorithm used for the generation of HANNA follows the HelioMont methodology, which was developed at MeteoSwiss [38]. A notable feature of the algorithm is its capacity to differentiate between clouds and snow. This is achieved by leveraging the fact that ground albedo changes slowly over time, whereas cloud albedo varies more rapidly [36,37]. Additionally, the algorithm also includes cloud shadow corrections [74] and adjustments to account for topographic effects, such as terrain shadowing, surface reflection, local horizon elevation angle, and sky view factor, which are particularly important in mountainous regions [37]. SARAH-3 tends to misclassify snow-covered surfaces as clouds, resulting in unreliable SSI estimates in Alpine regions [73,75]. Additionally, HANNA provides SSI fields at a spatial resolution five times higher than SARAH-3 and at twice the temporal sampling frequency.

Fig. 22 compares the mean absolute difference (MAD) and mean bias difference (MBD) between the HANNA and SARAH-3 datasets upscaled to the IFS-ENS spatial resolution. Deviations between HANNA and SARAH-3 remain low in low-elevation regions but increase sharply in mountainous areas such as the Alps and Jura, to the extent that even valleys can be distinguished in the MAD and MBD fields. Part of this deviation is that SARAH-3 tends to misclassify snow as clouds, which leads to persistently underestimated SSI values [75]. As a result, the MBD ($SSI_{HANNA} - SSI_{SARAH}$) becomes strongly positive, with differences exceeding 100 W m^{-2} in the Alps.

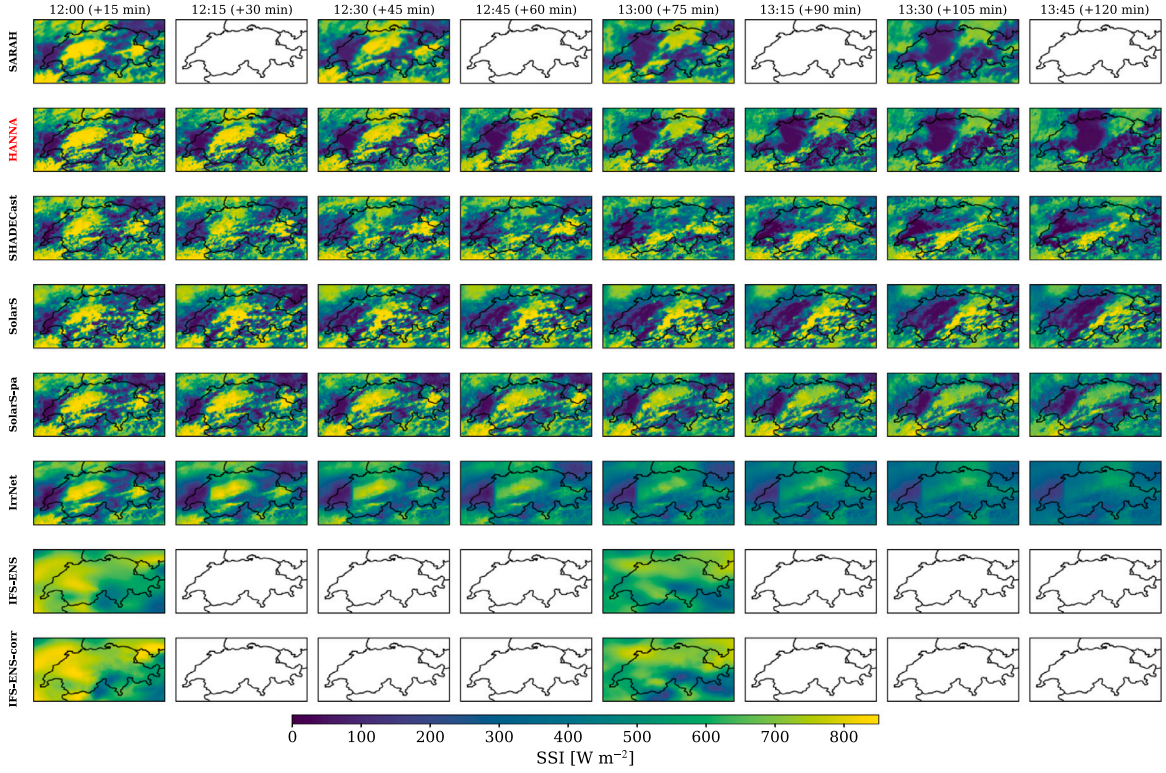


Fig. 20. Satellite-based SSI observation and model forecasts over the area of interest at lead times ranging from 15 to 120 min. The forecasts are issued at 12:00 UTC on 6 August 2019, a day with high-variability weather. For the probabilistic models, the ensemble member chosen is the one with the lowest RMSE. The black lines denote national borders. The row highlighted with the red label is used as the ground truth. Missing panels reflect differences in the temporal resolution of the satellite observation and forecast models.

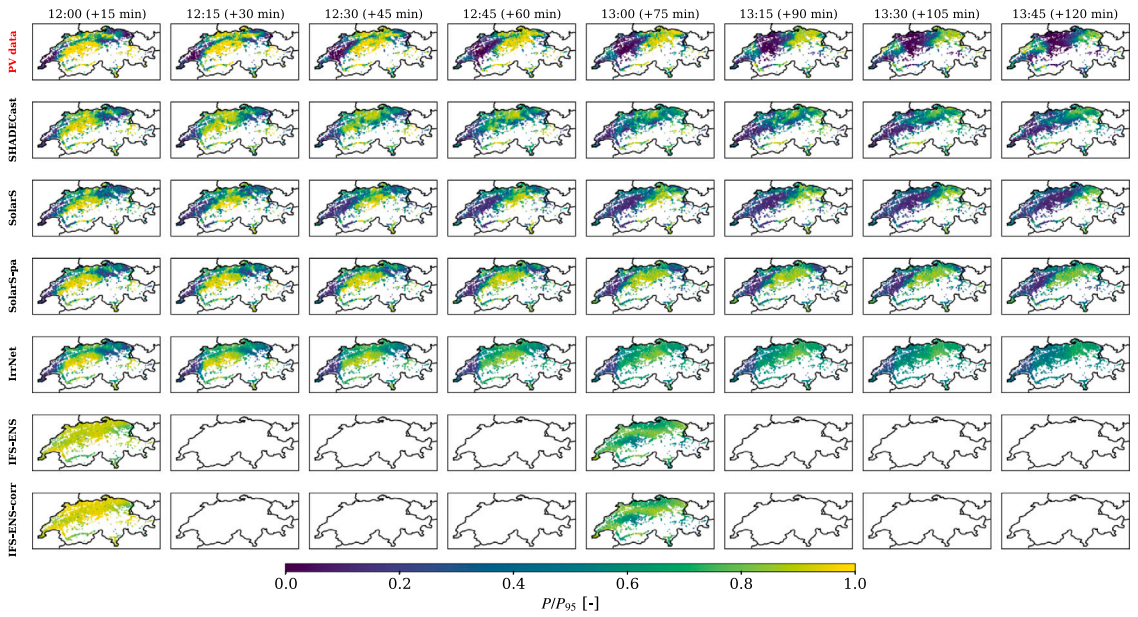


Fig. 21. PV power observation and model forecasts over the area of interest at lead times ranging from 15 to 120 min. The forecasts are issued at 12:00 UTC on 6 August 2019, a day with high-variability weather. For the probabilistic models, the ensemble member chosen is the one with the lowest RMSE. The black lines denote national borders. The row highlighted with the red label is used as the ground truth. Missing panels reflect differences in the temporal resolution of the satellite observation and forecast models. Note that the PV power output is normalized using the station-specific 95th percentile of the power time series.

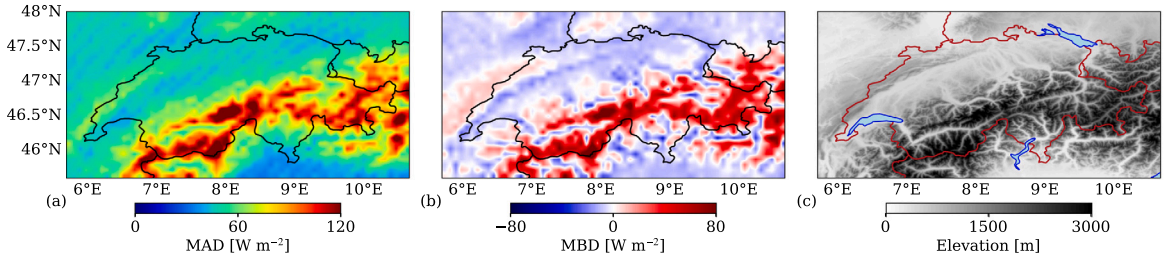


Fig. 22. (a) MAD and (b) MBD between SSI estimates from HANNA and SARAH-3 upscaled to the IFS-ENS spatial resolution. Results are averaged over all full-hour timestamps between sunrise and sunset during the two-year period 2019–2020. The MBE is defined as $\mathrm{SSI}_{\mathrm{HANNA}} - \mathrm{SSI}_{\mathrm{SARAH}}$. Finally, panel (c) illustrates the elevation map with values in meters above sea level.

Appendix C. Metrics

This appendix provides a detailed description of the metrics used in this study. Specifically, C.1 discusses the metrics adopted to evaluate the performance of the irradiance-to-power conversion model, while C.2 presents the metrics used for comparing forecasts in terms of SSI and PV power.

C.1. Irradiance-to-power conversion

This model takes as input an SSI value and several other geographical and temporal predictors to estimate the PV power output of a specific PV installation at a given time stamp. Here, we denote by $y_{s,n}$ and $\hat{y}_{s,n}$ the measured and predicted PV power at station s and time step n, respectively, where $s \in \{1, \dots, S\}$ indexes the PV installations and $n \in \{1, \dots, N\}$ indexes the time steps. The number of stations is fixed to $S = 6434$, while N depends on the number of time steps in the considered dataset (training, validation, or test).

For each station s, we evaluate three normalized error metrics: the normalized mean absolute error (nMAE_s), the normalized root mean square error (nRMSE_s), and the normalized mean bias error (nMBE_s), defined as

$$\mathrm{nMAE}_s = \frac{1}{N P_{95,s}} \sum_{n=1}^{N} \left| \hat{y}_{s,n} - y_{s,n} \right|,$$

$$\mathrm{nRMSE}_s = \frac{1}{P_{95,s}} \sqrt{\frac{1}{N} \sum_{n=1}^{N} \left(\hat{y}_{s,n} - y_{s,n} \right)^2},$$

$$\mathrm{nMBE}_s = \frac{1}{N P_{95,s}} \sum_{n=1}^{N} \left(\hat{y}_{s,n} - y_{s,n} \right),$$

where $P_{95,s}$ is the station-specific 95th percentile of the power time series, used here as a normalization factor. A positive value of nMBE_s indicates systematic overestimation of PV power production by the model, whereas a negative value indicates underestimation. In some applications, the metrics are aggregated across stations by simple averaging:

$$M = \frac{1}{S} \sum_{s=1}^{S} M_s,$$

where M denotes a generic metric (nMAE, nRMSE, or nMBE in this case). For simplicity, we use the same notation to refer both to the station-wise metric M_s and to its average across stations M. We note that these definitions apply to Section 5.1.

C.2. SSI and PV power comparison

This section illustrates the scores adopted to compare forecasts in terms of SSI and PV power. To start, we consider forecasts of PV power

for a generic PV installation, which consists of $L = 8$ lead times and $E = 10$ ensemble members. We denote with $\hat{y}_{s,n,e,l}$ the predicted PV power at station s, time step n, ensemble member e and lead time l. Here, $s \in \{1, \dots, S\}$ indexes the PV installations, $n \in \{1, \dots, N\}$ indexes the time steps, $e \in \{1, \dots, E\}$ indexes the ensemble member and $l \in \{1, \dots, L\}$ indexes the lead time. The corresponding observed PV power at station s and lead time l, associated with forecast n, is denoted by $y_{s,n,l}$.

Deterministic performance metrics are evaluated with respect to the ensemble mean, defined as

$$\bar{\hat{y}}_{s,n,l} = \frac{1}{E} \sum_{e=1}^{E} \hat{y}_{s,n,e,l}.$$

Deterministic models have $E = 1$, therefore $\bar{\hat{y}} = \hat{y}$.

The normalized station-wise metrics averaged over all lead times and all forecasts, shown for instance in Fig. 14, are defined as

$$\mathrm{nMAE}_s = \frac{1}{NLf_s} \sum_{n=1}^{N} \sum_{l=1}^{L} \left| \bar{\hat{y}}_{s,n,l} - y_{s,n,l} \right|,$$

$$\mathrm{nRMSE}_s = \frac{1}{f_s} \sqrt{\frac{1}{NL} \sum_{n=1}^{N} \sum_{l=1}^{L} \left(\bar{\hat{y}}_{s,n,l} - y_{s,n,l} \right)^2},$$

$$\mathrm{nMBE}_s = \frac{1}{NLf_s} \sum_{n=1}^{N} \sum_{l=1}^{L} \left(\bar{\hat{y}}_{s,n,l} - y_{s,n,l} \right),$$

where $f_s = P_{95,s}$. We note that it is important to apply this normalization, otherwise stations with higher $P_{95,s}$ values would inherently exhibit higher errors, as illustrated in Fig. 7(a).

In addition to deterministic scores, we also evaluate metrics tailored to probabilistic forecasts. For each forecast, the ensemble $\{\hat{y}_{s,n,e,l}\}_{e=1}^{E}$ can be used to construct predictive intervals or to approximate the full predictive distribution. We consider three standard metrics: the prediction interval coverage probability (PICP), the prediction interval normalized average width (PINAW), and the normalized continuous ranked probability score (nCRPS).

We define the central $(1 - \alpha)$ prediction interval at station s, forecast n, and lead time l via the empirical quantiles of the ensemble $\{\hat{y}_{s,n,e,l}\}_{e=1}^{E}$:

$$\left[\hat{y}^{\mathcal{L}}_{s,n,l},\ \hat{y}^{\mathcal{U}}_{s,n,l} \right] = \left[Q_{\alpha/2}(\{\hat{y}_{s,n,e,l}\}_{e=1}^{E}),\ Q_{1-\alpha/2}(\{\hat{y}_{s,n,e,l}\}_{e=1}^{E}) \right],$$

where Q_p denotes the empirical p-quantile. The quantity α denotes the nominal risk level, i.e. the probability mass outside the central prediction interval. In this work, we set $\alpha = 0.1$, which corresponds to a nominal 90% prediction interval. The coverage probability for station s is then defined as

$$\mathrm{PICP}_s = \frac{1}{NL} \sum_{n=1}^{N} \sum_{l=1}^{L} \mathbb{1}\left(\hat{y}^{\mathcal{L}}_{s,n,l} \le y_{s,n,l} \le \hat{y}^{\mathcal{U}}_{s,n,l} \right),$$

where $\mathbb{1}(\cdot)$ denotes the indicator function.

The average width of the same prediction intervals, normalized by the scaling factor f_s, is given by

$$\mathrm{PINAW}_s = \frac{1}{NLf_s} \sum_{n=1}^{N} \sum_{l=1}^{L} \left(\hat{y}^{\mathcal{U}}_{s,n,l} - \hat{y}^{\mathcal{L}}_{s,n,l} \right).$$

The nCRPS compares the predictive cumulative distribution function $F_{s,n,l}$, estimated from the ensemble, against the observation $y_{s,n,l}$, and is defined as

$$\mathrm{nCRPS}_s = \frac{1}{NLf_s} \sum_{n=1}^{N} \sum_{l=1}^{L} \int_{-\infty}^{\infty} \left(F_{s,n,l}(z) - \mathbb{1}(y_{s,n,l} \le z) \right)^2 dz.$$

In practice, $F_{s,n,l}$ is approximated from the empirical distribution of the ensemble members $\{\hat{y}_{s,n,e,l}\}_{e=1}^{E}$.

Both deterministic and probabilistic metrics can be further averaged across all stations, yielding the global nMAE, nRMSE, nMBE, PICP, PINAW, and nCRPS, as illustrated in Fig. 15. Depending on the application, the same metrics may instead be averaged across all forecasts and stations for a fixed lead time, as shown in Fig. 17. For clarity, we explicitly state in the text which averaging procedure is applied, while retaining a consistent notation throughout the manuscript. These metrics apply to Section 5.3.

When the comparison is carried out in terms of SSI, the same evaluation framework applies. The only difference is that $\hat{y}_{p,n,e,l}$ and $y_{p,n,l}$ denote the predicted and observed SSI values, and the station index s is replaced by the pixel index $p \in \{1, \dots, P\}$, where P is the total number of pixels covering the area of interest. In this case, we set $f_s = 1$, ensuring that the resulting metrics remain expressed in W m^{-2}. These metrics are adopted in Section 5.2.

Appendix D. ML-based bias correction methodology for IFS-ENS

To investigate whether bias correction could improve performance, we applied an ML-based correction to the IFS-ENS SSI forecasts. This is to mitigate systematic errors in the model outputs by learning the relationship between forecast and observed SSI values [76,77]. This section provides a detailed description of the adopted methodology and model architecture.

The bias correction model was trained using historical forecasts from IFS-ENS and corresponding SSI observations from HANNA for the period 2019–2020. The training, validation, and test sets were defined following the procedure outlined in Section 3.2. Specifically, the two-year period was divided into consecutive 12-day blocks. For each block, ten days were assigned to the training set, while the remaining 2 days were allocated to the validation and test sets.

The bias correction was implemented using a U-Net–based convolutional neural network. The model was configured with nine input channels containing the ensemble-mean forecast, elevation, latitude, longitude, year, and the sine and cosine transformations of both the DoY and HoD. Since temporal features such as DoY and HoD are scalar quantities, they were broadcast to the spatial domain and represented as two-dimensional grids in order to be compatible with the convolutional architecture. The architecture and hyperparameters were optimized with Optuna. The selected configuration consisted of a U-Net with two encoder–decoder levels, with 64 and 128 convolutional filters in the first and second layer, respectively, and incorporating dropout regularization with a rate of 0.19 to mitigate overfitting. Batch normalization and max-pooling operations were used to improve training stability and feature representation. The network was trained with an initial learning rate of 1.5×10^{-3} and an adaptive learning rate schedule, which reduced the rate by a factor of 0.12 after six consecutive epochs without improvement on the validation set. The model was optimized using the mean squared error loss function.

The trained model estimates the bias of the ensemble-mean IFS-ENS forecast. This correction is subsequently applied to all ensemble members, thereby adjusting the ensemble mean while preserving the ensemble spread. As a result, the bias-corrected IFS-ENS retains the same PINAW but exhibits a different PICP compared to the original forecasts.

Appendix E. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.egyai.2026.100786.

Data availability

The authors do not have permission to share data.

References

[1] International Energy Agency. Renewables 2024. 2024, https://www.iea.org/reports/renewables-2024, Licence: CC BY 4.0.

[2] Qin Jun, Jiang Hou, Lu Ning, Yao Ling, Zhou Chenghu. Enhancing solar PV output forecast by integrating ground and satellite observations with deep learning. Renew Sustain Energy Rev 2022;167:112680.

[3] Sharma Neetan, Puri Vinod, Mahajan Shubham, Abualigah Laith, Zitar Raed Abu, Gandomi Amir H. Solar power forecasting beneath diverse weather conditions using GD and LM-artificial neural networks. Sci Rep 2023;13(1):8517.

[4] Sameer Al-Dahidi, Mohammad Alrbai, Hussein Alahmer, Bilal Rinchi, Ali Alahmer. Enhancing solar photovoltaic energy production prediction using diverse machine learning models tuned with the chimp optimization algorithm. Sci Rep 2024;14.

[5] van der Meer DW, Munkhammar J, Widén J. Probabilistic forecasting of solar power, electricity consumption and net load: Investigating the effect of seasons, aggregation and penetration on prediction intervals. Sol Energy 2018;171:397–413.

[6] Xiuli Xiang, Xingyu Li, Yaoli Zhang, Jiang Hu. A short-term forecasting method for photovoltaic power generation based on the TCN-ECANet-GRU hybrid model. Sci Rep 2024;14:112680.

[7] Dácil Díaz-Bello, Carlos Vargas-Salgado, Manuel Alcazar-Ortega, David Alfonso-Solar. Optimizing photovoltaic power plant forecasting with dynamic neural network structure refinement. Sci Rep 2025;15.

[8] Agoua Xwégnon Ghislain, Girard Robin, Kariniotakis George. Probabilistic models for spatio-temporal photovoltaic power forecasting. IEEE Trans Sustain Energy 2019;10(2):780–9.

[9] Brester Christina, Kallio-Myers Viivi, Lindfors Anders V, Kolehmainen Mikko, Niska Harri. Evaluating neural network models in site-specific solar PV forecasting using numerical weather prediction data and weather observations. Renew Energy 2023;207:266–74.

[10] Perera Maneesha, De Hoog Julian, Bandara Kasun, Senanayake Damith, Halgamuge Saman. Day-ahead regional solar power forecasting with hierarchical temporal convolutional neural networks using historical power generation and weather data. Appl Energy 2024;361:122971.

[11] Wang Kai, Shan Shuo, Dou Weijing, Wei Haikun, Zhang Kanjian. A cross-modal deep learning method for enhancing photovoltaic power forecasting with satellite imagery and time series data. Energy Convers Manage 2025;323:119218.

[12] Karimi Ahmad Maroof, Wu Yinghui, Koyuturk Mehmet, French Roger H. Spatiotemporal graph neural network for performance prediction of photovoltaic power systems. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, (17):2021, p. 15323–30.

[13] Mayer Martin János. Benefits of physical and machine learning hybridization for photovoltaic power forecasting. Renew Sustain Energy Rev 2022;168:112772.

[14] Lorenz E, Heinemann D. 1.13 - prediction of solar irradiance and photovoltaic power. In: Sayigh Ali, editor. Comprehensive renewable energy. Oxford: Elsevier; 2012, p. 239–92.

[15] Perez Richard, Lorenz Elke, Pelland Sophie, Beauharnois Mark, Van Knowe Glenn, Hemker Karl, Heinemann Detlev, Remund Jan, Müller Stefan C, Traunmüller Wolfgang, Steinmauer Gerald, Pozo David, Ruiz-Arias Jose A, Lara-Fanego Vicente, Ramirez-Santigosa Lourdes, Gaston-Romero Martin, Pomares Luis M. Comparison of numerical weather prediction solar irradiance forecasts in the US, Canada and Europe. Sol Energy 2013;94:305–26.

[16] Aguiar L Mazorra, Pereira B, Lauret P, Díaz F, David M. Combining solar irradiance measurements, satellite-derived data and a numerical weather prediction model to improve intra-day solar forecasting. Renew Energy 2016;97:599–610.

[17] Lorenz Elke, Kühnert Jan, Heinemann Detlev, Nielsen Kristian Pagh, Remund Jan, Müller Stefan C. Comparison of global horizontal irradiance forecasts based on numerical weather prediction models with different spatio-temporal resolutions. Prog Photovolt, Res Appl 2016;24(12):1626–40.

[18] Haupt Sue Ellen, Kosović Branko, Jensen Tara, Lazo Jeffrey K, Lee Jared A, Jiménez Pedro A, Cowie James, Wiener Gerry, McCandless Tyler C, Rogers Matthew, Miller Steven, Sengupta Manajit, Xie Yu, Hinkelman Laura, Kalb Paul, Heiser John. Building the Sun4Cast system: Improvements in solar power forecasting. Bull Am Meteorol Soc 2018;99(1):121–36.

[19] Zhao Jing, Guo Zhen-Hai, Su Zhong-Yue, Zhao Zhi-Yuan, Xiao Xia, Liu Feng. An improved multi-step forecasting model based on WRF ensembles and creative fuzzy systems for wind speed. Appl Energy 2016;162:808–26.

[20] Roberts CD, Senan R, Molteni F, Boussetta S, Mayer M, Keeley SPE. Climate model configurations of the ECMWF integrated forecasting system (ECMWF-IFS cycle 43r1) for HighResMIP. Geosci Model Dev 2018;11(9):3681–712.

[21] Mathiesen Patrick, Kleissl Jan. Evaluation of numerical weather prediction for intra-day solar forecasting in the continental United States. Sol Energy 2011;85(5):967–77.

[22] European Centre for Medium-Range Weather Forecasts (ECMWF). Dissemination schedule. 2026, https://confluence.ecmwf.int/display/DAC/Dissemination+schedule. [Accessed 13 April 2026].

[23] Lucas Bruce D, Kanade Takeo. An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th international joint conference on artificial intelligence, vol. 2, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1981, p. 674–9.

[24] Carrière Thomas, Amaro e Silva Rodrigo, Zhuang Fuqiang, Saint-Drenan Yves-Marie, Blanc Philippe. A new approach for satellite-based probabilistic solar forecasting with cloud motion vectors. Energies 2021;14(16).

[25] Blanc Philippe, Remund Jan, Vallance Loïc. 6 - short-term solar power forecasting based on satellite images. In: Kariniotakis George, editor. Renewable energy forecasting. Woodhead publishing series in energy, Woodhead Publishing; 2017, p. 179–98.

[26] Urbich Isabel, Bendix Jörg, Müller Richard. A novel approach for the short-term forecast of the effective cloud albedo. Remote Sens 2018;10(6).

[27] Carpentieri A, Folini D, Nerini D, Pulkkinen S, Wild M, Meyer A. Intraday probabilistic forecasts of surface solar radiation with cloud scale-dependent autoregressive advection. Appl Energy 2023;351:121775.

[28] Lago Jesus, De Brabandere Karel, De Ridder Fjo, De Schutter Bart. Short-term forecasting of solar irradiance without local telemetry: A generalized model using satellite data. Sol Energy 2018;173:566–77.

[29] Brahma Banalaxmi, Wadhvani Rajesh. Solar irradiance forecasting based on deep learning methodologies and multi-site data. Symmetry 2020;12(11).

[30] Nielsen Andreas H, Iosifidis Alexandros, Karstoft Henrik. IrradianceNet: Spatiotemporal deep learning model for satellite-derived solar irradiance short-term forecasting. Sol Energy 2021;228:659–69.

[31] Carpentieri A, Folini D, Leinonen J, Meyer A. Extending intraday solar forecast horizons with deep generative models. Appl Energy 2025;377:124186.

[32] Yang Dazhi, Wang Wenting, Bright Jamie M, Voyant Cyril, Notton Gilles, Zhang Gang, Lyu Chao. Verifying operational intra-day solar forecasts from ECMWF and NOAA. Sol Energy 2022;236:743–55.

[33] Wang Wenting, Yang Dazhi, Hong Tao, Kleissl Jan. An archived dataset from the ECMWF ensemble prediction system for probabilistic solar power forecasting. Sol Energy 2022;248:64–75.

[34] Sebastianelli Alessandro, Serva Federico, Ceschini Andrea, Paletta Quentin, Panella Massimo, Le Saux Bertrand. Machine learning forecast of surface solar irradiance from meteo satellite data. Remote Sens Environ 2024;315:114431.

[35] Schmetz Johannes, Pili Paolo, Tjemkes Stephen, Just Dieter, Kerkmann Jochen, Rota Sergio, Ratier Alain. An introduction to meteosat second generation (msg). Bull Am Meteorol Soc 2002;83(7):977–92.

[36] Stöckli Reto. The HelioMont Surface Solar Radiation Processing (2022 Version). Scientific Report 93, MeteoSwiss; 2013, p. 126.

[37] Castelli M, Stöckli R, Zardi D, Tetzlaff A, Wagner JE, Belluardo G, Zebisch M, Petitta M. The HelioMont method for assessing solar irradiance over complex terrain: Validation and improvements. Remote Sens Environ 2014;152:603–13.

[38] EUMETSAT Satellite Application Facility on Climate Monitoring. Meteosat high resolution solar radiation (HANNA) demonstrational data record 2019–2020. 2025, https://www.cmsaf.eu/demo_hanna. [Accessed 24 July 2025].

[39] Tavares Ailton M, Conceição Ricardo, Lopes Francisco M, Silva Hugo G. Effect of solar irradiation inter-annual variability on PV and CSP power plants production capacity: Portugal case-study. Energies 2024;17(21).

[40] Hersbach Hans. Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather Forecast 2000;15(5):559–70.

[41] Brockwell Peter J, Davis Richard A. Time series: theory and methods. Springer science & business media; 1991.

[42] Pfeifroth U, Kothe S, Trentmann J, Hollmann R, Fuchs P, Kaiser J, Werscheck M. Surface radiation data set - heliosat (SARAH) - edition 2.1. 2019, Satellite Application Facility on Climate Monitoring. https://www.cmsaf.eu/EN/Home/home_node.html. [Accessed 29 July 2025].

[43] Ronneberger Olaf, Fischer Philipp, Brox Thomas. U-net: Convolutional networks for biomedical image segmentation. 2015.

[44] Yang Dazhi, Wang Wenting, Gueymard Christian A, Hong Tao, Kleissl Jan, Huang Jing, Perez Marc J, Perez Richard, Bright Jamie M, Xia Xiang'ao, van der Meer Dennis, Peters Ian Marius. A review of solar forecasting, its dependence on atmospheric sciences and implications for grid integration: Towards carbon neutrality. Renew Sustain Energy Rev 2022;161:112348.

[45] Yang Dazhi. Temporal-resolution cascade model for separation of 1-min beam and diffuse irradiance. J Renew Sustain Energy 2021;13(5):056101.

[46] Reindl DT, Beckman WA, Duffie JA. Evaluation of hourly tilted surface radiation models. Sol Energy 1990;45(1):9–17.

[47] Jimenez Pedro A, Hacker Joshua P, Dudhia Jimy, Haupt Sue Ellen, Ruiz-Arias Jose A, Gueymard Chris A, Thompson Gregory, Eidhammer Trude, Deng Aijun. WRF-solar: Description and clear-sky assessment of an augmented NWP model for solar power prediction. Bull Am Meteorol Soc 2016;97(7):1249–64.

[48] Chu Yinghao, Li Mengying, Coimbra Carlos FM, Feng Daquan, Wang Huaizhi. Intra-hour irradiance forecasting techniques for solar power integration: A review. IScience 2021;24(10):103136.

[49] Mayer Martin János, Gróf Gyula. Techno-economic optimization of grid-connected, ground-mounted photovoltaic power plants by genetic algorithm based on a comprehensive mathematical model. Sol Energy 2020;202:210–26.

[50] Nguyen Andu, Velay Maxime, Schoene Jens, Zheglov Vadim, Kurtz Ben, Murray Keenan, Torre Bill, Kleissl Jan. High PV penetration impacts on five local distribution networks using high resolution solar resource assessment with sky imager and quasi-steady state distribution system simulations. Sol Energy 2016;132:221–35.

[51] Li Caixia, Xu Yuanyuan, Xie Minglang, Zhang Pengfei, Zhang Bohan, Xiao Bo, Zhang Sujun, Liu Ziheng, Zhang Wenjie, Hao Xiaojing. Assessing solar-to-PV power conversion models: Physical, ML, and hybrid approaches across diverse scales. Energy 2025;323:135744.
[52] Das Utpal Kumar, Tey Kok Soon, Seyedmahmoudian Mehdi, Mekhilef Saad, Idris Moh Yamani Idna, Van Deventer Willem, Horan Bend, Stojcevski Alex. Forecasting of photovoltaic power generation and model optimization: A review. Renew Sustain Energy Rev 2018;81:912–28.
[53] Liu Luyao, Zhao Yi, Chang Dongliang, Xie Jiyang, Ma Zhanyu, Sun Qie, Yin Hongyi, Wennersten Ronald. Prediction of short-term PV power output and uncertainty analysis. Appl Energy 2018;228:700–11.
[54] Kim Gyu Gwang, Choi Jin Ho, Park So Young, Bhang Byeong Gwan, Nam Woo Jun, Cha Hae Lim, Park NeungSoo, Ahn Hyung-Keun. Prediction model for PV performance with correlation analysis of environmental variables. IEEE J Photovoltaics 2019;9(3):832–41.
[55] Ahmed R, Sreeram V, Mishra Y, Arif MD. A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization. Renew Sustain Energy Rev 2020;124:109792.
[56] Bamisile Olusola, Acen Caroline, Cai Dongsheng, Huang Qi, Staffell Iain. The environmental factors affecting solar photovoltaic output. Renew Sustain Energy Rev 2025;208:115073.
[57] Steger CR, Steger B, Schär C. HORAYZON v1.2: an efficient and flexible ray-tracing algorithm to compute horizon and sky view factor. Geosci Model Dev 2022;15(17):6817–40.
[58] Chen Tianqi, Guestrin Carlos. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. New York, NY, USA: Association for Computing Machinery; 2016, p. 785–94.
[59] Quinlan JR. Induction of decision trees. Mach Learn 1986;1.
[60] Smola Alex J, Schölkopf Bernhard. A tutorial on support vector regression. Stat Comput 2004;14.
[61] Ke Guolin, Meng Qi, Finley Thomas, Wang Taifeng, Chen Wei, Ma Weidong, Ye Qiwei, Liu Tie-Yan. LightGBM: A highly efficient gradient boosting decision tree. In: Guyon I, Luxburg U Von, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. In: Advances in neural information processing systems, vol. 30, Curran Associates, Inc.; 2017, p. 785–94.
[62] Pedregosa Fabian, Varoquaux Gaël, Gramfort Alexandre, Michel Vincent, Thirion Bertrand, Grisel Olivier, Blondel Mathieu, Prettenhofer Peter, Weiss Ron, Dubourg Vincent, Vanderplas Jake, Passos Alexandre, Cournapeau David, Brucher Mathieu, Perrot Matthieu, Duchesnay Édouard. Scikit-learn: Machine learning in python. J Mach Learn Res 2011;12:2825–30.
[63] Hochreiter Sepp, Schmidhuber Jürgen. Long short-term memory. Neural Comput 1997;9(8):1735–80.
[64] Akiba Takuya, Sano Shotaro, Yanase Toshihiko, Ohta Takeru, Koyama Masanori. Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. ACM; 2019, p. 2623–31.
[65] Ineichen Pierre. A broadband simplified version of the solis clear sky model. Sol Energy 2008;82(8):758–62.
[66] Mueller RW, Dagestad KF, Ineichen P, Schroedter-Homscheidt M, Cros S, Dumortier D, Kuhlemann R, Olseth JA, Piernavieja G, Reise C, Wald L, Heinemann D. Rethinking satellite-based solar irradiance modelling: The SOLIS clear-sky module. Remote Sens Environ 2004;91(2):160–74.
[67] Holmgren William F, Hansen Clifford W, Mikofski Mark A. Pvlib python: a python package for modeling solar energy systems. J Open Source Softw 2018;3(29):884.
[68] Agrawal Shreya, Hassen Mohammed Alewi, Brempong Emmanuel Asiedu, Babenko Boris, Zyda Fred, Graham Olivia, Li Di, Merchant Samier, Potes Santiago Hincapie, Russell Tyler, Cheresnick Danny, Kakkirala Aditya Prakash, Rasp Stephan, Hassidim Avinatan, Matias Yossi, Kalchbrenner Nal, Gupta Pramod, Hickey Jason, Bell Aaron. An operational deep learning system for satellite-based high-resolution global nowcasting. 2025.
[69] Schuurman KR, Meyer A. Surface solar radiation: AI satellite retrieval can outperform heliosat and generalizes well to other climate zones. Int J Remote Sens 2025.
[70] Gneiting Tilmann. Quantiles as optimal point forecasts. Int J Forecast 2011;27(2):197–207.
[71] Garg Piyush, Gergel Diana R, Shao Andrew E, Yacalis Galen J. The recipe matters more than the kitchen: Mathematical foundations of the AI weather prediction pipeline. 2026.
[72] Andenæs Erlend, Jelle Bjørn Petter, Ramlo Kristin, Kolås Tore, Selj Josefine, Foss Sean Erik. The influence of snow and ice coverage on the energy generation from photovoltaic solar cells. Sol Energy 2018;159:318–28.
[73] Pfeifroth U, Drücke J, Kothe S, Trentmann J, Schröder M, Hollmann R. SARAH-3 – satellite-based climate data records of surface solar radiation. Earth Syst Sci Data 2024;16(11):5243–65.
[74] Li Zhiwei, Shen Huanfeng, Weng Qihao, Zhang Yuzhuo, Dou Peng, Zhang Liangpei. Cloud and cloud shadow detection for optical satellite imagery: Features, algorithms, validation, and prospects. ISPRS J Photogramm Remote Sens 2022;188:89–108.
[75] Carpentieri A, Folini D, Wild M, Vuilleumier L, Meyer A. Satellite-derived solar radiation for intra-hour and intra-day applications: Biases and uncertainties by season and altitude. Sol Energy 2023;255:274–84.
[76] Cho Dongjin, Yoo Cheolhee, Im Jungho, Cha Dong-Hyun. Comparative assessment of various machine learning-based bias correction methods for numerical weather prediction model forecasts of extreme air temperatures in urban areas. Earth Space Sci 2020;7(4). e2019EA000740, e2019EA000740 2019EA000740.
[77] Han Lei, Chen Mingxuan, Chen Kangkaim, Chen Haonan, Zhang Yanbiao, Lu Bing, Song Linye, Qin Rui. A deep learning method for bias correction of ECMWF 24–240 h forecasts. Adv Atmospheric Sci 2021;38(9):1444–59.