Solar RRL

RESEARCH ARTICLE OPEN ACCESS

Enhancing Intra-Hour Solar Irradiance Forecasting for Solar Applications: A Blended Model of Satellite, Sky Imager, and Persistence

Bijan Nouri[1] | Jorge Lezaca[2] | Yann Fabel[1] | Annette Hammer[2] | Niklas Blum[1] | Stefan Wilbert[1]

[1]German Aerospace Center (DLR), Institute for Solar Research, Almería, Spain | [2]DLR, Institute of Networked Energy Systems, Oldenburg, Niedersachsen, Germany

Correspondence: Bijan Nouri (bijan.nouri@dlr.de)

Received: 26 June 2025 | **Revised:** 10 October 2025 | **Accepted:** 28 October 2025

Keywords: all-sky imagers | hybrid models | irradiance | machine learning | persistence | satellite data | solar forecasting

ABSTRACT

The increasing integration of solar power requires highly accurate intra-hour solar irradiance forecasts. This study aims to significantly improve intra-hour solar irradiance forecasts by developing and evaluating a blending approach that integrates distinct forecast sources. Our methodology involves extending the horizon of an All-sky imager (ASI) data-driven transformer-based model up to 1 h ahead. The outputs of this ASI model are blended with a Heliosat-3-based satellite forecast and a persistence forecast via linear regression as well as with distinct advanced machine learning algorithms. We assess the hybrid system's performance across varying sky conditions and analyze the impact of temporal aggregation schemes and the effective spatial coverage of a single ASI installation. Results demonstrate that this integrated multisource hybrid approach provides substantial benefits by reducing the overall root mean squared error and mean absolute error over the standalone satellite forecast by 13.6% and 17.0%, respectively. This is attributed to the complementary strengths of the individual models: ASI excels under dynamic conditions, satellite offers broader spatial coverage, and persistence provides a robust baseline for the immediate future. Furthermore, the strong generalization capability of the ASI model is shown through its effective performance across climatically distinct sites (training in southern Spain and validation in northern Germany).

1 | Introduction

Accurate intra-hour solar irradiance forecasts are essential for the seamless integration of solar energy into power grids [1]. The intermittent nature of solar resources increases reliance on costly reserves and reduces grid efficiency [2]. Advanced forecasting methods enable grid operators to anticipate these variabilities, optimize plant performance, improve plant and grid stability, and minimize operating costs, thereby supporting the wider deployment of renewable energy systems [3].

Improved local solar irradiance forecasts optimized for high accuracy within the area of commercial PV power plants can be achieved by combining satellite-based forecasts with complementary forecasts derived from ground-based observations such as All-Sky Imagers (ASIs) and/or radiometers. Satellite-based solar irradiance forecasting harnesses geostationary satellite imagery. These forecasts offer broad spatial coverage and can predict cloud movement patterns on a large scale, but they are limited by their temporal and spatial resolution, typically on the order of 1–5 km and 5–15 min [4]. Satellites like Meteosat Second Generation (MSG) and GOES provide continuous multispectral data. These are processed using methods such as the Heliosat-3 semiempirical approach, which derives a cloud index from visible band radiances to estimate ground-level Global Horizontal Irradiance (GHI) supported by clear-sky models

accounting for atmospheric and terrain factors [5]. Cloud motion vectors (CMVs), extracted via optical flow or phase correlation techniques from sequential cloud index images, enable forecasting of future cloud fields [6]. Increasingly, deep learning models, including Convolutional Neural Networks (CNNs) and UNet architectures, directly learn complex spatio-temporal cloud dynamics from raw imagery, improving accuracy for intra-hour to short-term (up to 6 h) forecasts compared to traditional CMV methods [7].

Radiometer-based ground observations of solar irradiance provide high temporal resolution and measurements, but they exhibit limited spatial coverage and do not capture future changes, such as upcoming cloud movements [3]. These systems, in the best case equipped with pyranometers or pyrheliometers, can deliver precise measurements of GHI, diffuse horizontal irradiance (DHI), and direct normal irradiance (DNI) at sub-minute intervals, enabling detailed monitoring of local irradiance conditions [8]. Forecasting based on radiometer data often relies on statistical or machine learning models, such as autoregressive integrated moving average (ARIMA) or neural networks, which use historical time-series data to predict short-term irradiance-fluctuations [9]. However, their point-specific nature restricts spatial representativeness, and their inability to directly observe clouddynamics limitsforecast horizons [3].

ASIs provide high-resolution images of the sky, allowing for detailed cloud observation, yet they are limited by their field of view and thus forecast horizon [4]. ASI-based forecasting systems are often paired with ground measurements and typically provide forecast horizons of 10–20 min ahead [10]. These systems typically operate by detecting and tracking cloud structures in images, employing techniques such as CMV extraction to forecast cloud trajectories and their impact on solar radiation [11]. Advanced image processing methods, including cloud segmentation and characterization [12], are applied to quantify cloud properties like opacity and coverage [11]. Increasingly, machine learning and deep learning algorithms, such as CNNs and recurrent neural networks (RNNs), are utilized to model complex relationships between sky images and future GHI or to directly predict future sky conditions [12]. These systems integrate projected cloud positions and characteristics to accurately estimate future solar irradiance at specific ground locations. Purely data-driven approaches provide forecasts for the ASI's direct location [13], whereas multicameraphysics-based models can generateir-radiance maps [14].

This study presents a significant advancement in intra-hour solar forecasting by enhancing a data-driven ASI model [15]. The model is based on a time-series [16] and vision transformer [17] and generates multistep forecasts with lead times (LTs) at 1 min intervals and updates every 30 s, and is originally designed for a 20 min horizon. We trained a new model which achieves a forecast horizon of up to 1 h and combined its outputs with a satellite-based forecasts, which uses the Heliosat-3 method [18]. These satellite forecasts produce multistep forecasts with 15 min intervals between LTs at a 15 min update rate. The blending process employs a linear-regression approach, divided into three tasks: homogenization, regression, and prediction [19]. Additionally, a persistence model is incorporated into the blending framework, enhancing robustness, particularly for very short lead times where its performance is often competitive. The combined forecast is referred to as the 'blended forecast' hereafter.

This research aims to demonstrate the combined strengths of different forecasting approaches and quantify their performance, while also identifying potential limitations. Thus, this research conducts a comprehensive evaluation of the proposed hybrid forecasting system. The performance of the system is analyzed under varying sky conditions, from clear to partly cloudy to overcast, to characterize its robustness.

We also investigate the impact of temporal aggregation schemes on forecast accuracy. Furthermore, the spatial influence of a single ASI installation is investigated, assessing the effective coverage of its forecast benefits. This includes evaluating how the accuracy of a hybrid forecast, relying on an ASI, changes with increasing distance from that sensor. This analysis aims to showcase the practical coverage within which a single ASI can provide an advantage. Finally, the study benchmarks the linear-regression blending approach against more advanced machine learning algorithms.

When integrating satellite and ASI data into blended forecasts, previous studies on similar approaches did observe some improvements [20]. Employed linear regression and random forest algorithms to combine multiple forecast sources, including low- and high-resolution satellite models, smart persistence, and ASI-based predictions. Their results showed reductions in root mean squared error (RMSE) and mean absolute error (MAE) of approximately 10–20% relative to the individual component models. However, these improvements were not statistically significant compared to simpler blends that used only satellite and persistence forecasts. This finding suggested that, under their experimental setup, the inclusion of ASI data provided only marginal additional benefit for irradiance nowcasting.

Subsequent work by Straub et al. [21] introduced a novel ASI-based forecasting approach built around a sequence of image-processing steps including a machine learning model that estimates pixel-wise clear-sky indices. This ASI model was then linearly blended with Heliosat-based satellite forecasts and persistence forecasts. The authors reported clear benefits of hybridization. Building on this, the same group, systematically evaluated a range of more advanced blending algorithms, including gradient boosting regression, extremely randomized trees, random forest, k-nearest neighbors, Lasso, and a feed-forward neural network [22]. Among these, only Lasso and gradient boosting achieved modest RMSE improvements relative to the simpler linear-regression blend, indicating that higher model complexity did not necessarily translate into substantially better performance. Both studies [21, 22] were limited to a 15 min forecast horizon and relied on an ASI system requiring external cloud-base-height measurements from a ceilometer. Moreover, neither study explicitly examined whether incorporating ASI data provides a statistically significant advantage over satellite–persistence hybrids.

These previous efforts collectively highlight the potential of blended ASI–satellite–persistence approaches but also reveal key open questions. In particular, the impact of modern, fully data-driven ASI architectures [10].

Building upon these developments, this research investigates the benefits of integrating data-driven ASI forecasts into a hybrid forecast with a horizon up to 60 min ahead. Such a hybrid forecast is expected to preserve the high temporal resolution characteristic of ASI systems. Furthermore, a key objective involves exploring the generalization capabilities of the ASI model across diverse geographic locations and varying climate conditions. The ASI model was originally trained in Spain which is classified as a desert climate under the Köppen climate classification, while the site in Germany used in this study falls under a temperate oceanic climate [23].

The objective of this study is not to advocate for the unconditional use of blended forecasts, but to examine their benefits and limitations across a range of operationally relevant conditions. By evaluating performance under varying sky states, temporal aggregation schemes, and spatial contexts, as well as testing generalization across climates, we aim to highlight both the potential and current bottlenecks of blended approaches. Forecast improvements are reported using standard error metrics (RMSE, MAE), including formal statistical significance testing. The emphasis is instead placed on the consistency and robustness of observed patterns, which provide practical insights for deployment.

2 | Used Datasets

Blended forecasts were developed and validated at the OLDON test site in northwestern Germany (53°8′46.96″N, 8°13′2.41″E). This site is part of the unique Eye2Sky ASI cloud monitoring network, which consists of 29 ASIs, 12 irradiance measurement stations, and 8 ceilometers distributed across an area of approximately 10,000 km^2 [24]. Section 6.4 uses three additional Eye2Sky network sites to evaluate the performance of blended forecasts at locations distant from the ground stations used. Table 1 illustrates all the sites that were used, their geographic coordinates positions and the distance to the main site, OLDON. Figure 1 shows a sky image for each site.

All four sites are equipped with rotating shadowband irradiometers (RSI) that measure GHI, DNI, and DHI [26]. The device uses a fast pyranometer with a small, motorized shadowband that swiftly moves across the detector. This enables the RSI to measure GHI when the detector is unshaded and DHI when it is shaded. DNI is then derived from these measurements and the sun's position. When appropriate calibration and correction functions are applied, accuracies in the range of 2.5–4% for GHI and 3–9% for DNI are obtained with RSIs (standard uncertainty, 1σ) [27]. A comprehensive evaluation of the site-specific uncertainties of the RSIs in Oldenburg was conducted in [28].

TABLE 1 | The geographic coordinates of the used sites and the distance to the main site, OLDON.

Site tag	Geographic coordinates	Distance to OLDON (km)	Elevation (m)
OLDON	53°8′46.96″N, 8°13′2.41″E	0.00	20
OLCLO	53° 6′42.96″N, 8°12’33.35″E	3.86	14
OLUOL	53° 9′12.53″N, 8° 9′42.92″E	3.79	15
OLJET	53° 7′13.38″N, 8° 5′44.95″E	8.63	4

The dataset used in this study spans July and August 2020. Data from July are reserved for regression/training tasks (50373 datapoints), while data from August are used only for validation (45,163 datapoints). Only timestamps with a sun elevation angle ≥10° are considered. The average irradiance for the 2 months is given in Table 2. August 2020 was notably sunnier, with higher average DNI and GHI but lower DHI. Detailed information on the irradiance distribution can be found in Appendix A1.

To evaluate forecast accuracy under various conditions, each timestamp is classified based on its DNI variability, as described in ref. [29]. This procedure originally distinguishes between eight distinct conditions, ranging from clear sky (class 1) to overcast (class 8), including six classes representing intermediate sky conditions, characterized by a distinct sky index and various variability indicators. For this work, these eight classes are consolidated into three overall bundled classes due to the low population of some original classes, which would otherwise lead to statistically insignificant results. Table 3 provides an overview of the original and bundled classes. Roughly, the bundled classes can be distinguished by high, intermediate, and low clear sky index. A detailed description of the original classification procedure, its corresponding variability indices, and thresholds can be found in ref. [29].

Figure 2 shows the present distribution of DNI variability classes at the OLDON site for the periods under consideration. Both months show a share of around 25% for bundled class 2, which describes partly cloudy weather with a mainly intermediate clear sky index and a high to intermediate variability. July was clearly the cloudier month, with only around 9% of observations in the clear-sky-conditions bundled class 1, compared to around 27% in August. Cloudy bundled class 3 conditions were the most common for both months, accounting for around 65% of July and 48% of August. This comes as no surprise, as this part of Germany is considered to have a temperate oceanic climate, which is often characterized by high cloud coverage, according to the Köppen climate classification [23]. The distribution of variability classes demonstrates that the dataset encompasses a wide range of diverse and complex conditions, including various cloud and irradiance scenarios. A more detailed analysis of the prevailing atmospheric conditions is provided in Appendix A1 (Meteorological Information on the Data Set Used). The appendix includes histograms of the solar irradiance components (DNI, GHI, DHI), air mass, Linke turbidity, and cloud coverage. Additionally, keograms derived from sky images are presented for each day. Given this diversity and complexity, the dataset is considered highly suitable for benchmarking different forecasting models against one another.

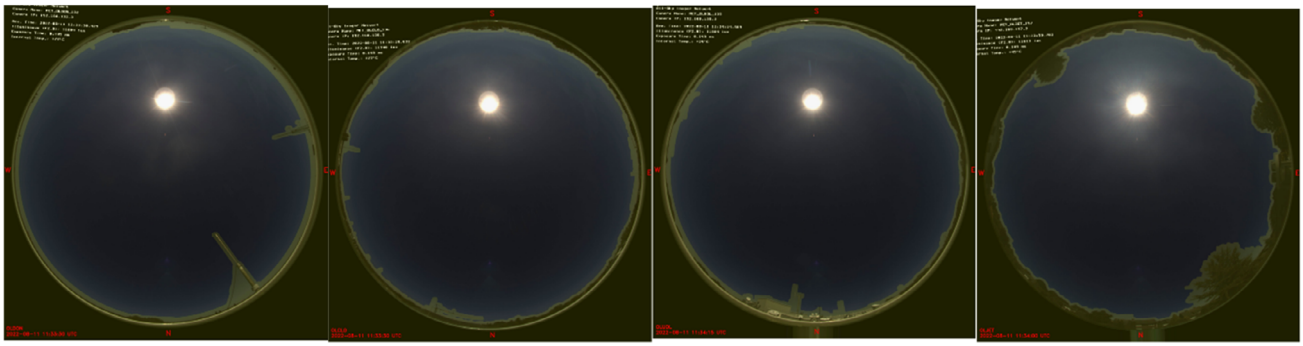

FIGURE 1 | Sky images with horizon and obstacle mask for the sites OLDON, OLCLO, OLUOL, and OLJET [25].

TABLE 2 | Average GHI, DHI, and DNI values for July and August 2020 taken from OLDON, considering only sun elevation angles ≥10°.

	GHI[a]		DHI		DNI	
	July	**August**	**July**	**August**	**July**	**August**
Mean (W/m^2)	321.7	358.1	192.1	166.7	209.9	323.5

[a]Data from July have been used for training, August for validation.

TABLE 3 | Brief description of DNI variability classes.

Original classes	Bundled classes	Sky conditions	Clear sky index	Variability
1	1	Mostly clear sky	Very high clear sky index	Low variability
2		Almost clear sky	High clear sky index	Low variability
3	2	Partly cloudy	High/intermediate clear sky index	Intermediate variability
4		Partly cloudy	Intermediate clear sky index	High variability
5		Partly cloudy	Intermediate clear sky index	Intermediate variability
6		Partly cloudy	Intermediate/low clear sky index	High variability
7	3	Almost overcast	Low clear sky index	Intermediate variability
8		Mostly overcast	Very low clear sky index	Low variability

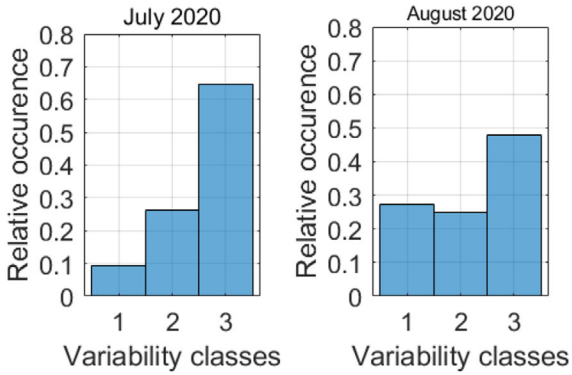

FIGURE 2 | Distribution of DNI bundled variability classes at OLDON.

3 | Used Forecasting Models

This work uses three distinct forecasting models: a satellite-based model (Section 3.1), an ASI-based model (Section 3.2), and a persistence baseline model (Section 3.3).

3.1 | Satellite Forecast

The satellite-based forecast model [18] utilizes geostationary Meteosat Second Generation (MSG) images. This method provides Europe-wide coverage with a spatial resolution of ∼2 km (for northwest Germany) and a temporal resolution and update interval of 15 min.

The model processes raw satellite imagery through a series of techniques to produce spatially resolved forecasts. The forecasting process starts with receiving raw satellite images from the MSG platform. These undergo radiometric correction and geometric rectification using Eumetsat's calibration routines [30]. Next, Cloud Index (CI) images are derived using methods from [18]. The CI is a measure of the cloudiness within a satellite pixel, summarizing cloud thickness and cloud cover. The cloud index is related to the average atmospheric transmissivity of the pixel to solar radiation.

To capture cloud dynamics, the model analyzes consecutive CI images to compute cloud motion vectors using block matching techniques. These vectors represent the displacement of cloud

formations over time. By extrapolating along these vectors, the model generates predicted CI images, accounting for advective cloud movement to represent future sky conditions dynamically.

The predicted CI images are then converted into solar irradiance values by using their relation to the clear sky index (k), as described in [18]. The clear sky index is the ratio of GHI to the GHI under theoretical clear-sky conditions (GHIcs). The GHIcs are calculated using the clear sky model [31], which incorporates climatological turbidity values from [32] to account for the effect of atmospheric aerosol and water vapor on clear-sky transmission. This multistep process yields pixel-level GHI forecasts.

In operational satellite-based forecasts, an adaptation of the clear sky model to recent observations is often conducted to reduce mean bias error (MBE). This step is not included in the baseline satellite forecast. Here, the adjustment to ground observations is conducted in the result Section 6 as the Satellite+Persistence approach. This work does not specifically evaluate the effect of high aerosol optical depth (AOD). Corresponding AOD evaluations are performed here [33].

3.2 | ASI Forecast

The ASI model, as described in ref. [15], is a deep learning model trained end-to-end to produce multistep solar irradiance forecasts without explicitly modeling physical processes such as cloud dynamics or transmittance. It employs a dual-branch transformer architecture to process both sky images and timeseries data, including historical irradiance measurements (GHI, DNI, DHI) and sun position data (elevation and azimuth), to predict scalar GHI and DNI values for lead times typically ranging from 1 to 20 min. The first branch uses an architecture from ref. [16] to process 30 min sequences of time-series data. This process extracts temporal patterns and generates feature representation vectors. The second branch utilizes the TimeSformer architecture [17], which applies a spatio-temporal attention mechanism to analyze 5 min sequences of ASI images, capturing cloud movement and behavior. These feature vectors from both branches are concatenated and fed into a multilayer perceptron to generate multistep forecasts for all lead times at once. The model has been trained on a comprehensive dataset from southern Spain (37°5′38.04″N, 2°21′32.40″W), consisting of over 3 million ASI images collected between 2015 and 2021 under diverse weather and irradiance conditions [15]. The dataset was filtered specifically to ensure balanced occurrences of clear skies, partial cloudiness, and overcast conditions. By training both branches in parallel without pretraining, the model optimizes the interplay between image and timeseries data, ensuring robust feature extraction and forecast accuracy across varying atmospheric conditions. For this work, the model was retrained to derive multistep forecasts with a horizon up to 60 min ahead. The comprehensive dataset from southern Spain was used once again for this purpose [15]. In operational practices, it is often meaningful to correct ASI forecasts with recent ground observations, which is presented by the ASI + persistence forecast (see Appendix A2).

3.3 | Persistence Forecast

The ground-based persistence approach utilizes a state-of-the-art scaled persistence method that inherently accounts for changes in solar position [1]. This method's core principle involves calculating a clear-sky index from real-time irradiance measurements. To achieve this, the Ineichen clear-sky model is employed [34]. Crucially, the Linke turbidity values required by the Ineichen model are derived dynamically from local DNI measurements, providing a site-specific atmospheric turbidity input. This entire process, including the clear-sky model and clear-sky index calculation, uses the PVlib implementation [35].

The derived clear-sky index then becomes the persisting magnitude, effectively propagating the current cloud conditions while accurately reflecting the expected irradiance under clear skies. This method uses both GHI and DNI measurements as input, and the target value for the used persistence forecasts is GHI. The generated forecasts have the same temporal resolution and update rate as the ASI model. Despite its relative simplicity, scaled persistence, particularly for very short-term forecast horizons, often provides a robust baseline and can even outperform more complex models due to its direct reliance on recent observations.

4 | Approaches used to Combine Forecasts

For effective multisource data blending, a crucial preprocessing step involves the homogenization of GHI forecasts from individual sources to a consistent temporal resolution and forecast horizon, as well as a standardized update frequency. This process primarily focuses on aligning the satellite forecasts with the target ASI forecast characteristics with an update rate of 30 s and a forecast horizon of 1 h, consisting of 60 lead times at 1 min intervals. First, bilinear interpolation is applied to each required site on the GHI satellite map. Notably, each Oldenburg site is situated within a distinct pixel of the map. Second, within each issued satellite forecast, temporal interpolation is applied across the lead times. Third, to address the coarser update rate of the satellite forecasts and align them with the desired higher frequency, a forecast forwarding mechanism is implemented. For each new, high-frequency timestamp for which a satellite forecast is not available, the most recent existing satellite forecast is reused. This involves selecting and re-indexing the appropriate lead-time predictions from this older forecast by effectively shifting its entire sequence of lead-time values. For example, a prediction LT2 of a forecast issued at time T0 can serve as the LT1 prediction for a new forecast instance at T0 + 1 min. This forwarding strategy is combined with linear interpolation for any remaining gaps. Crucially, throughout this entire homogenization process, all interpolation and forwarding steps are strictly performed on forecasts already available at each given forecast instance, thereby preventing any use of future information that would compromise the integrity of real-time forecasting.

This study investigates four distinct approaches to combine individual forecasting systems. The primary method is linear regression (LR), which is detailed in Section 4.1. This approach is used for all evaluations in Sections 6.1–6.4. Additionally, three

machine learning approaches are assessed: XGBoost (Section 4.2), long short-term memory network (LSTM) (Section 4.3), and transformer (Section 4.4). The results of a benchmark comparing the distinct approaches are presented in Section 6.5.

Each approach employs an ensemble of models, where individual models are tailored to specific lead times. This specialization enables each model to capture the unique temporal dynamics of its designated forecast horizon. The data preprocessing involves preparing time-series sequences for the machine learning models. For each lead time, the models are provided the comprising forecasts from three distinct systems (satellite-based, ASI-based, and persistence) and relevant solar position information (azimuth and elevation). These raw features, along with the target GHI reference values, are then scaled using standard scaler for each individual lead time column. This method transforms the data by subtracting its mean and dividing by its standard deviation, resulting in a distribution with a mean of zero and a standard deviation of one. The models then perform internal feature extraction from these scaled inputs. This data, organized as time-aware sequences with a duration of 15 min, is then presented as a flattened vector. This enables the models to learn a mapping from this comprehensive context to the target GHI reference values, effectively creating a merged and optimized forecast.

All machine learning models were configured to predict continuous values. For the deep learning models (LSTM and transformer), optimization was performed using MSE as the loss function. For XGBoost, the RMSE was used as the primary evaluation metric during optimization. To optimize model performance, hyperparameter tuning was conducted using Optuna, an open-source framework for efficient hyperparameter optimization [36]. Optuna ran 50 trials, systematically evaluating hyperparameter combinations to minimize forecasting errors. Deep learning models were trained for up to 75 epochs, with early stopping applied to prevent overfitting.

Optimization, training, and inference were conducted on two NVIDIA GeForce RTX 3090 GPUs utilizing multiprocessing parallelization. For the machine learning models, individual lead-time models are bundled in a single ensemble of models. This unified ensemble of models streamlines deployment by functioning as a cohesive single model during inference. The computational requirements for inference are minimal. Inference for a single timestamp takes less than 1 s on a standard CPU, tested on an Intel Xeon Silver 4215R CPU @ 3.20 GHz with 16 cores. This negligible computational overhead enables seamless integration with the existing hardware infrastructure of ASI systems, requiring no additional high-performance hardware and ensuring high practicality for operational deployment.

4.1 | Linear Regression

GHI forecasts are used as features, while GHI ground observations serve as targets in the linear regression model. For each forecast lead time k ($\mathrm{LT_k}$), the model is fitted independently to predict the ground truth GHI, using the forecast values from N sources ($\vec{F}_n$). The model is defined as

$$\widehat{\mathrm{GHI}} = \sum_{n=1}^{N} a_n \cdot \vec{F}_n + b$$

where a_n represents the scalar weight for forecast source n, and b is the scalar bias correction term. The model parameters (a_n and b) are determined by minimizing the RMSE between the predicted GHI and the ground observations

$$\min_{a_n, b} \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(\sum_{n=1}^{N} a_n \cdot F_{n,i} + b - \mathrm{GHI}_i\right)^2}_{\mathrm{LT_k}}$$

where M is the number of observations, $F_{n,i}$ is the i-th value from forecast source n, and GHI_i is the corresponding ground observation at lead time k. The optimization is performed using the 'trust-constr' method from SciPy's minimize function [37], with constraints ensuring that the weights $a_n \in [0, 1]$ and their sum satisfy $0.9 \leq \sum_{n=1}^{N} a_n \leq 1.1$. The bias term b is constrained to the range -200 to $200\ \mathrm{W/m^2}$.

The weights determined for the two combined systems, Satellite+Persistence and Satellite+ASI + Persistence, are shown in Figure 3. For the Satellite+Persistence syem, ground persistence initially holds a stronger weight, dominating up to LT9. This indicates that for shorter forecast horizons, ground-based data offers greater reliability. Beyond LT9, a significant shift occurs; the satellite forecast becomes increasingly dominant, with its weights stabilizing between 70% and 80%. Simultaneously, the bias correction term exhibits a rise until LT15, suggesting that initial systematic errors become more pronounced and require greater adjustment over time. After LT15,

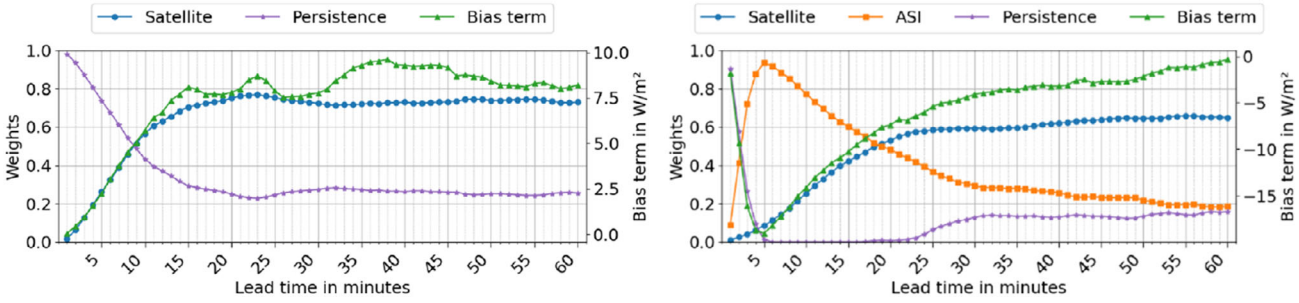

FIGURE 3 | Derived weights and bias term used for the two combinations Satellite+Persistence (left) and Satellite+ASI + Persistence (right) discretized over lead times. The values have been derived by linear regression with data from July 2020 (OLDON).

the bias correction stabilizes in the range of 7–10 W/m^2. This stabilization implies that a correction is consistently needed for longer lead times. However, the systematic biases also become stable, allowing for more reliable and consistent adjustments.

In the case of the Satellite+ASI + Persistence system, the persistence forecast is initially given a very high weighting. This drops rapidly to almost 0% by LT5 and remains negligible until LT22. In contrast, the ASI forecast exhibits almost inverse behavior, rapidly gaining prominence and overtaking the persistence forecast between LT2 and LT3. It then reaches a dominant weight of over 90% at LT5. This emphasizes the critical role of the ASI for very short-term forecasts. Subsequently, the ASI's weight decreases continuously, stabilizing at around 19% by LT60. Meanwhile, the satellite-based forecast, starting near 0%, steadily increases its influence, surpassing the ASI forecast at LT19. After LT20, its curve flattens, stabilizing at around 63%, thereby establishing satellite data as the primary contributor for longer lead times. Interestingly, persistence shows a resurgence after LT22, reaching around 16% at LT60. The bias correction term starts at around -1.8 W/m^2 but rapidly decreases to a strong negative value of approximately -19 W/m^2 at LT5. Thereafter, its magnitude decreases steadily, eventually moving towards 0 W/m^2 at LT60, which suggests a diminishing bias over the forecast horizon.

4.2 | XGBoost

The core of this approach is an ensemble of XGBoost models, a gradient boosting framework known for its efficiency and performance in structured data prediction tasks [38].

The tuned hyperparameters of the XGBoost model are listed in Table 4. These parameters control crucial aspects of the boosting process. *n_estimators* determines the number of boosting rounds, *max_depth* limits the complexity of individual trees, and *learning_rate* scales the contribution of each tree. Additionally, *subsample* and *colsample_bytree* control the proportion of training instances and features used for building each tree, respectively, helping to reduce variance. *min_child_weight* specifies the minimum sum of instance weight required in a child node, influencing the tree's structure, while gamma determines the minimum loss reduction needed to make a further partition on a leaf node.

TABLE 4 | Used hyperparameter ranges on XGBoost for the optimization of the training process.

Parameter	Hyperparameter ranges	Used tuned parameters
n_estimators	[50, 500]	158
max_depth	[3, 10]	3
learning_rate	[0.01, 0.3]	0.1837
subsample	[0.5, 1.0]	0.9062
colsample_bytree	[0.5, 1.0]	0.9884
min_child_weight	[1, 10]	1
gamma	[0, 5]	0.9297

4.3 | LSTM

LSTMs are a type of RNN particularly well-suited for processing sequential data, making them a natural fit for time-series forecasting tasks [39]. LSTMs inherently process a sequence of past observations to make a future prediction.

The tuned hyperparameters are presented in Table 5. These include *num_layers*, which determines the depth of the LSTM network, *hidden_dim*, specifying the number of memory units in each LSTM layer, and *dropout*, a regularization technique to prevent overfitting by randomly omitting a fraction of neurons during training. The *learning_rate* controls the step size during model weight updates, and *batch_size* defines the number of samples processed before the model's weights are updated.

4.4 | Transformer

The transformer model represents a deep learning architecture based on the self-attention mechanism, which has demonstrated significant success in sequential data processing [40]. Unlike recurrent networks like LSTMs, transformers process sequences in parallel, allowing them to capture long-range dependencies more efficiently.

To prepare the inputs, a linear input embedding layer first transforms the raw features into a higher-dimensional space. A Positional Encoding layer is then applied to infuse the transformer model with information about the order or position of the elements within the sequence, as transformers inherently lack this sequential understanding. The transformer then uses its encoder layers, composed of multihead self-attention mechanisms and feed-forward networks, to learn complex relationships and interdependencies across the historical input features.

The optimized hyperparameters are detailed in Table 6. Key parameters include *d_model*, which defines the dimensionality of the input and output of the attention layers and the feed-forward networks, *nhead*, which represents the number of attention heads in the multihead attention mechanism, and *num_encoder_layers*, which determines the number of stacked encoder layers in the transformer architecture. *dim_feedforward* specifies the size of the hidden layer in the feed-forward networks, and *dropout* is used for regularization. The *learning_rate* controls the optimization step size, while *batch_size* dictates the number of samples processed per training iteration.

TABLE 5 | Used hyperparameter ranges on LSTM for optimization of the training process.

Parameter	Hyperparameter ranges	Used tuned parameters
num_layers	[1, 3]	2
dropout	[0.1, 0.5]	0.2465
hidden_dim	[16, 256]	183
learning_rate	[1e−4, 1e−2]	0.0002591
batch_size	[16, 32, 64, 128]	64

TABLE 6 | Used hyperparameter ranges on transformer for optimization of the training process.

Parameter	Hyperparameter ranges	Used tuned parameters
d_model	[16, 256]	32
n_head	—	4
num_encoder_layers	[1, 4]	2
dropout	[0.1, 0.5]	0.2188
learning_rate	[1e−4, 1e−2]	0.002386
batch_size	[16, 32, 64, 128]	64
dim_feedforward	[64, 1024]	166

5 | Used Error Metrics

The validation uses conventional error metrics, including RMSE, MAE, and MBE, as well as skill scores based on RMSE and MAE, with persistence serving as the baseline [10]:

$$\mathrm{RMSE} = \sqrt{1/n \sum_{i=1}^{n} (\hat{y}_i - y_i)^2}$$

$$\mathrm{MAE} = 1/n \sum_{i=1}^{n} |\hat{y}_i - y_i|$$

$$\mathrm{MBE} = 1/n \sum_{i=1}^{n} \hat{y}_i - y_i$$

$$\mathrm{SS} = 1 - \frac{S_{\mathrm{model}}}{S_{\mathrm{benchmark}}}$$

where $\hat{y}$ is the predicted irradiance and y the corresponding measured irradiance. The skill score is defined as SS. S represents a suitable metric, such as RMSE or MAE.

6 | Results

This section presents the benchmark results. Section 6.1 provides an overview of the general findings. Section 6.2 assesses the impact of varying irradiance conditions, while Section 6.3 examines the effects of temporal aggregation. Section 6.4 analyzes spatial influences. Section 6.1–6.4 consider only linear regression to combine the distinct sources. Section 6.5 evaluates the performance of different merging methods.

6.1 | General Benchmark Forecasts

Figure 4 illustrates the error metrics and skill scores of the evaluated forecast systems over the lead times, while Table 7 presents their aggregated results. For this aggregation analysis, the forecast horizon is divided into five distinct lead time (LT) ranges. Four of these cover individual quarters of the total horizon, while the fifth spans the entire range from LT1 to LT60. The average error metrics and their corresponding standard deviations for these ranges are listed. If aggregated metrics for multiple lead times are specified directly in the main text, the standard deviation across those lead times is always provided.

The persistence forecast serves as a fundamental baseline. As anticipated, it exhibits the highest overall RMSE (160.6 ± 28.2 W/m^2) and MAE (106.0 ± 24.0 W/m^2), with its accuracy rapidly degrading beyond the very shortest lead times. However, it maintains a consistently very small slightly positive MBE (2.0 ± 1.1 W/m^2). The small bias observed in persistence is expected due to its fundamental nature, indicating it lacks any inherent systematic tendency to consistently over- or under-predict GHI.

Individual advanced forecasts present complementary characteristics. The ASI forecast demonstrates high accuracy for very short lead times (LT1−15), achieving the lowest RMSE (99.5 ± 17.5 W/m^2) and MAE (58.8 ± 11.0 W/m^2) in this range, coupled with strong positive skill scores. For longer LTs, the ASI system continues to outperform the satellite system in terms of RMSE up to

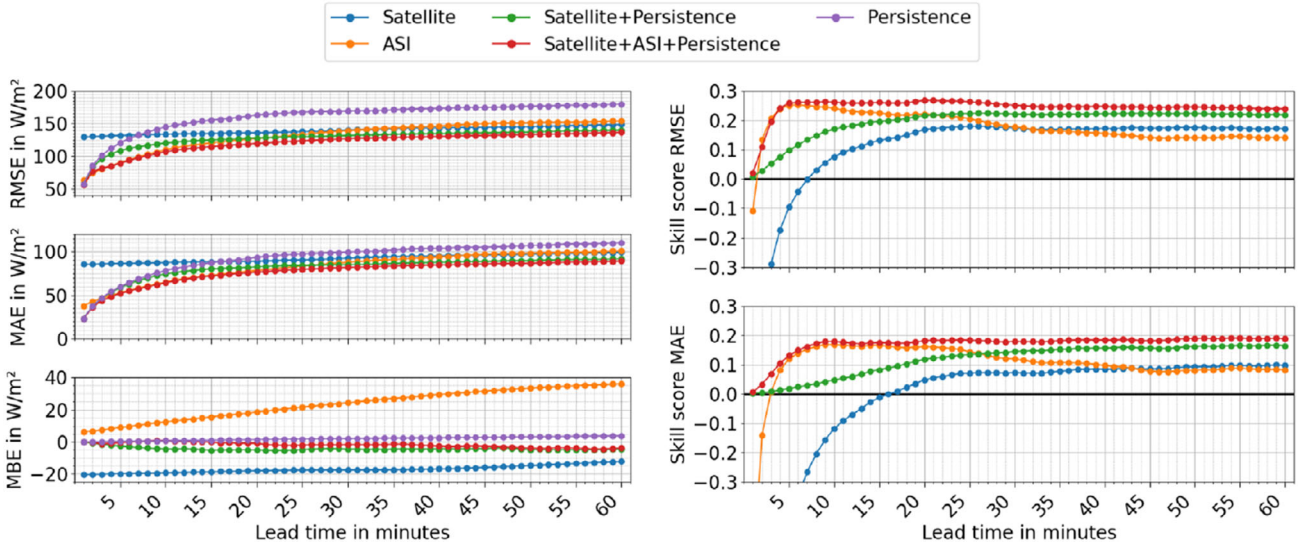

FIGURE 4 | Error metrics and skill scores over lead time. In terms of skill scores, some of the lower LTs for the ASI and satellite forecasts have been cropped. This is done to improve the readability of the skill scores.

TABLE 7 | Average ± standard deviation of error metrics over different LT ranges.

Error metric[a]	Used LT	Satellite	ASI	Satellite+ Persistence	Satellite+ASI+ Persistence	Persistence
RMSE [W/m²] ↓	LT1 to LT15	132.8±1.7	99.5±17.5	109.8±18.6	96.9±16.6	127.9±28.2
	LT16 to LT30	137.1±1.5	130.8±5.6	129.0±2.0	121.8±3.5	164.8±4.3
	LT31 to LT45	142.9±1.2	145.2±3.1	134.2±1.2	129.9±1.5	172.4±1.8
	LT46 to LT60	146.5±1.3	152.1±1.0	138.1±1.3	134.2±1.4	177.3±1.4
	LT1 to LT60	139.8±5.5	131.9±22.3	127.8±14.3	120.7±16.8	160.6±24.0
MAE [W/m²] ↓	LT1 to LT15	87.1±0.9	58.8±11.0	63.8±17.0	56.6±13.9	66.7±19.1
	LT16 to LT30	90.1±1.3	81.3±4.4	83.4±1.6	78.1±2.8	95.2±3.7
	LT31 to LT45	94.7±1.0	92.8±2.6	87.2±1.1	84.3±1.3	103.0±1.8
	LT46 to LT60	97.7±0.9	99.0±1.0	90.4±0.9	87.6±1.0	107.9±1.4
	LT1 to LT60	92.4±4.3	83.0±16.5	81.2±13.4	76.6±14.0	93.2±18.7
MBE [W/m²] ↓	LT1 to LT15	-19.4±0.6	10.9±3.0	-3.3±1.6	0.0±0.6	0.5±0.3
	LT16 to LT30	-17.7±0.3	20.3±2.6	-4.9±0.4	-1.5±0.6	1.6±0.3
	LT31 to LT45	-16.9±0.5	28.4±2.1	-4.5±0.2	-2.1±0.6	2.5±0.3
	LT46 to LT60	-13.9±1.1	34.2±1.3	-4.7±0.4	-3.8±0.5	3.4±0.3
	LT1 to LT60	-17.0±2.1	23.4±9.1	-4.3±1.0	-1.9±1.5	2.0±1.1
Skill Score (RMSE)↑	LT1 to LT60	0.10±0.22	0.18±0.05	0.20±0.05	0.24±0.04	-
Skill Score (MAE)↑	LT1 to LT60	-0.06±0.42	0.10±0.11	0.12±0.05	0.17±0.03	-

[a]The color coding evaluates each row separately and ranges from dark green (best outcome) to dark red (worst outcome).

LT32 and in terms of MAE up to LT43. Beyond these lead times, the satellite system very slightly outperforms the ASI system.

The ASI forecast exhibits a constantly increasing positive MBE, ranging from around 6 W/m^2 at LT1 to approximately 36 W/m^2 at LT60, averaging 23.4 ± 9.1 W/m^2 overall. As described in Section 3.2, the ASI model was trained using data from a location in southern Spain and applied directly to a location in northern Germany without local training. At the training location in Spain, the ASI system exhibits a barely noticeable MBE [15]. This contrast to the significant positive MBE in Germany may indicate weaknesses in generalization for sites with notably different ambient conditions, mainly due to differences in aerosol and cloud-type characteristics. The site in southern Spain features lower Linke turbidity and predominantly cumulus and cirrus clouds [15, 41], while Germany exhibits higher turbidity and more frequent stratocumulus and nimbostratus clouds (see Appendix A1). Section 6.2 discusses the influence of different conditions on the forecast in more detail. To mitigate the inherent bias of the ASI forecasting model, a hybrid ASI + Persistence approach was also developed. This combination is practical, as both models rely on local irradiance time series data as input, allowing for straightforward integration. The results of this state-of-the-art [15] ASI + Persistence hybrid model are presented in Appendix A2 (see Figures A5, A6). The ASI + Persistence hybrid dramatically reduces the MBE. Over the full 60 min horizon, this hybrid achieves an overall negative bias (−3.43 ± 5.2 W/m^2). This substantial bias reduction suggests that the integration with the persistence component effectively counteracts ASI's inherent tendency to over-predict, leading to a more centered and balanced forecast.

Conversely, the satellite forecast offers greater stability at longer LTs. While the satellite component gains dominance in the optimal combined forecast structure around LT19 (see weights in Section 4.1), its standalone error performance in terms of RMSE and MAE as already stated overtakes that of ASI at much later lead times. The overall MAE Skill Score is negative (−0.06 ± 0.42 W/m^2). This negative skill is primarily driven by its performance in the very short lead times LT1−15, where the MAE Skill Score is −0.24 ± 0.81 W/m^2. However, beyond LT15, the satellite's MAE Skill Score becomes positive, reaching values up to around 0.1.

The integration of distinct forecasts yields significant performance enhancements. The Satellite+Persistence system shows a general improvement in accuracy, with lower overall RMSE (127.8 ± 14.3 W/m^2) and MAE (81.2 ± 13.4 W/m^2) than either individual component. Notably, its MBE is substantially reduced (−4.3 ± 1.0 W/m^2) compared to standalone satellite. This reflects previously found results and therefore, the state-of-the-art in satellite-based forecasting.

The Satellite+ASI + Persistence system, however, consistently emerges as the strongest performer across all lead times and metrics. It achieves the lowest overall RMSE (120.7 ± 16.8 W/m^2) and MAE (76.6 ± 14.0 W/m^2), alongside the highest overall RMSE (0.24 ± 0.04) and MAE (0.17 ± 0.03) skill scores. Maximum values of 0.27 (RMSE skill score at LT20) and 0.19 (MAE skill score at LT54) are achieved. This superior accuracy is underpinned by its bias mitigation, with its MBE consistently remaining closest to zero across all lead times.

The analysis of standard deviations across lead times offers insight into the plausibility and consistency of the model's performance. The low standard deviations for RMSE, MAE, and particularly MBE in the Satellite+ASI + Persistence system indicate not only superior average accuracy but also a high degree of plausibility and consistency in its performance. This means the errors are less volatile across different lead times, providing a more reliable and trustworthy forecast. This synergistic effect, leveraging complementary strengths (short-term from ASI, longer-term from satellite, and baseline from persistence) and balancing inherent biases, is crucial for developing high-quality GHI forecasts.

To confirm statistical significance, Wilcoxon signed-rank tests [42] were conducted on RMSE values across all lead times, with results listed in Table 8. All approaches significantly

TABLE 8 | Wilcoxon signed-rank test results based on paired RMSE values.

	Persistence[a]		Satellite		ASI		Satellite+ Persistence		Satellite+ASI + Persistence	
Persistence	—	—	<0.001	−20.8 W/m^2	<0.001	−28.7 W/m^2	<0.001	−32.8 W/m^2	<0.001	−39.9 W/m^2
Satellite	—	—	—	—	0.023	−7.9 W/m^2	<0.001	−12 W/m^2	<0.001	−19.1 W/m^2
ASI	—	—	—	—	—	—	0.005	−4.1 W/m^2	<0.001	−11.2 W/m^2
Satellite + Persistence	—	—	—	—	—	—	—	—	<0.001	−7.1 W/m^2

[a]The first value in each cell represents the *p*-value, while the second represents the mean RMSE difference. The system indicated in the rows is always subtracted from the system indicated in the columns of the matrix to calculate the differences.

outperformed the Persistence. The hybrid models also outperformed the individual Satellite and ASI system.

The impact of the satellite's enhanced resolution and update frequency is also examined. For this purpose, only the forecasts available at the issue times of the satellite system are utilized. The analysis reveals no significant differences in the results. The skill scores and overall tendencies of the different forecasting systems are comparable. The corresponding plots are provided in Appendix A3.

To provide context for the results discussed in this work, a comparison is made with the findings of Straub et al. [21]. It should be noted that such comparisons are generally constrained by differences in the ambient conditions of the datasets employed. These differences may considerably influence the performance of models. The skill score of a model may vary notably depending on the distribution of variable and less variable conditions within the datasets chosen. Thus, comparisons should focus on identifying general trends. Figure 5 presents a comparison of the RMSE-based skill score from this evaluation with that reported by Straub et al. [21]. Given that their forecast extends only to a 15 min horizon, this comparison is similarly limited to that horizon. Both studies utilize the Heliosat method, relying on MSG data for satellite-based forecasting. Nevertheless, ref. [21] reports significantly higher skill score than this study. On the other hand, the ASI system in this study achieves a significantly higher skill score than that in ref. [21]. As noted in the introduction, the two studies use fundamentally different ASI model implementations. For hybrid systems, the skill scores appear broadly similar between the two studies. Although the methodology for satellite forecasting is identical, the varying skill scores highlight the difficulties involved in comparing results from different datasets. Nevertheless, general trends in hybrid forecasting systems remain consistent between the studies.

6.2 | Influence of Prevailing Sky Conditions

The analysis of forecast system performance under distinct sky conditions provides a more granular understanding of their strengths and limitations. The classification approach of sky conditions into three-bundled classes is explained in Section 2. Figure 6 presents the error metrics and corresponding skill scores, discretized by the distinct classes. Tables with aggregated results are provided in Appendix A4. This also includes results from Wilcoxon signed-rank tests, which confirm statistical significance.

6.2.1 | Bundled Class 1 - Mostly/Almost Clear Sky with Low Variability

Under the highly clear and low-variability conditions of bundled class 1, RMSE and MAE values are generally the lowest across all systems, as expected for easier forecasting scenarios. The persistence model, in particular, demonstrates low errors in the lead time range LT1–15, with an RMSE of $19.0 \pm 4.4\,\mathrm{W/m^2}$ and MAE of $7.9 \pm 2.4\,\mathrm{W/m^2}$. However, its performance rapidly

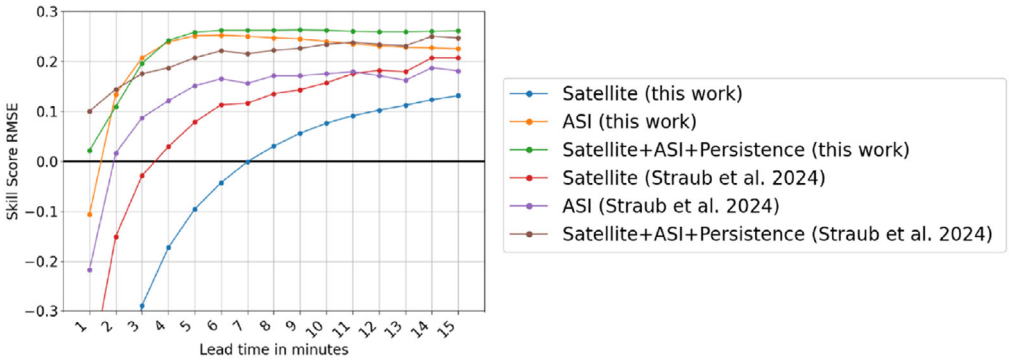

FIGURE 5 | RMSE-based skill score over lead time. A comparison of this work with that of Straub et al. [21]. Some of the lower LTs for the satellite forecasts have been cropped. This is done to improve the readability of the skill scores.

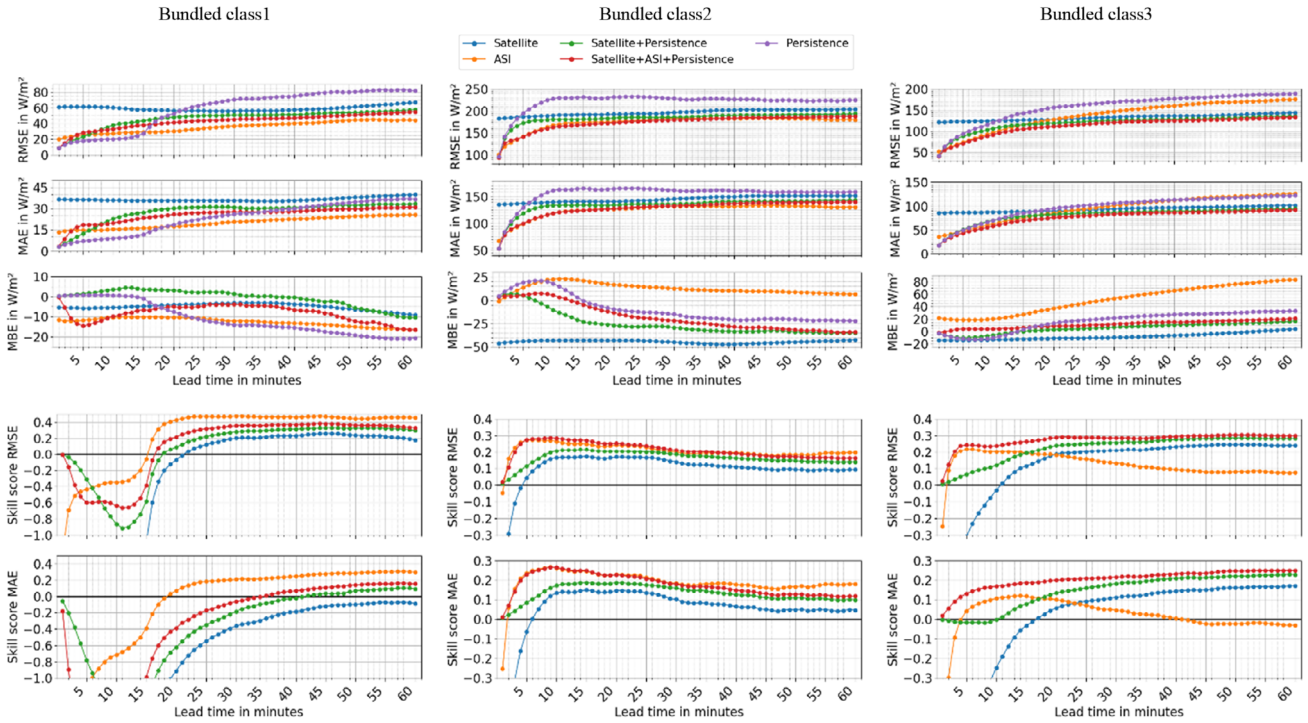

FIGURE 6 | Error metrics and skill scores over lead time discretized in the three variability conditions high clear sky index (bundled class1), intermediate clear sky index (bundled class2), and low clear sky index (bundled class3). In terms of skill scores, some of the lower LTs for the ASI and satellite forecasts have been cropped. This is done to improve the readability of the skill scores.

degrades thereafter, with RMSE jumping significantly. This behavior is not visible for MAE. This disparity of good MAE but higher RMSE for longer LTs illustrates nicely how some outliers are strongly penalized by RMSE but have a limited effect on MAE. Consequently, the overall RMSE for persistence increases to $58.1 \pm 10.4\,\mathrm{W/m^2}$, while the MAE remains relatively lower at $23.4 \pm 10.7\,\mathrm{W/m^2}$. This underscores how, under highly stable clear-sky conditions, a simple persistence model can be very difficult to outperform in terms of MAE. Additionally, persistence maintains a small positive MBE.

The ASI model shows the best overall performance in terms of RMSE, achieving $35.6 \pm 7.2\,\mathrm{W/m^2}$ and yielding the highest RMSE-based skill score (∼0.39). It also performs best in MAE, with a value of $20.2 \pm 3.9\,\mathrm{W/m^2}$ and a corresponding skill score of ∼0.14. Notably, ASI exhibits a negative overall MBE ($-12.5 \pm 2.0\,\mathrm{W/m^2}$), which is a significant contrast to its overall positive MBE seen across all conditions.

The satellite forecast performs poorly in bundled class 1. Its overall RMSE ($59.2 \pm 3.0\,\mathrm{W/m^2}$) is on par with persistence, resulting in a slightly negative RMSE skill score. Its MAE is significantly worse at $36.5 \pm 1.3\,\mathrm{W/m^2}$, and it shows a low negative MBE ($-5.0 \pm 1.6\,\mathrm{W/m^2}$).

The Satellite+Persistence hybrid outperforms persistence in terms of overall RMSE ($46.3 \pm 11.3\,\mathrm{W/m^2}$), but not in MAE ($27.8 \pm 7.3\,\mathrm{W/m^2}$). Similarly, the Satellite+ASI + Persistence hybrid system ranks as the second-best performer in terms of RMSE ($42.7 \pm 9.6\,\mathrm{W/m^2}$), yet its MAE ($25.6 \pm 5.5\,\mathrm{W/m^2}$) remains somewhat higher than that of persistence. However, it does outperform persistence in terms of MAE after LT34.

These results highlight that, under highly stable and clear-sky conditions, the simplicity of the persistence model makes it a robust baseline for forecasting the immediate future. This is especially true when the model's performance is evaluated using MAE. This can be attributed to the inherent nature of persistence, which, by not introducing additional variability. In contrast, more complex models optimized through least squares tend to learn an averaged response to variability. While beneficial for dynamic scenarios, this can introduce a degree of average variability even into highly stable clear-sky conditions.

6.2.2 | Bundled Class 2 - Intermediate Clear Sky Index Conditions (Partly Cloudy with Intermediate/High Variability)

These partly cloudy conditions represent the most challenging forecasting scenario, as evidenced by the highest RMSE and MAE values across all systems compared to other classes. Persistence, in particular, shows an increase in errors compared to bundled class 1, with its RMSE ($220.9 \pm 19.7\,\mathrm{W/m^2}$) and MAE ($155.5 \pm 19.7\,\mathrm{W/m^2}$) being several times higher than in clear-sky conditions. This surge in error highlights persistence's inability in handling conditions with significant variability.

The Satellite+ASI + Persistence hybrid system consistently delivers extremely strong performance, achieving the lowest RMSE ($173.8 \pm 17.8\,\mathrm{W/m^2}$) and the highest RMSE skill score (0.213).

In terms of RMSE, the ASI forecast is nearly on par with the Satellite+ASI + Persistence hybrid, showing only a slightly higher overall RMSE of $174.0 \pm 17.0\,\mathrm{W/m^2}$. In terms of MAE, the ASI forecast slightly outperforms the hybrid system, with

a lower MAE of 125.1 ± 14.0 W/m^2 and a corresponding higher MAE skill score of ∼0.20, compared to the hybrid's MAE of 127.8 ± 16.5 W/m^2 and skill score of ∼0.18.

In terms of bias, the satellite forecast displays a very large negative overall MBE (−44.3 ± 1.4 W/m^2), indicating significant under-prediction. In contrast, the ASI forecast maintains a positive MBE (12.8 ± 5.4 W/m^2), consistent with its general tendency. Despite these opposing tendencies, the Satellite + ASI + Persistence hybrid still exhibits a significant negative MBE (−17.8 ± 14.2 W/m^2), suggesting that the large negative bias of the satellite component strongly influences the combined output under these specific conditions.

6.2.3 | Bundled Class 3 - Low Clear Sky Index Conditions (Mostly/Almost Overcast with Intermediate/Low Variability)

In predominantly overcast conditions, RMSE and MAE values are generally intermediate compared to the other classes. Notably, for lead times greater than LT20, the satellite-based forecasts often demonstrate a clear advantage over the ASI forecasts. The Satellite+ASI + Persistence hybrid system exhibits the best overall performance, achieving the lowest error magnitudes (RMSE: 112.4 ± 21.4 W/m^2 and MAE: 76.1 ± 17.5 W/m^2) and the highest skill scores (RMSE: ∼0.28 and MAE: ∼0.22).

Bias behavior in this class is particularly striking. The ASI forecast exhibits an extremely large positive overall MBE (51.2 ± 22.4 W/m^2), indicating substantial over-prediction under overcast conditions. This considerable over-prediction directly impacts its MAE performance, especially at longer lead times. For example, while ASI's MAE is 56.5 ± 13.1 W/m^2 for LT1−15, it steadily increases, reaching 122.2 ± 2.4 W/m^2 for LT46−60. At these higher lead times, ASI's MAE actually surpasses that of persistence (119.4 ± 2.2 W/m^2) and is notably higher than the satellite forecast (99.9 ± 1.3 W/m^2) and the Satellite+ASI + Persistence hybrid (89.9 ± 1.3 W/m^2). This degradation in performance under overcast conditions is likely exacerbated by the fact that the ASI model was trained on data from southern Spain (see Section 3.2), a region characterized by a comparatively low frequency of overcast sky conditions. While the model exhibits strong generalization capabilities across varying conditions, its direct application to northwest Germany, a region characterized by significantly more frequent and prolonged overcast periods, initially presents a systematic over-prediction. As noted in Section 6.1, the pronounced MBE of the ASI system can be effectively mitigated by combining it with persistence, which primarily functions as a bias correction mechanism (see Appendix A2).

The satellite forecast has an overall negative MBE (−7.7 ± 5.3 W/m^2). Due to ASI's very large positive MBE, the Satellite+ASI + Persistence hybrid system still results in an overall positive MBE (11.6 ± 5.9 W/m^2).

6.3 | Temporal Aggregation

To further evaluate the performance of GHI forecasts, a temporal aggregation procedure was implemented to align forecast and reference data over equivalent time windows. This approach is particularly relevant for many solar energy applications that rely on smoothed power values rather than instantaneous predictions [43]. Shorter intervals (5–10 min) support real-time inverter control and ramp mitigation operations, while longer intervals (15–30 min) align with energy market decisions, such as real-time dispatch or back-up generator activation in hybrid systems.

For each forecast lead time (LT), both the GHI forecast and a persistence forecast are aggregated over a specified historical window of duration (W). Consequently, the corresponding reference GHI observations are aggregated over the same window to ensure identical time intervals for direct comparison. The aggregation is performed using a sliding window keeping the temporal resolution.

Let $F_{t_{issue},LT}$ denotes the instantaneous forecast issued at time t_{issue} for a valid time of $t_{issue} + LT$. Let $R_{t_{obs}}$ represents the instantaneous reference (observed) GHI at time t. For an aggregation window of duration W ending at time t, the aggregated reference value ($\overline{R}_{t,W}$) is calculated as the mean of all instantaneous reference observations within the window $(t - W, t]$:

$$\overline{R}_{t,W} = \frac{1}{N_R} \sum_{k=0}^{N_R - 1} R_{t - k \cdot \Delta t_R}$$

here, N_R is the number of reference observations in the window, and Δt_R is the sampling interval of the reference data.

Similarly, the aggregated forecast value ($\overline{F}_{t,W,L}$) for lead time (LT) is computed by averaging instantaneous forecasts whose valid times fall within the same window $(t - W, t)$. The aggregated forecast at valid time (t) is thus the mean of forecasts whose issue times correspond to valid times within the aggregation window.

$$\overline{F}_{t,W,LT} = \frac{1}{N_F} \sum_{k=0}^{N_F - 1} F_{(t - LT - k \cdot \Delta t_F), LT}$$

here, N_F is the number of forecasts aggregated, and Δt_F is the frequency with which the forecasts are issued. The term $(t - LT - k \cdot \Delta t_F)$ represents the issue time of each instantaneous forecast included in the sum, ensuring that its corresponding valid time $((t - LT - k \cdot \Delta t_F) + LT)$ falls within the aggregation window $(t - W, t)$ ending at t.

After temporal alignment, the error metrics are computed to quantify forecast performance. In this work the aggregation windows W of 5, 10, 15, and 30 min are used. The error metrics of this analysis are shown in Figure 7. Furthermore, Appendix A5 contains tables with averaged results, along with an additional Figure A8 displaying the skill scores. A clear and consistent trend observed across most models and error metrics is that temporal aggregation universally reduces forecast errors as the aggregation window increases. This confirms that averaging forecasts over longer periods effectively smooths out short-term fluctuations and noise, leading to more stable and accurate predictions in terms of power.

The satellite system generally outperforms the persistence model in overall RMSE. However, for shorter aggregation windows

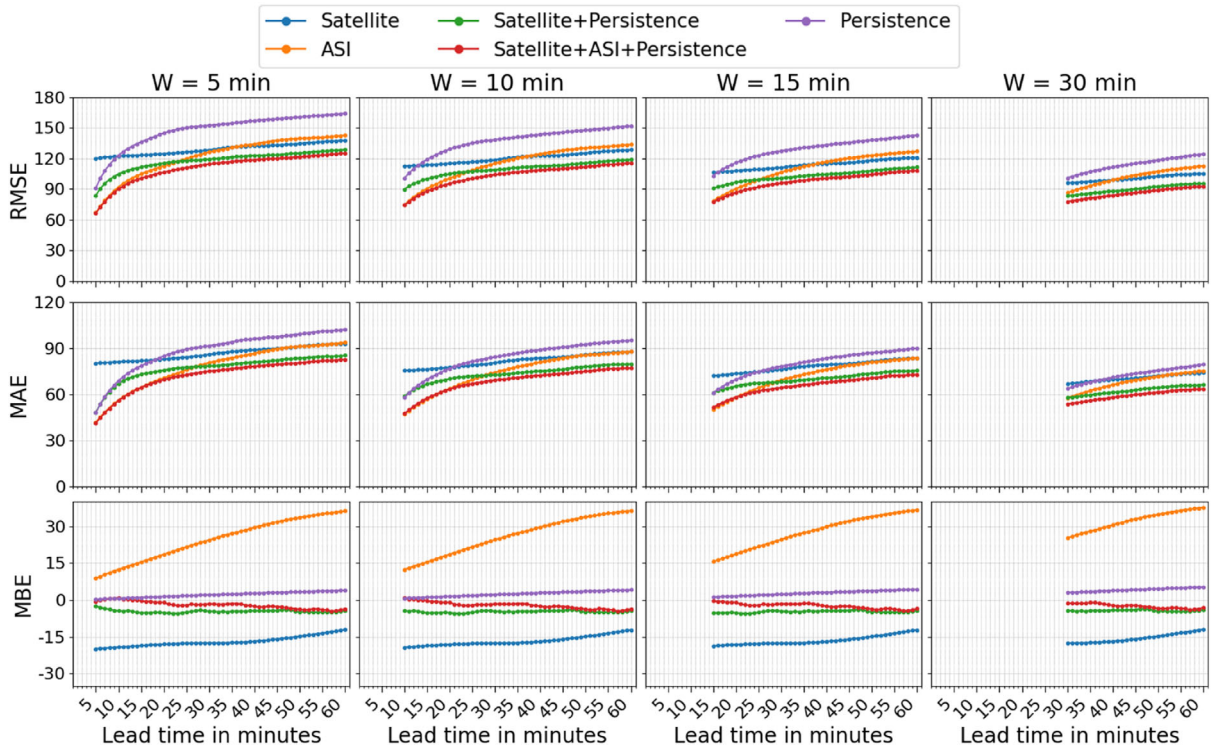

FIGURE 7 | Error metrics over lead time, considering different time windows for temporal aggregation. In terms of skill scores, some of the lower LTs for the ASI and satellite forecasts have been cropped. This is done to improve the readability of the skill scores.

(5 and 10 min) and lead times ≤15 min (LT15), the persistence model achieves lower RMSE than the satellite system. A similar trend is observed for MAE, where persistence outperforms the satellite system for lead times up to 30 min (LT30), depending on the window size.

The ASI system consistently surpasses the persistence model, exhibiting particularly strong performance for window sizes up to 15 min and lead times ≤LT15. For lead times >LT30, its RMSE performance is occasionally weaker than the satellite system, but its MAE performance remains comparable.

The Satellite+Persistence hybrid outperforms the persistence model but is generally surpassed by the ASI system for lead times <LT30.

The Satellite+ASI + Persistence hybrid delivers the lowest overall RMSE and MAE across all aggregation levels, demonstrating robust performance even with smoothed forecasts.

6.4 | Spatial Influence

This evaluation focuses on assessing the spatial transferability of the Hybrid forecast derived from the OLDON site to three additional, nearby sites. The primary objective of this evaluation is to understand the effective spatial coverage of a Hybrid forecast based on a single ASI installation. This sheds light on the practical considerations for large solar power plants, specifically determining the distance at which the advantage of a central ASI in a hybrid forecast diminishes, potentially necessitating the installation of additional ASIs within the solar field to maintain forecasting edge.

The three additional sites, OLCLO, OLUOL, and OLJET, are equipped with reference GHI measurements, allowing for a comparison against site-specific satellite forecasts. The OLDON site, conversely, serves as the base for the hybrid forecast, utilizing an ASI and ground station data as input for its ASI component, along with persistence and satellite data. The distances between OLDON and these evaluation sites are given in Table 1.

Figure 8 shows the error metrics of this analysis over the lead times, and Table 9 lists aggregated results over the entire lead time range. For OLCLO and OLUOL, which are at similar distances from OLDON (∼3.8 km), the hybrid consistently shows lower RMSE and MAE than the satellite-only forecast across nearly all lead times (exception lead times <LT3 for RMSE). For OLCLO, it achieves a reduction of about 9% in RMSE and 13% in MAE. For OLUOL, the hybrid yields an overall reduction of about 4% in RMSE and 7% in MAE. OLCLO is located south of OLDON, while OLUOL lies to the west. The predominant cloud movement in this region is from west to east. The lateral positioning of OLCLO relative to OLDON and the primary cloud movement direction may partially explain its superior performance, as ASI observations at OLDON are likely more relevant to OLCLO than to OLUOL.

For OLJET (8.63 km distance), the RMSE of the satellite system is slightly lower than that of the hybrid system for most lead times, while the hybrid consistently achieves a lower MAE. Overall, the satellite system outperforms the hybrid by approximately 1% in RMSE, but the hybrid surpasses the satellite system by about 5% in MAE.

This analysis suggests that even for large solar fields, a single ASI can improve the overall forecast for the entire plant. However,

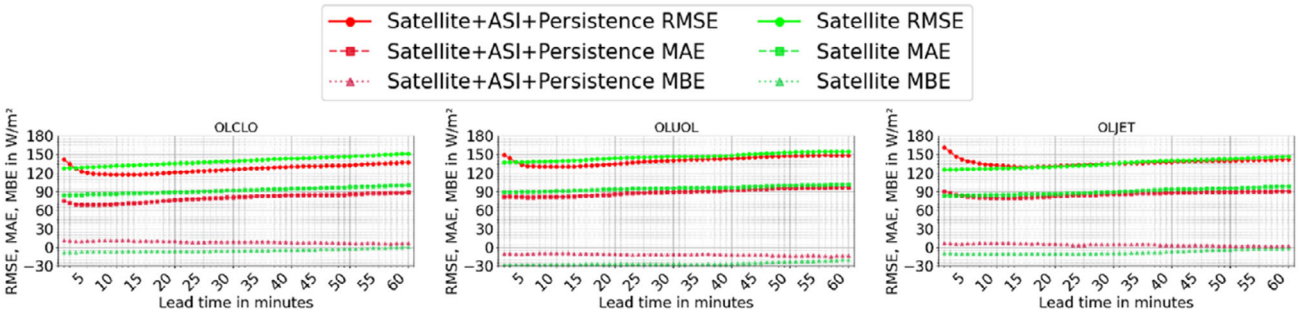

FIGURE 8 | Error metrics over lead time. The red metrics represent the errors of the merged forecasts derived from the OLDON site and validated at the OLCLO, OLUOL, or OLJET sites. The corresponding error metrics for the satellite forecasts, derived for and validated at the sites OLCLO, OLUOL, and OLJET, are shown in green.

TABLE 9 | Average ± standard deviation of error metrics of error metrics over the entire range from LT1 to LT60.

Error metric[a]	Hybrid OLDON@ OLCLO	Satellite OLCLO	Hybrid OLDON@ OLUOL	Satellite OLUOL	Hybrid OLDON@ OLJET	Satellite OLJET
RMSE [W/m²] ↓	126.8±6.5	139.4±6.8	140.2±6.5	146.3±5.6	137.1±5.6	135.4±6.8
MAE [W/m²] ↓	79.2±6.4	92.0±4.8	88.9±5.6	95.5±3.7	85.7±3.6	90.2±4.9
MBE [W/m²] ↓	9.0±1.6	-4.7±2.4	-11.3±1.3	-25.5±2.2	4.8±1.5	-7.4±3.0

[a]The color coding evaluates each row and site pair separately and ranges from dark green (best outcome) to dark red (worst outcome).

accuracy decreases as the distance increases. Therefore, for very large solar fields, installing additional ASIs may be necessary to maintain significant improved accuracies across the entire area.

6.5 | Benchmark Approaches to Combine Forecasts

Figure 9 and Table 10 show the evaluation of the hybrid forecast's performance when using different merging algorithms: LR, XGBoost, LSTM, and transformer. The aim is to determine if more advanced machine learning models can enhance the merging process compared to a standard linear regression.

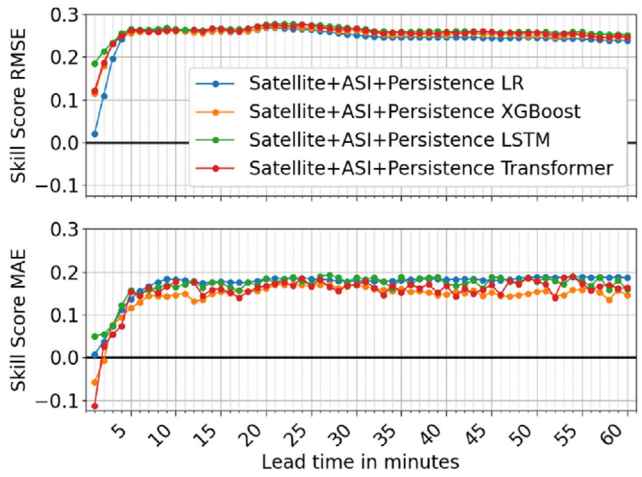

FIGURE 9 | Skill scores over lead time based on different models used to merge the individual forecasts.

TABLE 10 | Average ± standard deviation of error metrics over the entire range from LT1 to LT60.

Error metric[a]	LR	XGBoost	LSTM	Transformer
RMSE [W/m²] ↓	121.8 ±16.9	120.8 ±17.2	119.7 ±17.4	120.4 ±17.2
MAE [W/m²] ↓	77.6 ±14.2	80.3 ±14.9	78.1 ±14.5	79.5 ±14.6
MBE [W/m²] ↓	-1.7 ±1.6	-5.5 ±1.5	-2.8 ±3.0	-1.8 ±5.1
Skill Score (RMSE) ↑	0.24 ±0.04	0.25 ±0.02	0.26 ±0.01	0.26 ±0.02
Skill Score (MAE) ↑	0.17 ±0.03	0.14 ±0.04	0.17 ±0.03	0.15 ±0.05

[a]The color coding evaluates each row separately and ranges from dark green (best outcome) to dark red (worst outcome).

All merging algorithms demonstrate positive skill scores across most lead times, peaking between LT10 and LT15 before stabilizing or slightly declining at longer horizons. For very short lead times (<LT5 minutes), XGBoost, LSTM, and transformer ramp up more quickly and maintain higher skill than LR, which starts lower. Beyond this, LSTM and transformer sustain slightly higher RMSE skill scores. Overall, XGBoost, LSTM and transformer reduce RMSE by approximately 0.8%, 1.8% and 1.1%, respectively.

For MAE skill, LR and LSTM are highly competitive, generally maintaining slightly higher scores across most lead times, while XGBoost and transformer exhibit lower MAE skill initially. Overall, LR outperforms XGBoost, LSTM, and transformer in MAE by approximately 3.4%, 0.6%, and 2.3%, respectively.

The evaluation of merging algorithms for the hybrid indicates no single algorithm excels across all metrics (Table 10) Advanced machine learning models, such as LSTM and transformer, achieve slightly lower RMSE, especially at shorter lead times. However, Linear Regression remains highly competitive, surpassing others in overall MAE and demonstrating the smallest bias. XGBoost consistently underperforms compared to the other methods.

The analysis indicates that for the simple merging task, the added complexity of advanced machine learning models yields modest RMSE improvements, while the simpler LR model remains

highly effective, particularly for minimizing MAE. Deep learning models, especially the transformer, may benefit from larger training datasets. Furthermore, additional features could further improve such deep learning models. However, these findings align with similar studies, which suggest that outperforming LR in this straightforward task is challenging [22].

7 | Conclusion

This study presents the development and validation of a data-driven forecasting methodology that effectively blends satellite forecasts with ground-based ASI and persistence forecasts. It quantifies the performance benefits of this integrated approach and characterizes the generalization capabilities of the underlying ASI model.

A key finding concerns the generalization performance of the ASI system. The ASI model, trained on a dataset from southern Spain (desert climate), demonstrated operational capability when deployed in northwest Germany, a region exhibiting a temperate oceanic climate with significantly different cloud cover characteristics. The ASI system maintained positive skill scores for both RMSE and MAE in the German validation environment, affirming its cross-climatic applicability. This climatic transfer of the ASI system introduced a positive bias, which can be compensated for by combining ASI with persistence.

Furthermore, this study demonstrates that local solar irradiance forecasts derived from satellite data are significantly enhanced when combined with the ASI model.

The individual forecasting components possess complementary characteristics, ASI exhibits superior performance at very short lead times, satellite data provide greater stability for longer forecast horizons, and persistence serves as a fundamental baseline. The hybrid model utilizes the respective advantages, which are reflected in the performance of the Satellite+ASI + Persistence hybrid system. It consistently exhibited optimal performance across all evaluated lead times and metrics, yielding statistically significant improvements over each individual component. Crucially, the hybrid provides forecasts with the high temporal resolution of the ASI system, encompassing both its update rate and the number of leads. But the hybridization process, which optimizes based on metrics like RMSE or MSE, inherently leads to a partially smoothed performance curve.

This superior performance is evidenced by the lowest overall RMSE (120.7 ± 16.8 W/m^2) and MAE (76.6 ± 14.0 W/m^2), alongside the highest skill scores for RMSE (0.24 ± 0.04) and MAE (0.17 ± 0.03), on average over the entire 60 min forecast horizon. This corresponds to an approximate reduction of the RMSE by 13.6% and of the MAE by 17.0% for this particular location compared to the satellite-only forecast, while at the same time increasing the RMSE skill score by 0.14 and the MAE skill score by 0.23. This enhancement is directly attributable to the effective bias mitigation facilitated by the blending process and the strategic utilization of ASI's granular, short-term cloud dynamics alongside satellite data's broader, longer-range cloud tracking capabilities.

The analysis of forecast performance under varying sky conditions confirms these conclusions. Under mainly clear sky conditions, the persistence model demonstrated low errors in both RMSE and MAE for very short lead times (LT1−15). However, for the overall horizon, the ASI model exhibited superior performance in both overall RMSE and MAE, achieving the lowest values among all individual forecasting systems under these conditions. The ASI model also demonstrated superior performance under challenging partly cloudy conditions, where it consistently delivered high skill. This strength significantly benefited the blended hybrid system. The hybrid system's robust performance in these scenarios, including its consistently low RMSE despite inherently high variability, directly leveraged the ASI's capabilities. Moreover, its ability to effectively counteract the ASI's inherent positive bias under predominantly overcast conditions further validates the blending methodology.

Temporal aggregation analyses confirmed the consistent superior performance of the Satellite+ASI + Persistence hybrid, thereby ensuring its applicability for various operational requirements in solar energy.

Spatial analysis provided quantitative insights into the effective spatial coverage of a single ASI installation within a hybrid forecasting framework. An ASI-based hybrid forecast offers performance advantages that remain visible at distances of approximately 4 km. This range is significant, as it encompasses the spatial extension of the majority of existing utility-scale PV plants/parks. However, while these benefits are still clearly present at 4 km, their magnitude is reduced compared to forecasts made directly at the ASI's location, and they begin to diminish further beyond this range. This finding has direct implications for the design and deployment of ASIs in solar power plants. Multiple ASI installations may be required to sustain optimal forecast accuracy across larger, more extensive areas. In general, it is clear that the ASI offers benefits beyond its immediate location. However, our specific quantitative findings are dependent on the characteristics of the used dataset and the site in question.

The comparison of merging algorithms showed that advanced machine learning models like LSTM and transformer yielded in some RMSE improvements, especially at shorter lead times. However, the linear regression method demonstrated comparable performance and even slightly superior performance in MAE. This suggests that for this specific blending task, the increased complexity of more sophisticated models may not translate into substantial performance gains.

The strategic integration of satellite and data-driven ASI forecasts, complemented by persistence-based bias correction, represents an effective strategy for accurate intra-hour solar irradiance forecasting. The demonstrated generalization capabilities of the ASI model, coupled with the quantified performance enhancements of the blended system, provide advanced forecast capabilities. These advancements can contribute directly to optimizing plant performance, enhancing grid stability, minimizing operating costs, and ultimately supporting more reliable deployment of solar energy systems.

However, the presented study did not analyze the accuracy of the forecasts for the detection of short-term ramp events. Each participating model inherently possesses characteristics which are not ideal for ramp forecasting. In particular, persistence models, by their very nature, are incapable of anticipating rapid changes, thus missing ramp events by definition. Purely data-driven ASI models as used in this work are extensively optimized for RMSE, which typically results in smoother output curves, making sharp ramp detections challenging [44]. Similarly, the satellite-based forecast lacks the necessary temporal or spatial resolution required for accurate identification of such transient events. Furthermore, combining these models through approaches that optimize RMSE or MSE tends to further smooth the combined forecast curves, making short-term ramp event detection even more difficult. For the accurate and timely detection of very short-term ramp events, generative ASI models have been shown to be more appropriate [45], and such approaches and their combination with other forecasts will be studied in the future.

Yet, this combined forecast excels at providing accurate and stable prediction of the overall solar irradiance profile across all lead times, a strength demonstrated by its low RMSE, MAE, and correspondingly high skill scores. This makes it a valuable tool for applications requiring accurate total energy yield estimations over a given period.

Acknowledgments

This work was conducted at the DLR. The research was funded through internal support from DLR as part of the DLR In2Action Impuls Project (https://www.dlr.de/en/so/research-transfer/projects-1/current-projects-1/in2action). The authors would like to express their sincere gratitude to their DLR colleagues, Mr. Eduardo Saez Martinez and Dr Thomas Schmidt, for their valuable contributions toward processing sky images to derive cloud coverage and creating keograms. Both of these contributions have greatly advanced the analysis of atmospheric conditions in this study. [Correction added on 24 November 2025 after first online publication: "mean squared error" replaced with "mean absolute error." has been updated.]

Open Access funding enabled and organized by Projekt DEAL.

Funding

This work was supported by DLR.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Y. Chu, M. Li, C. F. Coimbra, D. Feng, and H. Wang, "Intra-Hour Irradiance Forecasting Techniques for Solar Power Integration: A Review," *iScience* 24, no. 10 (2021): 103136.

2. A. D. Mills and R. H. Wiser, "Strategies to Mitigate Declines in the Economic Value of Wind and Solar at High Penetration in California," *Applied Energy* 147 (2015): 269–278.

3. J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. J. Martinez-de-Pison, and F. Antonanzas-Torres, "Review of Photovoltaic Power Forecasting," *Solar Energy* 136 (2016): 78–111.

4. P. Blanc, J. Remund, and L. Vallance, "Short-Term Solar Power Forecasting Based on Satellite Images," in Renewable Energy Forecasting (Woodhead Publishing, 2017), 179–198.

5. Z. Qu, A. Oumbe, P. Blanc, et al., "Fast Radiative Transfer Parameterisation for Assessing the Surface Solar Irradiance: The Heliosat-4 Method," *Meteorologische Zeitschrift* 26 (2017): 33–57.

6. T. Carrière, Amaro, R. Silva, F. Zhuang, Y. M. Saint-Drenan, and P. Blanc, "A New Approach for Satellite-Based Probabilistic Solar Forecasting with Cloud Motion Vectors," *Energies* 14, no. 16 (2021): 4951.

7. Z. Si, M. Yang, Y. Yu, and T. Ding, "Photovoltaic Power Forecast Based on Satellite Images considering Effects of Solar Position," *Applied Energy* 302 (2021): 117514.

8. A. Habte, M. Sengupta, A. Andreas, S. Wilcox, and T. Stoffel, "Intercomparison of 51 radiometers for Determining Global Horizontal Irradiance and Direct Normal Irradiance Measurements," *Solar Energy* 133 (2016): 372–393.

9. H. T. Pedro, C. F. Coimbra, M. David, and P. Lauret, "Assessment of Machine Learning Techniques for Deterministic and Probabilistic Intra-Hour Solar Forecasts," *Renewable Energy* 123 (2018): 191–203.

10. E. Lorenz, B. Nouri, S. Cros, et al., Forecasting Solar Radiation and Photovoltaic Power. Best Practices Handbook for the Collection and Use of Solar Resource Data for Solar Energy Applications, 4th ed. (2024), https://iea-pvps.org/key-topics/best-practices-handbook-for-the-collection-and-use-of-solar-resource-data-for-solar-energy-applications-fourth-edition/.

11. B. Nouri, P. Kuhn, S. Wilbert, et al., "Cloud Height and Tracking Accuracy of Three All Sky Imager Systems for Individual Clouds," *Solar Energy* 177 (2019): 213–228.

12. Y. Fabel, B. Nouri, S. Wilbert, et al., "Applying Self-Supervised Learning for Semantic Cloud Segmentation of All-Sky Images," *Atmospheric Measurement Techniques* 15, no. 3 (2022): 797–809.

13. C. Feng, J. Zhang, W. Zhang, and B. M. Hodge, "Convolutional Neural Networks for Intra-Hour Solar Forecasting Based on Sky Image Sequences," *Applied Energy* 310 (2022): 118438.

14. Y. Nie, Q. Paletta, A. Scott, et al., "Sky Image-Based Solar Forecasting Using Deep Learning with Heterogeneous Multi-Location Data: Dataset Fusion versus Transfer Learning," *Applied Energy* 369 (2024): 123467.

15. N. B. Blum, S. Wilbert, B. Nouri, et al., "Analyzing Spatial Variations of Cloud Attenuation by a Network of All-Sky Imagers," *Remote Sensing* 14, no. 22 (2022): 5685.

16. Y. Fabel, B. Nouri, S. Wilbert, et al., "Combining Deep Learning and Physical Models: A Benchmark Study on All-Sky Imager-Based Solar Nowcasting Systems," *Solar RRL* 8, no. 4 (2024): 2300808.

17. G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, "*A Transformer-Based Framework for Multivariate Time Series Representation Learning*," in Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (2021), 2114–2124.

18. G. Bertasius, H. Wang, and L. Torresani, "*Is Space-Time Attention All You Need for Video Understanding?*," in ICML, Vol. 2 (2021), 4.

19. A. Hammer, J. Kühnert, K. Weinreich, and E. Lorenz, "Short-Term Forecasting of Surface Solar Irradiance Based on Meteosat-SEVIRI Data Using a Nighttime Cloud Index," *Remote Sensing* 7, no. 7 (2015): 9070–9090.

20. J. Lezaca, N. Blum, B. Nouri, A. Hammer, and O. Lünsdorf, *Deliverable 2.3: Methodologies for Short-Term Solar Resource Forecasting by Merging Various Inputs*, 2023, https://www.smart4res.eu/wp-content/uploads/2023/01/Smart4RES_Deliverable_D2.3.pdf.

21. M. López-Cuesta, R. Aler-Mur, I. M. Galván-León, F. J. Rodríguez-Benítez, and A. D. Pozo-Vázquez, "Improving Solar Radiation Nowcasts by Blending Data-Driven, Satellite-Images-Based and All-Sky-Imagers-Based Models Using Machine Learning Techniques," *Remote Sensing* 15, no. 9 (2023): 2328.

22. N. Straub, W. Herzberg, A. Dittmann, and E. Lorenz, "Blending of a Novel All Sky Imager Model with Persistence and a Satellite Based Model for High-Resolution Irradiance Nowcasting," *Solar Energy* 269 (2024): 112319.

23. N. Straub, W. Herzberg, E. Lorenz, and A. Dittmann, "*Combination of a Novel All Sky Imager Based Approach for High-resolution Solar Irradiance Nowcasting with Persistence and Satellite Nowcasts for Increased Accuracy*," in European Photovoltaic Solar Energy Conference and Exhibition (2023).

24. H. E. Beck, N. E. Zimmermann, T. R. McVicar, N. Vergopolan, A. Berg, and E. F. Wood, "Present and Future Köppen-Geiger Climate Classification Maps at 1-Km Resolution," *Scientific Data* 5, no. 1 (2018): 1–12.

25. T. Schmidt, J. Stührenberg, N. Blum, et al., "Eye2Sky–a Network of All-Sky Imager and Meteorological Measurement Stations for High Resolution Nowcasting of Solar Irradiance," *Meteorologische Zeitschrift* 34 (2025): 35.

26. T. Schmidt, J. Stührenberg, N. Blum, et al., "Eye2Sky Dataset - All-Sky Images and Meteorological Measurements (1.0) [Data Set], Zenodo," *Meteorologische Zeitschrift* 34 (2025): 35.

27. S. Wilbert, S. Kleindiek, B. Nouri, et al., "*Uncertainty of Rotating Shadowband Irradiometers and Si-Pyranometers including the Spectral Irradiance Error*," in AIP Conference Proceedings (AIP Publishing, Vol. 1734, 2016).

28. S. Wilbert, T. Stoffel, D. Myers, et al., Measuring Solar Radiation. Best Practices Handbook for the Collection and Use of Solar Resource Data for Solar Energy Applications, 4th ed. (2024).

29. A. Forstinger, S. Wilbert, A. Driesse, and B. Kraas, "Uncertainty Calculation Method for Photodiode Pyranometers," *Solar RRL* 6, no. 5 (2022): 2100468.

30. B. Nouri, S. Wilbert, P. Kuhn, et al., "Real-Time Uncertainty Specification of All Sky Imager Derived Irradiance Nowcasts," *Remote Sensing* 11, no. 9 (2019): 1059.

31. Eumetsat, *Conversion from Radiances to Reflectances for SEVIRI Warm Channels, Technical Report* (EUMETSAT, Eumetsat-Allee 1, D-64295 Darmstadt, Germany, 2012), https://user.eumetsat.int/s3/eup-strapi-media/pdf_msg_seviri_rad2refl_d6cd0c663f.pdf#:~:text=The%20present%20document%20describes%20the%20conversion%20from%20radiance,for%20SEVIRI%20on%E2%80%93board%20MSG1%2C%20MSG2%2C%20MSG3%20and%20MSG4 .

32. D. Dumortier, *Modelling Global and Diffuse Horizontal Irradiances under Cloudless Skies with Different Turbidities. Daylight II, jou2-ct92-0144, Final Report Vol. 2* (1995).

33. J. Remund and D. Domeisen, *Aerosol Optical Depth and Linke Turbidity Climatology. Description for Final Report of IEA SHC Task, 36* (2009).

34. A. Roy, D. Heinemann, and A. Hammer, *Effect of Dust Aerosol on Surface Solar Irradiance Retrieval from Satellite Observations* (2022).

35. P. Ineichen and R. Perez, "A New Airmass Independent Formulation for the Linke Turbidity Coefficient," *Solar Energy* 73, no. 3 (2002): 151–157.

36. W. F. Holmgren, C. W. Hansen, and M. A. Mikofski, "Pvlib Python: A Python Package for Modeling Solar Energy Systems," *Journal of Open Source Software* 3, no. 29 (2018): 884.

37. T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "*Optuna: A next-generation hyperparameter optimization framework*," in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2019), 2623–2631.

38. R. H. Byrd, M. E. Hribar, and J. Nocedal, "An Interior Point Algorithm for Large-Scale Nonlinear Programming," *SIAM Journal on Optimization* 9, no. 4 (1999): 877–900.

39. T. Chen and C. Guestrin, *Xgboost: A Scalable Tree Boosting System. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining* (2016): 785–794.

40. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation* 9, no. 8 (1997): 1735–1780.

41. Ashish Vaswani, Noam Shazeer, Niki Parmar, et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems* 30 (2017).

42. F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin* 1, no. 6 (1945): 80–83.

43. P. Kuhn, B. Nouri, S. Wilbert, et al., "Validation of an All-Sky Imager–based Nowcasting System for Industrial PV Plants," *Progress in Photovoltaics: Research and Applications* 26, no. 8 (2018): 608–621.

44. Q. Paletta, G. Arbod, and J. Lasenby, "Benchmarking of Deep Learning Irradiance Forecasting Models from Sky Images–An in-Depth Analysis," *Solar Energy* 224 (2021): 855–867.

45. B. Nouri, Y. Fabel, N. Blum, et al., "Ramp Rate Metric Suitable for Solar Forecasting," *Solar RRL* 8, no. 24 (2024): 2400468.

46. D. Magiera, Y. Fabel, B. Nouri, N. Blum, D. Schnaus, and L. F. Zarzalejo, "Advancing Semantic Cloud Segmentation in All-Sky Images: A Semi-Supervised Learning Approach with Ceilometer-Driven Weak Labels," *Solar Energy* 300 (2025): 113822.

APPENDIX A1

Meteorological Information on the Dataset used

The distribution of solar irradiance components, air mass, Linke turbidity, and cloud cover within the dataset is presented in Figure A1 and Figure A2. Linke turbidity and clear-sky irradiance were calculated following the methodology outlined in [34]. Cloud cover was derived from all-sky images using the procedure described in [46]. Additionally, Figure A3 and Figure A4–A6 display keograms generated from all-sky images at the OLDON site, illustrating the daily atmospheric conditions throughout the dataset in an high resolution.

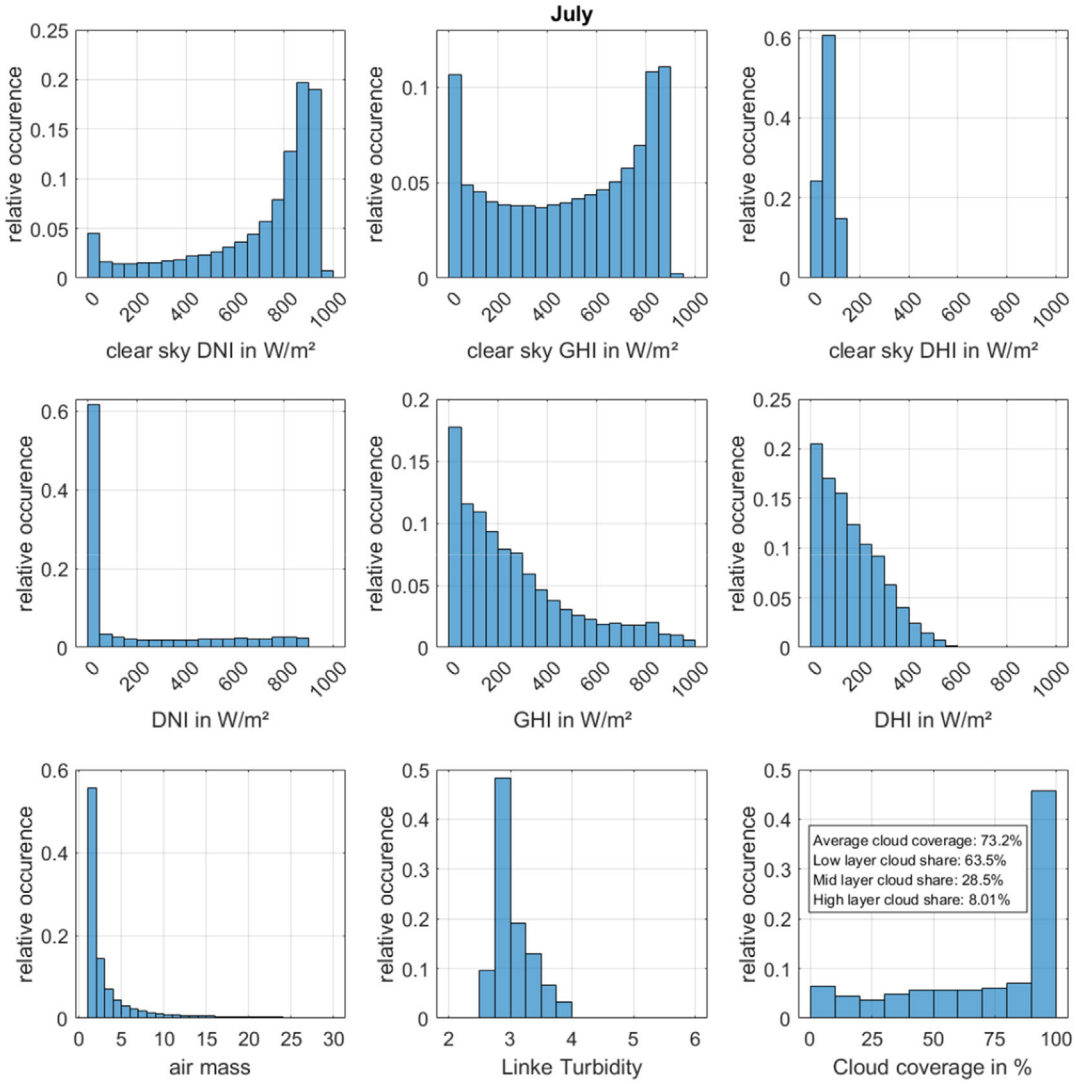

FIGURE A1 | Distribution of solar irradiance, air mass, Linke turbidity, and cloud coverage at the OLDON site in July 2020.

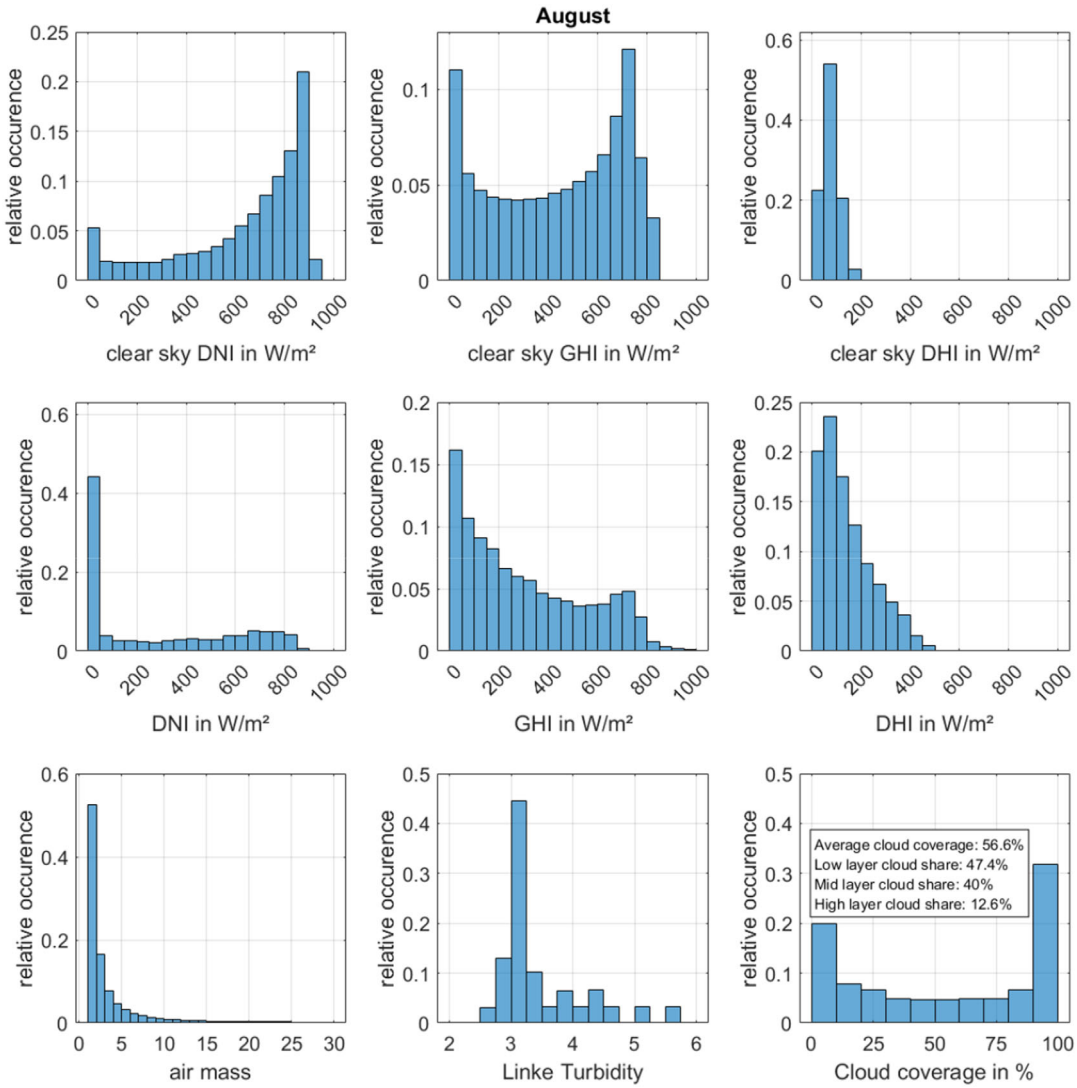

FIGURE A2 | Distribution of solar irradiance, air mass, Linke turbidity, and cloud coverage at the OLDON site in August 2020.

FIGURE A3 | Keograms based on all images in July 2020.

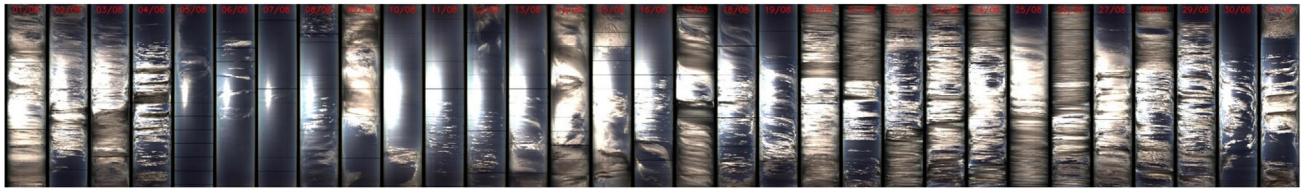

FIGURE A4 | Keograms based on all images in August 2020.

APPENDIX A2

Combined ASI+Persistence Forecasts with Linear Regression

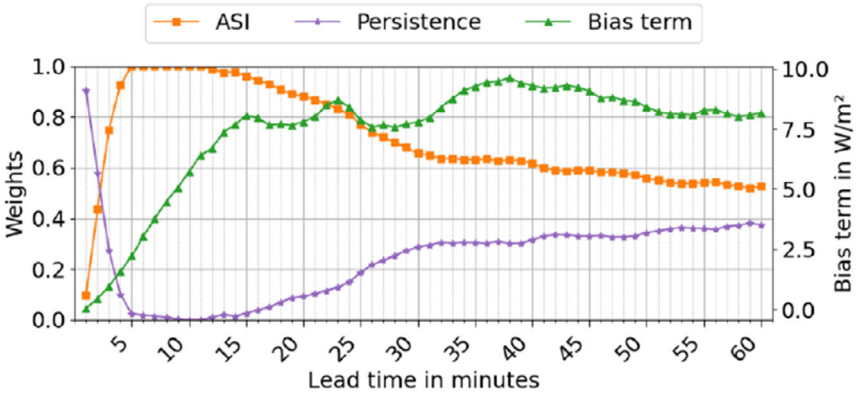

FIGURE A5 | Derived weights and bias term used for the combination ASI + Persistence discretized over lead times.

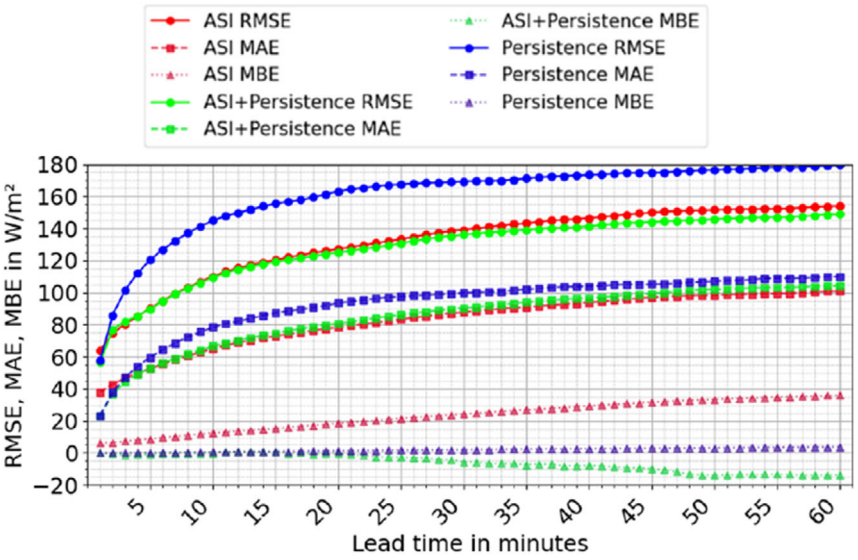

FIGURE A6 | Error metrics over lead time for the combination ASI+Persistence.

APPENDIX A3

Assessment of Forecast Performance at Operational Satellite Update Frequency

For this evaluation, only the ASI and persistence forecasts available at the time of issue of the satellite system are used. Additionally, only the lead times provided by the satellite system are considered (Figure A7).

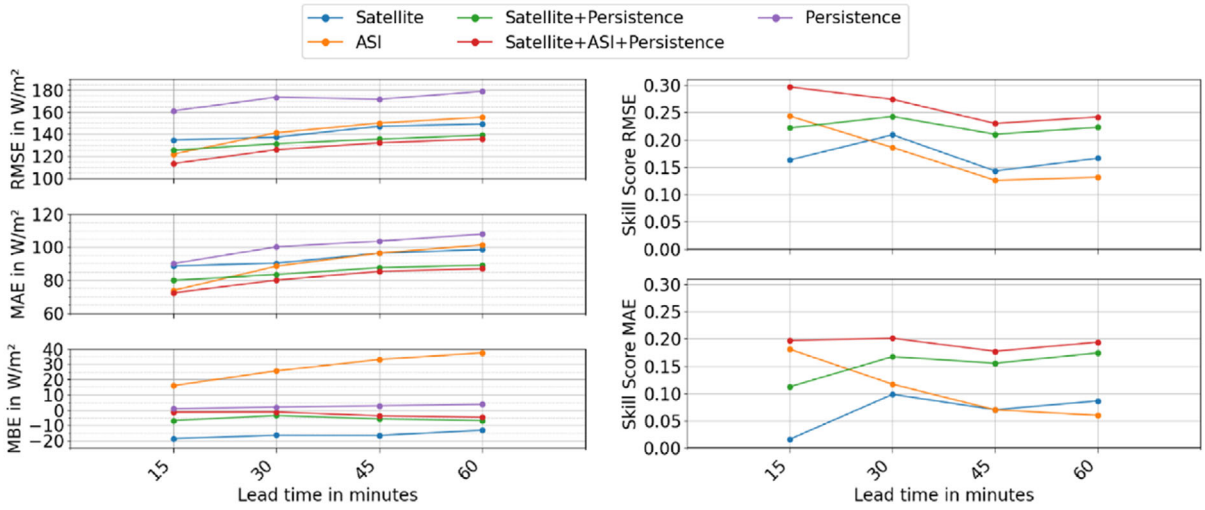

FIGURE A7 | Error metrics and skill scores over lead time for operational satellite update frequency.

APPENDIX A4

Averaged Metrics Concerning: Influence of Prevailing Sky Conditions

See Tables A1–A6.

TABLE A1 | Average ± standard deviation of error metrics over different LT ranges under high clear sky index conditions (bundled class 1).

Bundled class 1 conditions[a]						
Error metric	Used LT	Satellite	ASI	Satellite+ Persistence	Satellite+ASI+ Persistence	Persistence
RMSE [W/m²] ↓	LT1 to LT15	60.2±1.4	26.1±2.7	30.0±11.3	29.1±8.3	19.0±4.4
	LT16 to LT30	56.5±0.4	32.5±2.8	48.8±1.9	42.5±2.2	58.1±10.4
	LT31 to LT45	57.1±0.8	39.4±1.4	51.4±0.8	46.9±1.1	74.5±2.8
	LT46 to LT60	62.9±2.5	44.3±0.7	55.1±1.3	52.2±1.7	81.5±1.1
	LT1 to LT60	59.2±3.0	35.6±7.2.	46.3±11.3	42.7±9.6	58.3±25.1
MAE [W/m²] ↓	LT1 to LT15	36.0±0.3	15.1±0.8	17.4±8.2	17.7±5.4	7.9±2.4
	LT16 to LT30	35.6±0.1	18.4±1.4	30.4±0.9	26.4±1.3	20.9±3.9
	LT31 to LT45	35.7±0.4	22.3±0.9	30.6±0.5	28.1±0.5	29.3±2.1
	LT46 to LT60	38.5±0.9	24.9±0.6	32.7±0.4	30.3±0.6	35.4±1.3
	LT1 to LT60	36.5±1.3	20.2±3.9	27.8±7.3	25.6±5.5	23.4±10.7
MBE [W/m²] ↓	LT1 to LT15	-5.4±0.3	-10.9±0.8	2.5±1.4	-9.5±3.8	0.4±0.5
	LT16 to LT30	-3.8±0.4	-10.8±0.7	2.6±0.6	-4.9±0.9	-9.6±3.5
	LT31 to LT45	-3.6±0.6	-12.9±0.6	-0.5±1.0	-5.7±1.3	-15.2±1.1
	LT46 to LT60	-7.1±1.2	-15.4±0.8	-6.9±2.8	-13.1±2.5	-19.7±1.1
	LT1 to LT60	-5.0±.1.6	-12.5±2.0	-0.6±4.2	-8.3±4.0	-11.0±7.8

[a]The color coding evaluates each row separately and ranges from dark green (best outcome) to dark red (worst outcome).

TABLE A2 | Wilcoxon signed-rank test results based on paired RMSE values under high clear sky index conditions (bundled class 1).

Bundled class 1 conditions[a]										
	Persistence		Satellite		ASI		Satellite+ Persistence		Satellite+ASI+ Persistence	
Persistence	—	—	0.757	0.9 W/m^2	<0.001	−22.7 W/m^2	<0.001	−11.9 W/m^2	<0.001	−15.6 W/m^2
Satellite	—	—	—	—	<0.001	−23.6 W/m^2	<0.001	−12.9 W/m^2	<0.001	−16.5 W/m^2
ASI	—	—	—	—	—	—	<0.001	10.8 W/m^2	<0.001	7.1 W/m^2
Satellite+Persistence	—	—	—	—	—	—	—	—	<0.001	−3.7 W/m^2

[a]The first value in each cell represents the *p*-value, while the second represents the mean RMSE difference. The system indicated in the rows is always subtracted from the system indicated in the columns of the matrix to calculate the differences.

TABLE A3 | Average ± standard deviation of error metrics over different LT ranges under intermediate clear sky index conditions (bundled class 2).

Bundled class 2 conditions[a]						
Error metric	Used LT	Satellite	ASI	Satellite+ Persistence	Satellite+ASI+ Persistence	Persistence
RMSE [W/m²] ↓	LT1 to LT15	188.4±2.9	152.1±23.0	168.7±23.6	149.9±20.6	201.6±39.7
	LT16 to LT30	193.7±1.3	178.3±3.0	184.8±1.5	175.4±3.3	231.0±1.6
	LT31 to LT45	201.3±1.9	183.7±0.5	189.3±1.7	183.2±1.8	227.0±0.9
	LT46 to LT60	203.4±0.2	182.2±1.5	192.1±0.8	186.7±0.8	224.0±0.6
	LT1 to LT60	196.7±6.3	174.0±17.0	183.7±14.8	173.8±17.8	220.9±22.5
MAE [W/m²] ↓	LT1 to LT15	138.7±2.0	106.9±18.0	119.0±23.8	106.1±20.3	137.7±34.1
	LT16 to LT30	142.4±1.6	129.0±2.3	135.3±1.0	129.5±2.7	164.3±1.3
	LT31 to LT45	149.5±1.7	132.8±0.5	139.6±1.9	136.1±1.9	161.0±0.8
	LT46 to LT60	151.6±0.1	131.8±1.3	142.3±0.6	139.5±0.6	158.8±0.5
	LT1 to LT60	145.5±5.5	125.1±14.0	134.1±14.8	127.8±16.5	155.5±19.7
MBE [W/m²] ↓	LT1 to LT15	-43.5±1.0	16.4±7.9	-5.3±10.8	3.7±3.6	13.3±7.1
	LT16 to LT30	-43.3±0.6	15.8±2.3	-27.0±1.5	-15.9±4.7	-10.6±3.5
	LT31 to LT45	-46.3±0.6	10.8±0.8	-32.3±1.1	-26.6±2.4	-19.3±1.5
	LT46 to LT60	-43.9±1.0	8.1±1.0	-34.0±0.9	-32.3±1.5	-20.9±0.8
	LT1 to LT60	-44.3±1.4	12.8±5.4	-24.6±12.7	-17.8±14.2	-9.4±14.3

[a]The color coding evaluates each row separately and ranges from dark green (best outcome) to dark red (worst outcome).

TABLE A4 | Wilcoxon signed-rank test results based on paired RMSE values under intermediate clear sky index conditions (bundled class 2).

Bundled class 2 conditions[a]										
	Persistence		Satellite		ASI		Satellite+Persistence		Satellite+ASI+ Persistence	
Persistence	—	—	<0.001	−24.2 W/m^2	<0.001	−46.9 W/m^2	<0.001	−37.2 W/m^2	<0.001	−47.1 W/m^2
Satellite	—	—	—	—	<0.001	−22.7 W/m^2	<0.001	−13 W/m^2	<0.001	−22.9 W/m^2
ASI	—	—	—	—	—	—	<0.001	9.7 W/m^2	0.324	−0.3 W/m^2
Satellite+Persistence	—	—	—	—	—	—	—	—	<0.001	−9.9 W/m^2

[a]The first value in each cell represents the *p*-value, while the second represents the mean RMSE difference. The system indicated in the rows is always subtracted from the system indicated in the columns of the matrix to calculate the differences.

TABLE A5 | Average ± standard deviation of error metrics over different LT ranges under low clear sky index conditions (bundled class 3).

Bundled class 3 conditions[a]						
Error metric	Used LT	Satellite	ASI	Satellite+ Persistence	Satellite+ASI+ Persistence	Persistence
RMSE [W/m²] ↓	LT1 to LT15	124.0±1.5	86.3±20.5	93.8±21.5	81.2±19.4	106.0±29.7
	LT16 to LT30	129.6±2.4	133.7±9.2	121.0±3.0	114.1±4.5	159.7±7.3
	LT31 to LT45	135.8±0.8	158.0±5.6	127.8±0.9	124.4±1.4	175.6±3.4
	LT46 to LT60	140.5±2.2	171.4±2.6	133.0±1.9	129.8±2.0	185.9±2.6
	LT1 to LT60	132.5±6.5	137.4±34.6	118.9±18.6	112.4±21.4	156.8±34.5
MAE [W/m²] ↓	LT1 to LT15	87.1±0.8	56.5±13.1	58.6±18.8	50.3±15.5	59.5±20.4
	LT16 to LT30	91.2±1.8	90.4±7.4	83.8±2.2	78.1±3.7	98.1±5.7
	LT31 to LT45	96.5±0.9	110.3±4.8	89.1±1.0	86.2±1.3	111.6±2.6
	LT46 to LT60	99.9±1.3	122.2±2.4	92.8±1.2	89.9±1.3	119.4±2.2
	LT1 to LT60	93.7±5.1	94.9±26.3	81.1±16.3	76.1±17.5	97.2±25.5
MBE [W/m²] ↓	LT1 to LT15	-13.3±0.8	21.2±2.7	-5.7±3.4	3.6±2.3	-7.4±4.9
	LT16 to LT30	-10.6±0.7	41.9±7.5	4.1±1.7	9.4±1.4	15.8±4.9
	LT31 to LT45	-7.0±.1.6	63.4±5.3	9.9±1.4	14.8±1.1	26.2±1.9
	LT46 to LT60	-0.0±2.6	78.4±3.6	14.0±.1.5	18.5±1.4	31.1±1.5
	LT1 to LT60	-7.7±5.3	51.2±22.4	5.6±7.8	11.6±5.9	16.4±15.4

[a]The color coding evaluates each row separately and ranges from dark green (best outcome) to dark red (worst outcome).

TABLE A6 | Wilcoxon signed-rank test results based on paired RMSE values under low clear sky index conditions (bundled class 3).

Bundled class 3 conditions[a]										
	Persistence		Satellite		ASI		Satellite+ Persistence		Satellite+ASI+ Persistence	
Persistence	—	—	<0.001	−24.4 W/m^2	<0.001	−19.5 W/m^2	<0.001	−37.9 W/m^2	<0.001	−44.5 W/m^2
Satellite	—	—	—	—	0.069	4.9 W/m^2	<0.001	−13.6 W/m^2	<0.001	−20.1 W/m^2
ASI	—	—	—	—	—	—	<0.001	−18.4 W/m^2	<0.001	−25 W/m^2
Satellite+Persistence	—	—	—	—	—	—	—	—	<0.001	−6.55 W/m^2

[a]The first value in each cell represents the *p*-value, while the second represents the mean RMSE difference. The system indicated in the rows is always subtracted from the system indicated in the columns of the matrix to calculate the differences.

APPENDIX A5

Averaged Metrics Concerning: Temporal Aggregation

See Tables A7–A10 and Figure A8.

TABLE A7 | Average ± standard deviation of error metrics over different LT ranges, considering a 5 min time window for temporal aggregation.

5 min window for temporal aggregation[a]						
Error metric	Used LT	Satellite	ASI	Satellite+ Persistence	Satellite+ASI+ Persistence	Persistence
RMSE [W/m²] ↓	LT5 to LT15	121.1±1.1	89.4±12.5	101.9±9.0	87.3±11.3	119.3±14.6
	LT16 to LT30	125.6±1.5	116.8±6.3	116.4±2.2	108.8±4.0	147.0±4.7
	LT31 to LT45	131.5±1.2	132.6±3.3	121.9±1.3	117.8±1.6	155.8±2.0
	LT46 to LT60	135.5±1.5	140.2±1.3	126.2±1.5	122.6±1.6	161.5±1.6
	LT5 to LT60	129.1±5.3	121.9±19.6	117.7±9.6	110.7±13.8	147.8±16.7
MAE [W/m²] ↓	LT5 to LT15	81.3±0.6	55.0±7.6	64.5±8.2	54.7±7.7	66.8±10
	LT16 to LT30	83.9±1.2	73.8±4.7	76.7±1.6	71.3±3.0	87.1±3.8
	LT31 to LT45	88.5±1.1	85.7±2.6	80.6±1.2	77.7±1.3	95.4±1.8
	LT46 to LT60	92.0±0.9	92.1±1.1	84.2±0.8	81.4±1.0	100.4±1.4
	LT5 to LT60	86.8±4.2	78.2±14.1	77.4±7.9	72.5±10.3	88.9±13.0
MBE [W/m²] ↓	LT5 to LT15	-19.2±0.4	12.3±2.2	-4.1±0.8	0.2±0.4	0.7±0.2
	LT16 to LT30	-17.8±0.3	20.3±2.7	-4.9±0.4	-1.5±0.6	1.6±0.3
	LT31 to LT45	-17.0±0.5	28.5±2.2	-4.5±0.2	-2.1±0.6	2.6±0.3
	LT46 to LT60	-14.0±1.1	34.4±1.3	-4.7±0.4	-3.8±0.5	3.5±.0.3
	LT5 to LT60	-16.8±2.0	24.7±8.3	-4.6±0.5	-1.9±1.5	2.2±1.1

[a]The color coding evaluates each row separately and ranges from dark green (best outcome) to dark red (worst outcome).

TABLE A8 | Average ± standard deviation of error metrics over different LT ranges, considering a 10 min time window for temporal aggregation.

10 min window for temporal aggregation[a]						
Error metric	Used LT	Satellite	ASI	Satellite+ Persistence	Satellite+ASI+ Persistence	Persistence
RMSE [W/m²] ↓	LT10 to LT15	113.0±0.4	83.2±5.8	95.6±3.8	81.7±5.1	110.8±6.8
	LT16 to LT30	116.1±1.5	104.9±7.1	106.0±2.4	97.9±4.5	131.8±5.4
	LT31 to LT45	121.8±1.3	122.6±3.7	111.7±1.3	107.6±1.7	142.4±2.2
	LT46 to LT60	126.2±1.5	131.1±1.6	116.5±1.6	113.0±1.7	148.9±1.9
	LT10 to LT60	120.4±5.0	115.3±16.4	109.5±6.9	103.3±10.4	137.5±12.5
MAE [W/m²] ↓	LT10 to LT15	76.1±0.3	52.3±3.6	63.3±2.9	52.9±3.8	64.5±4.4
	LT16 to LT30	78.6±1.2	67.1±5.1	71.0±1.7	65.2±3.3	79.2±4.2
	LT31 to LT45	83.1±1.0	79.8±2.7	74.8±1.1	71.8±1.3	88.2±1.9
	LT46 to LT60	86.6±1.1	86.3±1.2	78.7±1.0	75.9±1.1	93.5±1.3
	LT10 to LT60	82.0±3.9	74.8±11.7	73.5±5.0	68.9±7.6	84.3±9.6
MBE [W/m²] ↓	LT10 to LT15	-19.0±0.2	14.0±1.1	-4.7±0.4	0.2±0.4	0.9±0.1
	LT16 to LT30	-17.9±0.3	20.4±2.7	-5.0±0.4	-1.5±0.6	1.7±0.3
	LT31 to LT45	-17.0±0.5	28.6±2.2	-4.5±0.2	-2.1±0.6	2.8±0.3
	LT46 to LT60	-14.0±1.2	34.6±1.3	-4.7±0.4	-3.7±0.5	3.7±0.3
	LT10 to LT60	-16.6±1.9	26.2±7.4	-4.7±0.4	-2.1±1.3	2.5±1.0

[a]The color coding evaluates each row separately and ranges from dark green (best outcome) to dark red (worst outcome).

TABLE A9 | Average ± standard deviation of error metrics over different LT ranges, considering a 15 min time window for temporal aggregation.

15 min window for temporal aggregation[a]						
Error metric	Used LT	Satellite	ASI	Satellite+ Persistence	Satellite+ASI+ Persistence	Persistence
RMSE [W/m²] ↓	LT16 to LT30	108.8±1.4	95.1±8.0	97.7±2.7	89.2±5.1	119.1±6.6
	LT31 to LT45	114.1±1.3	114.6±4.0	103.7±1.4	99.7±1.9	132.0±2.4
	LT46 to LT60	118.9±1.5	124.1±1.9	109.0±1.7	105.6±1.8	139.3±2.1
	LT15 to LT60	111.5±5.8	97.0±28.7	92.7±23.1	86.5±23.6	113.4±33.6
MAE [W/m²] ↓	LT16 to LT30	74.4±1.2	61.5±5.5	66.3±1.9	60.1±3.6	72.3±4.8
	LT31 to LT45	78.9±1.0	75.0±2.8	70.3±1.1	67.2±1.4	82.5±2.2
	LT46 to LT60	82.3±1.1	81.8±1.4	74.4±1.0	71.6±1.2	88.1±1.3
	LT15 to LT60	76.6±4.5	63.5±18.7	62.7±16.2	58.3±16.2	70.0±21.9
MBE [W/m²] ↓	LT16 to LT30	-18.0±0.3	20.5±2.7	-5.0±0.4	-1.5±0.6	1.9±0.4
	LT31 to LT45	-17.0±0.5	28.8±2.2	-4.5±0.2	-2.0±0.5	3.0±0.3
	LT46 to LT60	-14.1±1.2	34.8±1.3	-4.6±0.4	-3.7±0.4	4.0±0.3
	LT15 to LT60	-17.2±2.1	23.8±9.3	-4.3±1.0	-1.8±1.4	2.4±1.3

[a]The color coding evaluates each row separately and ranges from dark green (best outcome) to dark red (worst outcome).

TABLE A10 | Average ± standard deviation of error metrics over different LT ranges, considering a 30 min time window for temporal aggregation.

30 min window for temporal aggregation[a]						
Error metric	Used LT	Satellite	ASI	Satellite+ Persistence	Satellite+ASI+ Persistence	Persistence
RMSE [W/m²] ↓	LT31 to LT45	98.2±1.3	96.4±5.1	87.0±1.9	82.7±2.5	109.4±4.0
	LT46 to LT60	103.4±1.4	108.9±2.5	93.4±1.6	90.1±1.9	120.1±2.7
	LT30 to LT60	96.5±5.1	79.9±27.3	75.9±22.1	70.5±21.8	88.1±33.9
MAE [W/m²] ↓	LT31 to LT45	69.1±1.1	64.6±3.4	60.6±1.5	57.2±1.8	69.7±2.8
	LT46 to LT60	72.8±1.0	73.0±1.7	65.1±0.9	62.2±1.1	76.9±1.7
	LT30 to LT60	67.4±4.2	53.7±17.9	52.8±15.5	48.6±15.0	55.8±22.0
MBE [W/m²] ↓	LT31 to LT45	-17.0±0.5	29.5±2.3	-4.2±0.2	-1.7±0.5	3.7±0.4
	LT46 to LT60	-13.9±1.2	35.7±1.4	-4.3±0.4	-3.3±0.4	4.8±0.3
	LT30 to LT60	-17.1±2.2	24.3±9.6	-4.1±1.0	-1.5±1.3	2.9±1.6

[a]The color coding evaluates each row separately and ranges from dark green (best outcome) to dark red (worst outcome).

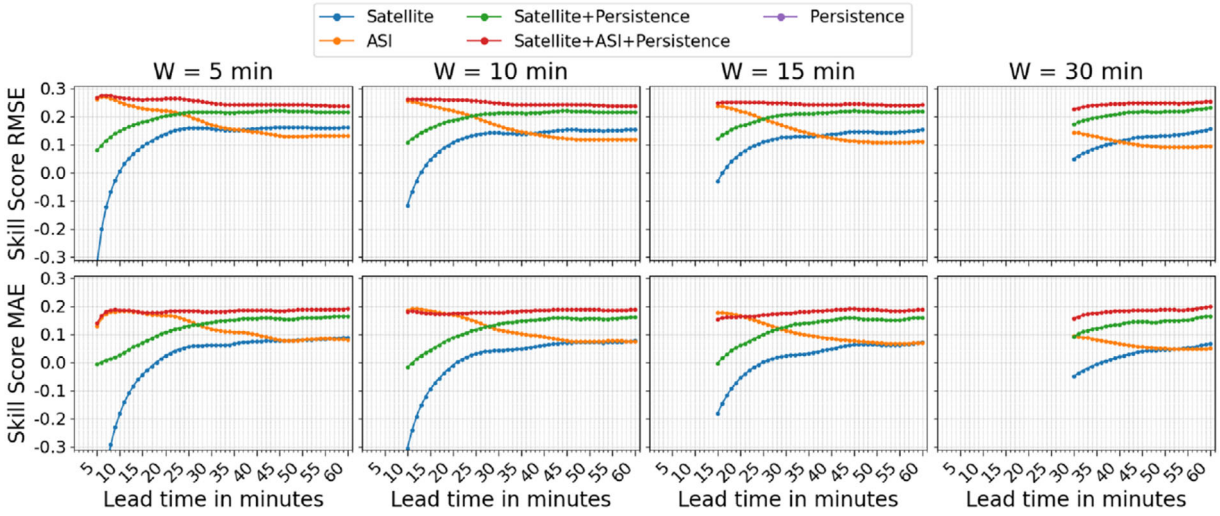

FIGURE A8 | Skill scores over lead time, considering different time windows for temporal aggregation. In terms of skill scores, some of the lower LTs for the ASI and satellite forecasts have been cropped. This is done to improve the readability of the skill scores.