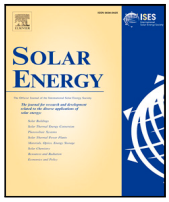
Solar Energy 272 (2024) 112463

Contents lists available at ScienceDirect

Solar Energy

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



All sky imaging-based short-term solar irradiance forecasting with Long Short-Term Memory networks

N.Y. Hendrikx [a,1], K. Barhmi [b,*], L.R. Visser [b], T.A. de Bruin [b], M. Pó [c], A.A. Salah [a], W.G.J.H.M. van Sark [b]

[a] Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands
[b] Copernicus Institute of Sustainable Development, Utrecht University, Utrecht, The Netherlands
[c] EKO Instruments Europe B.V., The Hague, The Netherlands

ARTICLE INFO

Dataset link: https://github.com/nielsyh/ASI_playground

Keywords:
Time-series
Solar forecasting
Machine learning
Global horizontal irradiance
All sky imaging
Deep neural networks
LSTM

ABSTRACT

The intermittent nature of solar irradiance, primarily due to cloud movements, leads to rapid short-term fluctuations in the power output of photovoltaic (PV) systems. These fluctuations pose a significant challenge for integrating this renewable energy source into the power grid. Accurate forecasting of solar irradiance is not only crucial but also multi-beneficial. It enables more precise grid management by allowing operators to anticipate power output fluctuations and adjust energy distribution and storage strategies accordingly. This proactive approach reduces the reliance on backup power sources, which are often less sustainable and more expensive. Furthermore, accurate forecasts enhance the overall efficiency and reliability of energy systems by minimizing the impact of power variability on the grid, thereby supporting a more stable and sustainable energy supply.

Addressing this need, our study focuses on the development of a forecasting model through innovative feature engineering, systematic design of specific attributes, and optimization of sequence length. The model is tailored to perform efficiently across various weather conditions and offers predictions for a time horizon of 0 to 20 min ahead. Utilizing a Long Short-Term Memory (LSTM) model, we achieve a remarkable ramp Forecast Skill Score of 39% in sunny and 25% in partially cloudy conditions. This work not only contributes to the existing literature but also presents a pioneering methodology for solar energy integration, highlighting the importance and application of accurate short-term solar irradiance forecasting.

1. Introduction

The integration of increasing amounts of photovoltaic (PV) systems presents technical challenges to grid operation. The variable nature of solar irradiance means that PV power output can experience significant fluctuations within a short time. These rapid changes, commonly referred to as "ramp events", necessitate that grid operators maintain sufficient regulating and reserve capacity to uphold grid stability. Operating extensive balancing reserves is not only costly but also carbon-intensive. Additionally, rapid fluctuations in PV power can provoke local voltage swings, adversely affecting power quality and service reliability.

Accurate solar forecasting is crucial in addressing these challenges. By predicting solar irradiance, and, consequently, PV power output, grid operators can more effectively prepare for and respond to energy production fluctuations. Such foresight diminishes the reliance on expensive and carbon-intensive balancing reserves, thereby enhancing grid stability and improving power quality [1].

A promising tool in solar forecasting is the employment of sky cameras, also known as All-Sky Imagers (ASI). These devices, which capture images of the sky, provide valuable data on cloud movements and formations. Clouds significantly influence the variability of sunlight reaching the earth's surface, making this information essential for predicting short-term changes in solar irradiance. The integration of ASI data into forecasting models has demonstrated improved accuracy in short-term solar forecasts [2].

In this context, our study aligns with the growing interest in leveraging deep learning techniques for solar forecasting. Specifically, we draw upon the innovative approaches of studies such as in [3], which utilizes LSTM networks for photovoltaic power prediction using sky images and historical power values. Our study extends this by integrating a novel

* Corresponding author.
E-mail address: k.barhmi@uu.nl (K. Barhmi).
[1] ISES member.

https://doi.org/10.1016/j.solener.2024.112463
Received 23 September 2023; Received in revised form 1 March 2024; Accepted 12 March 2024
Available online 1 April 2024

combination of all-sky imaging and advanced Long Short-Term Memory (LSTM) convolutional neural networks, showcasing their efficacy in various weather conditions .

This study proposes using an LSTM model for forecasting Global Horizontal Irradiance (GHI) up to 20 min in advance, utilizing data derived from ASI input. LSTM models have garnered attention in various fields for their forecasting capabilities and are now being explored in the context of solar irradiance prediction. Our approach, which pioneers the application of LSTM models in solar forecasting, builds on findings from two prior studies.

We compare the proposed LSTM model against persistence and smart persistence models and other neural network and Random Forest (RF) models. The study also investigates the impact of incorporating additional variables and selecting optimal lagged values on forecast accuracy.

Our comparative analysis includes the SKIPP'D model, a deep-learning method for solar forecasting, by applying it to our test data and fine-tuning it on our dataset for comparative evaluation. This approach is vital for establishing the model's effectiveness and its place in the deep learning landscape for solar forecasting [3].

Our findings suggest that the LSTM model outperforms other classifier models and persistence methods in providing superior short-term solar irradiance forecasts under various weather conditions. The remainder of this paper is organized as follows: Section 2 presents related work on solar forecasting. Section 3 describes our methodology, benchmarks, and the experimental data. Section 4 discusses the experimental results, while Section 5 delves into our analysis of these findings. The paper concludes with Section 6.

2. Background and related work

As the integration of photovoltaic (PV) capacity into the power grid continues to grow, the challenges and costs associated with grid management are escalating. For instance, this trend prompted the Puerto Rico Electric Power Authority to enact legislation, allowing for a maximum ramp event of 10% for grid-connected utility-scale PV power plants [4]. Accurate PV power forecasts have become essential tools for grid operators, aiding in the integration of substantial PV capacity and enabling timely measures to address power fluctuations. This, in turn, leads to more dependable and cost-effective grid operations. Given that balancing reserves typically require up to 15 min to adjust [5,6], the prediction of ramp events at this timescale is crucial.

A promising approach in solar forecasting revolves around All-Sky Imager (ASI)-based models [7–9]. These models have proven effective in generating accurate short-term forecasts with high spatial and temporal resolutions, often predicting the Global Horizontal Irradiance (GHI) locally up to 20 or 30 min in advance. Beyond this 30-minute horizon, the accuracy of ASI-based forecasts rapidly declines due to cloud dynamics [10]. Moreover, the precision achieved by ASI-based forecast models at the proposed resolution surpasses that of alternative solar forecasting methods, including Numerical Weather Prediction (NWP) and satellite-based models [7,11].

An ASI captures images of the sky at regular intervals, creating sequences that illustrate cloud movement and dynamics. Various methods are employed to extract information from these images or sequences, enabling the prediction of GHI at specific time horizons. Subsequently, PV power production can be estimated from these GHI predictions using a GHI-to-PV power conversion model. Several approaches with varying degrees of success have been developed thus far [7].

Recent advancements in the field, such as the study by [12], have demonstrated the capability of convolutional neural networks (CNN) to predict PV output from sky images directly. This study underscores the potential of using contemporaneous sky images for "now-cast" predictions, achieving significant accuracy. The success of this approach in correlating PV output with sky images suggests the feasibility of direct PV power predictions, expanding the scope of data-driven methods in solar forecasting.

A common first step in these models involves the use of algorithms to extract features from ASI images, i.e., the process of transforming raw image pixel values to other meaningful information. Examples of features are the amount of cloud pixels obtained by means of a cloud pixel detection method or their brightness level. Once extracted, the features may be used as an input for statistical models like regression or support vector machines (SVM), which are trained to predict the future GHI on the extracted features [13,14].

Recent studies in solar forecasting show significant variation in prediction horizons. For example, [13] used a four-week dataset, achieving the best results with a 5-minute prediction horizon using images taken 5 min prior. However, their study presents data duration and image resolution limitations without comparing them to persistence-based forecasts. Similarly, [14] employed SVM with a 4-hour feature timespan aiming for a 1-hour prediction horizon, categorizing data into day types. Additionally, [15] investigated optimal time frames for feature extraction with the CNN model "SUNSET", focusing on 15-minute predictions, offering valuable insights for improved forecast accuracy. These studies highlight the importance of identifying the right timeframe for effective feature extraction in solar forecasting.

ASI-based GHI forecasting models often use Cloud Motion Vectors (CMVs), derived from cloud movement in images, for predicting cloud positions and estimating GHI. Techniques like cross-correlation and optical flow, especially variational optical flow at the pixel level, aid in CMV estimation [6,16–20]. However, CMVs can be less effective under unstable cloud conditions, leading to lower accuracy with increased lead time [21]. Integrating local GHI and temperature data, alongside cloud movement tracking, enhances prediction accuracy [22].

[23] developed a method for predicting solar irradiance one hour in advance using data from sensors across five locations, measuring variables like GHI, DNI, and atmospheric conditions over a two-year period. The study tested three models: ARIMA [24], a well-known time series forecasting method; MLP, an Artificial Neural Network with a single hidden layer; and XGBoost [25], a gradient-boosted decision tree model. While XGBoost was the most effective, integrating multiple-site data was challenging. The researchers proposed investigating Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for future work, given their potential in sequential data analysis [23].

Another data-driven forecast has been reported by [26], employing a 30-minute forecast horizon and a network of 65 GHI measurement sites, along with a comprehensive suite of ANN models. One of their notable findings was the degradation of forecasting performance with increasing forecasting horizons.

In a noteworthy contribution to solar forecasting using LSTM [27], this research explores various deep learning architectures, including MLP, CNN, and LSTM networks, for short-term photovoltaic power output prediction using sky images and historical power data. The study, conducted in Kyoto, Japan, highlights that LSTM models, known for their proficiency in processing temporal sequences, outperform MLP and CNN models in forecasting accuracy. Specifically, the LSTM model achieves an impressive Root Mean Square Error (RMSE) skill score of 21%. This underscores the potential of LSTM networks in accurately capturing temporal dependencies essential for precise solar power forecasting, aligning with the approaches emphasized in our research.

In the same context a recent contribution in this domain is a model developed by [3], which has significantly advanced the field by providing a comprehensive dataset for short-term solar forecasting using deep learning models and sky images. This benchmark dataset offers a unique opportunity to rigorously test and compare the performance of different deep learning models, including our own. In our research, we aim to leverage this dataset to validate the effectiveness and accuracy of our proposed method, ensuring a thorough comparison with current state-of-the-art models.

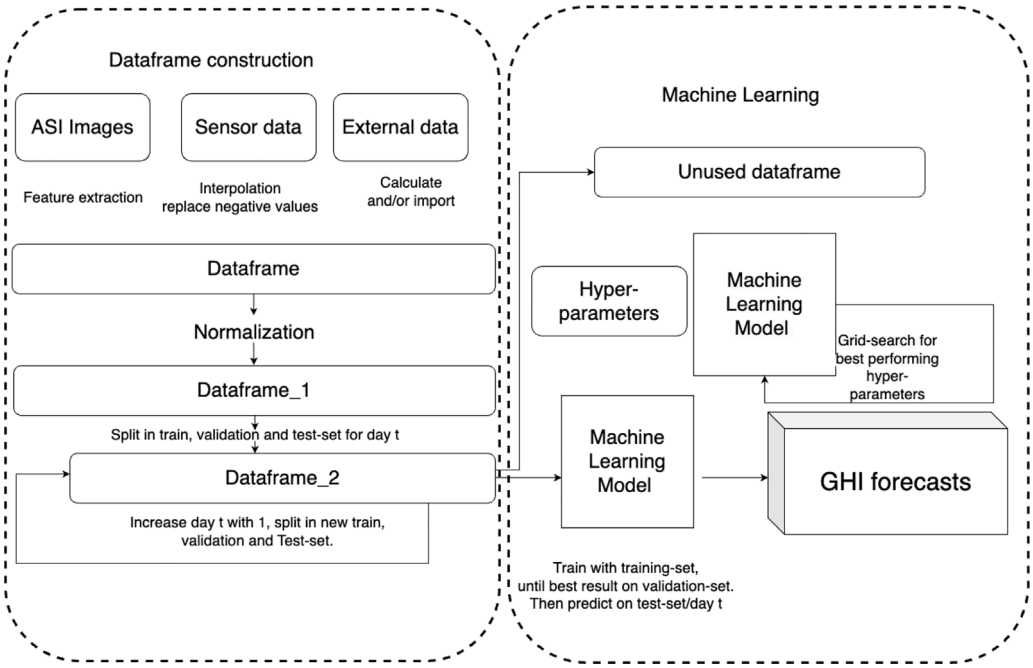


Fig. 1. Flowchart of the proposed approach for solar irradiance prediction. The 'Machine learning model' in this chart refers to the proposed LSTM method but also represents other approaches tested for comparison.

Recent studies use large models like Pangu-Weather for enhanced weather forecasting. Pangu-Weather, with its 3D deep neural networks and Earth-specific priors, excels in analyzing complex weather patterns and minimizing errors, outperforming systems like ECMWF [28]. This trend towards combining diverse data with sophisticated models for accuracy mirrors our use of advanced LSTM convolutional neural networks in solar irradiance forecasting. LSTM's proficiency in processing sequential solar data highlights the uniqueness of our methodology.

In our research, we advance the field of solar irradiance forecasting by leveraging cutting-edge deep learning methods specifically tailored to address the challenges posed by the intermittent nature of solar energy due to cloud cover. Recognizing the pivotal role of Convolutional Neural Networks (CNNs) in image processing, our study harnesses their potential for accurate cloud detection and movement prediction, critical for short-term GHI forecasting [29]. We also explore the possibilities offered by Generative Adversarial Networks (GANs) in image synthesis and enhancement, which is particularly beneficial for simulating diverse sky conditions and thus enriching the dataset for our models [30]. A core aspect of our study is the innovative use of Long Short-Term Memory (LSTM) networks, a subset of Recurrent Neural Networks (RNNs), known for their exceptional ability to model temporal sequences. By integrating LSTM with all-sky imager images and local meteorological data, our research formulates a novel approach that method enables us to create future representations of the sky's evolution, which, when processed through our LSTM neural network, results in highly accurate short-term solar irradiance predictions. The LSTM model, enhanced through feature engineering and sequence length optimization, demonstrates remarkable proficiency in forecasting GHI across various weather conditions and a prediction horizon of 0 to 20 min ahead, achieving a ramp Forecast Skill Score of 39% in sunny and 25% in partially cloudy scenarios.

3. Proposed forecasting approach

This section describes the method we propose for solar forecasting and the data used to validate it. A flowchart summarizing our approach is shown in Fig. 1, and the methods are explained in the following.

The flowchart 1 illustrates our proposed methodology for forecasting Global Horizontal Irradiance (GHI) through the use of machine learning techniques. While the Long Short-Term Memory (LSTM) model is our primary focus, given its proficiency in handling sequential data, we have also employed Random Forest (RF) and Artificial Neural Network (ANN) models for benchmarking purposes. The raw data comprising ASI images, sensor data, and external data sources undergo preprocessing, which includes feature extraction, interpolation to address missing values, and normalization of variables to ensure uniformity in scale. The preprocessed data is then split into training, validation, and test sets for day t, referred to as `Dataframe_2`, with subsequent iterations incrementing the day by 1 to facilitate continuous evaluation. The machine learning models are trained to discern patterns within historical data, and upon achieving the optimal performance on the validation set, they are employed to generate GHI forecasts, which are crucial for optimizing the performance of solar power generation systems.

3.1. Data collection

Two CMS-Schreder ASI-16/50 cameras from [8] are installed at Plataforma Solar de Almería (PSA) in southern Spain, at coordinates 37.091549 °N, −2.363556 °E and 37.095253 °N, −2.354785 °E, as shown in Fig. 2. The optimal siting of these cameras, approximately 880.2 meters apart, was determined based on achieving comprehensive sky coverage while minimizing the overlap between their fields of view, which is crucial for the stereoscopic cloud analysis when used.

These sites were carefully selected to provide a clear, unobstructed field of view for each camera, ensuring the highest quality of data for solar irradiance and cloud movement analysis. The robust construction of the camera hardware, including features such as ventilation to prevent condensation and a double-cover design, provides resilience in harsh environmental conditions, as stated in the ASI manual [8].

3.1.1. Data input

The data used in this study can be divided into three categories: (1) image data from the cameras, (2) measurements by installed sensors, and (3) external data collected from open sources (see Table 1). The cameras have been operational since July 23, 2019. We studied three months of data from August to November 2019 under different weather

Fig. 2. Location of the CMS-Schreder ASI-16/50 cameras near Almería, Spain.

Table 1
Sets of features used in the study.

Features/Subsets	All data	Image	Onsite	Meteo
Time and date	✓	✓	✓	✓
GHI	✓		✓	
Temperature	✓		✓	
Humidity	✓		✓	
Clear sky GHI	✓			✓
Clear sky index (CSI)	✓			✓
Azimuth	✓			✓
Zenith	✓			✓
Sun–earth distance	✓			✓
Number of cloud pixel	✓	✓		
Brightness	✓	✓		
Number of edges	✓	✓		
Number of corners	✓	✓		

conditions. The images cover a 180° field of view, with a sampling rate of 15 s. Additional sensor data, including ambient temperature (°C), GHI (W/m^2), and relative humidity (%), are acquired at the same time and location. Details for the various parameters implemented in this study are further indicated with a checkmark in Table 1.

As a data set, We consider data from the 1st of August 2019 to 31 December 2019. These comprise 121 sunny days, 29 partially cloudy days, and three cloudy days (a total of 153 days). We take a random sample of five sunny days, three partially cloudy days, and two cloudy days to construct an independent test set; these are not seen during model selection and training.

The classes sunny, partially cloudy, and cloudy are distinguished based on a daily averaged clear sky index (CSI):

- Sunny: CSI > 0.75
- Partially cloudy: 0.75 > CSI > 0.25
- Cloudy: CSI < 0.25

where CSI is defined as the ratio of measured irradiance and estimated irradiance in clear sky conditions [18].

The predictions are made for days between 25 September 2019 and 21 November 2019. This includes 4 partially cloudy days and 24 sunny days. Due to weather circumstances in the south of Spain, fully cloudy days are not present in this period. Fig. 3 shows monthly averages of GHI and its variance over the months of August until December 2019. We observe that all GHI graphs closely follow a bell-like curve. Higher variance is seen for GHI in September and October (Fig. 3), as a result of some moving clouds. Consequently, GHI for these months might be harder to predict. Temperature and humidity are less affected by moving clouds, as can be seen in Figs. A.1 and A.2 in Appendix.

The selected timeframe for prediction, specifically the months of September to November, typically experiences a high frequency of cloud cover, particularly moving clouds. This introduces significant variability in solar irradiance, posing a substantial challenge for the forecasting models. The increased cloud movement during these months is evidenced by the higher variance in the GHI values, as depicted in Fig. 3, making it a pertinent period for evaluating the robustness and accuracy of our forecasting models under less-than-optimal conditions for solar generation.

To assess the effectiveness of various parameters, we conducted an extensive comparison over a nine-day period. We used the preliminary dataset as a 'validation set' to identify the models with the best performance. The validation set comprises specific days in 2019 with different weather conditions, including sunny days (September 15th, October 15th, November 15th, and December 15th), partially cloudy days (October 21st, November 17th, and December 16th), and cloudy days (October 22nd and December 3rd).

3.1.2. All sky images

From the camera images see Fig. 4, several features are extracted. The intensity I is defined as the mean grayscale of an image (with $0 < p_r, p_g, p_b < 255$):

$$I = \frac{\sum_{p \in image} (p_r + p_g + p_b)}{N} \tag{1}$$

where:

$$\begin{aligned} N &= \text{the amount of pixels} \\ p_r &= \text{intensity of red pixel (R)} \\ p_g &= \text{intensity of green pixel (G)} \\ p_b &= \text{intensity of blue pixel (B)} \end{aligned}$$

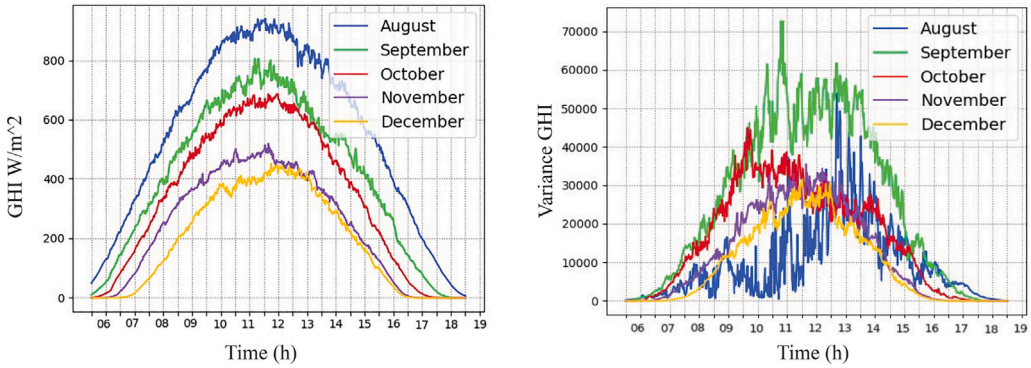


Fig. 3. Monthly average (left) and variance (right) GHI (in kWh/m^2) at camera site 1.

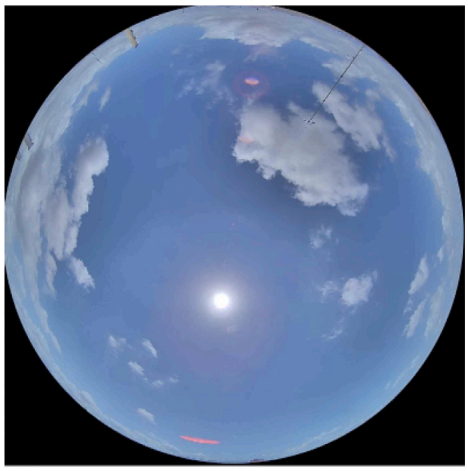


Fig. 4. Initial ASI image capturing cloud coverage at a specific time point.

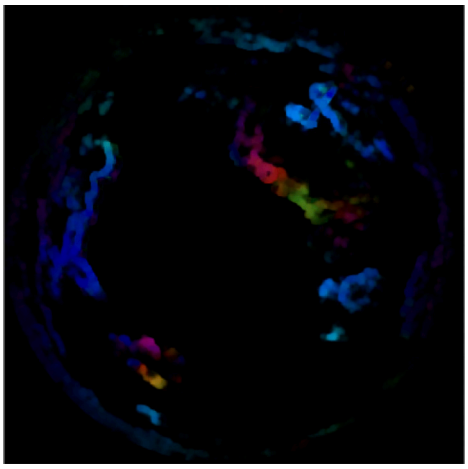


Fig. 5. Visualization of cloud movements with Farneback dense optical flow.

Clouds are extremely important when predicting GHI [9,10]. Image-based forecasting requires the extraction of information from relevant images, particularly in determining whether a pixel is part of a cloud. There are multiple ways to ascertain if a pixel belongs to a cloud. While Chauvin et al. [11] identified three categories for existing cloud detection algorithms—threshold methods, neural networks, and dedicated algorithms—a more recent and extensive benchmarking by Hasenbalg et al. [13] evaluated six cloud segmentation algorithms, including a neural network-based approach and several threshold and hybrid methods. In this study, we focus on threshold methods, as they offer a balance between computational efficiency and accuracy for our specific application, where an artificial neural network-based approach is not feasible since the dataset lacks ground truth on clouded and non-clouded pixels necessary to train a neural network. Additionally, dedicated algorithms would be too computationally expensive and require dedicated hardware to detect haze, thin clouds, and opaque clouds, which is not justified given the marginal improvement in segmentation accuracy for our use case as indicated by the findings of Hasenbalg et al. [13].

Threshold algorithms are employed to classify pixels as cloud or clear-sky based on pre-defined thresholds. Among various methods for pixel classification using RGB pixel values [13,14], the following algorithms stand out: (1) red–blue ratio (RBR), (2) blue–red–blue–green ratio (BRBG), and (3) normalized red–blue ratio (NRBR). The RBR ($= R/B$) method uses a fixed threshold T_f, typically ranging from 0.6 to 0.8 to differentiate cloud and non-cloud pixels ($R/B > T_f$). Studies have indicated that BRBG ($= B/R + B/G$) performs well for pixel classification [13]. In the case of NRBR ($= (R - B)/(R + B)$), an adaptive thresholding approach has been proposed using minimum cross-entropy (MCE) [16]. The MCE method determines the threshold by minimizing cross-entropy between the original and segmented images. The segmentation is achieved by calculating the mean and standard deviation of the normalized B/R ratio values.

In our paper, the RBR algorithm is used to identify cloud pixels, employing a threshold value of $T_f = 0.8$. To detect edges, we apply the Canny edge detection algorithm [17], and for corner detection, we utilize the Harris algorithm [31].

3.1.3. Optical flow

The integration of optical flow techniques is instrumental in the estimation of cloud velocity, a critical factor in solar forecasting. We first tested the Lukas–Kanade method, as delineated in [9], to analyze cloud dynamics using high-resolution All-Sky Imager (ASI) images. These images, with a resolution of 3456 × 3456 pixels, were captured at 30-second intervals for a comprehensive period of a month. By subdividing the images into 400 × 400 pixel blocks, we were able to track cloud movements in proximity to the sun.

Using optical flow, specifically the Farneback method [32], we predicted cloud block locations based on the sun's position. The Farneback method calculates the flow per pixel by employing a two-frame flow estimation algorithm, which uses quadratic polynomials to approximate the motion between two frames, as depicted in Fig. 5. This approach, utilizing the polynomial expansion transform, allowed us to quantify cloud coverage as variable C. Subsequently, we improved Global Horizontal Irradiance (GHI) predictions by subtracting the cloud coverage fraction, weighted by a constant, from the clear sky GHI baseline forecast.

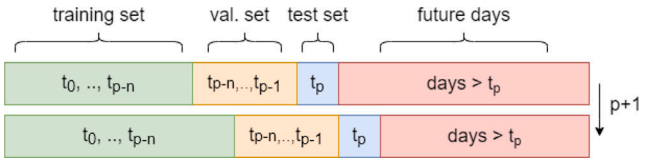


Fig. 6. Example training and test set. Here p is the day to predict (test) and n is the number of validation days. t_0 is the first data in the data set.

Our methodology augmented the data on time intervals utilized for calculating cloud velocity, addressing the gaps in previous studies [9]. Additionally, the robustness of our approach was validated across an extensive spectrum of weather conditions and scenarios.

Moreover, Forecast Skill Scores, which are discussed in Section 4, provide a benchmark for our model's performance. When comparing these scores to other weather prediction methods, it is imperative to account for the variability of meteorological conditions, which can significantly influence forecast accuracy.

3.1.4. Data: Availability, train, validation and test data

To predict for the day t_p we can only use data observed before t_p to train our models. We consider all data before predicting time $t_p - 3$ as training data. See Fig. 6 (note, we have taken $n = 3$ as the number of validation days, as test data). Furthermore, the validation set will exist out of $t_p - 3, \ldots, t_p - 1$, thus not including t_p itself. Data for the day t_p will be used to test the predictions. Therefore, when time passes, the training set will contain a large amount of data assuming a certain starting date t_{start}. Let 11 September 2019 be a day to predict. All data prior to 11 September is taken as training data, excluding 3 days before 11 September (8, 9, and 10 September). These three days are included as a validation set.

3.1.5. Single (S) and multi models (M)

The current use of ANN and LSTM for single (S) models results in unpredictable performance. Single models, denoted by 'S', are designed to forecast a specific prediction horizon, leading to a total of 20 unique models for 20 different prediction horizons. The models often predict 0 and require 20 times more computation time than multi-models (M), which are capable of forecasting across all prediction horizons simultaneously, significantly enhancing computational efficiency. Only multi-models are capable of making accurate predictions on the test-set. One specific example of unstable predictions is shown in Fig. 10 (Section 4) for the 'LSTM M 5 all data' model, and other single models were even less reliable. While the RF 'single' models do not appear to have this issue, the computational problem persists. As a result, due to their inefficiency and less reliable performance, all 'single' models will be excluded from predictions on the test-set.

3.2. Overview of methods

In this study, we use several machine learning (ML) approaches, including a Random Forest, which is a collection of decision trees [19], Artificial Neural Networks (ANN) [33], and Long Short-Term Memory (LSTM) [23]. We trained these models with stochastic gradient descent (SGD) and Adam optimizer [24]. SGD uses back-propagation until some defined minimum error is found. Adam is an extension of SGD, where it makes use of past values to change the momentum adaptively [24].

Each of the used ML approaches have a number of hyperparameters to optimize. These were selected based on the training set, and the test set is not seen during hyperparameter selection. We report the full range of parameters tested during learning here for transparency.

For RF, we tested different numbers of decision trees (50, 100, 150, 200, 250), minimum samples per leaf (1, 2, 4, 12, 24, 64), and the maximum depth of a tree (25, 50, 75, 100, 200). For ANN, we have chosen a three-layer architecture, testing different numbers of nodes per layer ((32, 64, 32), (64, 128, 64), (128, 256, 128)), different learning rates (0.001, 0.01, 0.1), dropout (0, 0.1, 0.5), and sigmoid vs. rectified linear unit activation functions.

We use a naming convention with the machine learning models presented in this paper as M-T-I-D, where M (model) can be LSTM, ANN, or RF; T (model type) is S or M, as discussed in Section 3.1.5; I (input length) can be selected from {5, 10, 20, 40, 60}, and D (data) can either be 'all data' or 'on site' or 'On site + all data' (Table 1).

3.3. Benchmarking

A common baseline model in solar irradiance forecasting is the persistence model [34]. 'Regular' persistence predicts a $value_t$ for time $t + h$, where t is time and h is the prediction horizon. An improvement on this is called smart persistence [26], and assumes a stable Clear Sky Index (CSI). The current CSI is calculated by the ratio of the current direct normal irradiance (DNI) and GHI. Subsequently, considering the solar zenith angle and time, the future DNI is calculated. The prediction, then, is the multiplication of the future DNI with the current CSI.

In addition to these baseline models, our study also incorporates a comparison with the SKIPP'D model developed by [3]. This model represents a significant advancement in solar irradiance forecasting, utilizing a sophisticated approach based on sky images and PV power generation data. To comprehensively evaluate the effectiveness of our proposed methodology, we perform a two-fold comparison with the SKIPP'D model:

1. **Direct Application**: In this setting, we apply the trained SKIPP'D model directly to our test data. This allows us to assess the model's out-of-the-box performance and its generalizability to different datasets.
2. **Fine-Tuning**: Here, we fine-tune the SKIPP'D model using our training data before applying it to the test data. This approach helps us understand the adaptability of the SKIPP'D model to new datasets and conditions, as well as the improvements in accuracy that can be achieved through fine-tuning.

We observed (see Fig. 7) that a long time frame as input made performance worse, while selecting the observed variables from a short period before the prediction moment performed better. The optimal number of minutes depends on the model, but we note that, in general, five to ten minutes of observations is most informative.

Since we had access to two other identical setups, we examined the contribution of adding training data from other sites during the training stage, but this did not improve our models.

We predicted CSI and calculated the GHI accordingly. Eventually, this did not improve our models. Thus, we suggest that predicting GHI directly is a better approach.

3.4. Error metrics

We employ common metrics for the evaluation of forecasting methods and compare these with a baseline. Most often, the following evaluation techniques are used in similar studies: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Forecast Skill Score (FS) [9,25,34], of which the latter is increasingly being used. However, as typically forecasting models are built to optimize on RMSE, we consider this the most important performance metric. The metrics are defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(e_i^2\right)} \tag{2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|e_i\right| \tag{3}$$

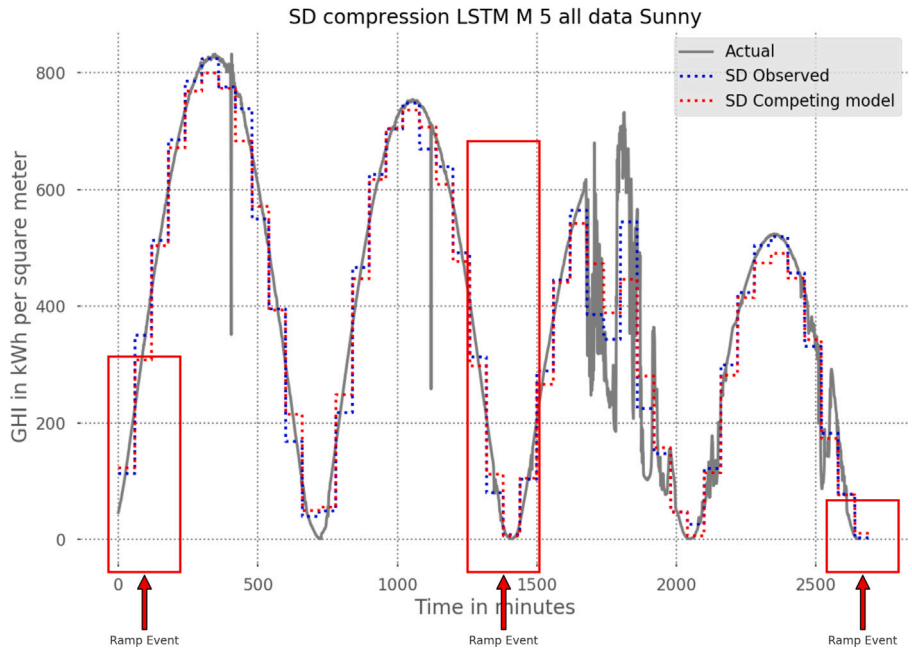


Fig. 7. Swinging door compression for the 'LSTM M 5 all data' model (for model definition, see Section 3.1) under sunny weather conditions. Markers indicate ramp events where significant GHI fluctuations were observed. SD output is averaged by the hour in this figure for clarity. This is based on a more granular 5-minute averaging interval. On this continuous timeline, four non-consecutive days are displayed, with the third day showing considerable variability.

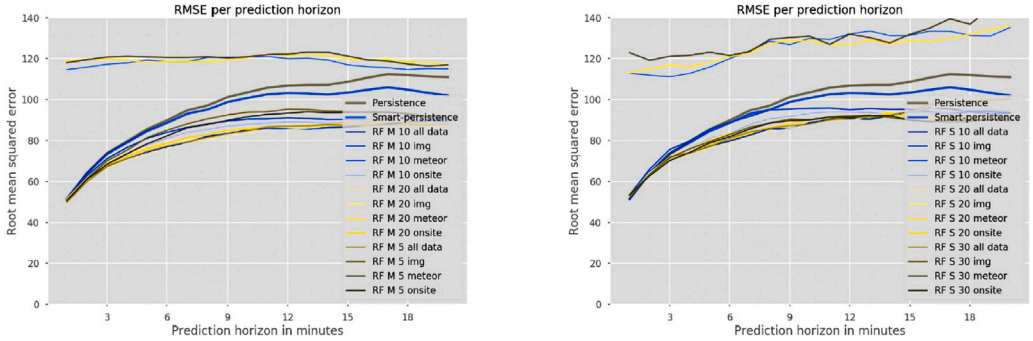


Fig. 8. RMSE per prediction horizon for preliminary data-set-RF. Left graph: single model, right graph: multi-model.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|o_i - \bar{p}_i|}{o_i} \times 100\% \tag{4}$$

$$FS = 1 - \frac{Error_{forecast}}{Error_{baseline}} \tag{5}$$

where:

e_i = The absolute difference $o_i - p_i$,
o_i = observed output,
p_i = predicted output,
N = the amount of samples considered,
Error = One of the error metrics (RMSE, MAE, MAPE) was used.

As an additional error metric, the ramp score (RS) is utilized to measure the forecasting ability for significant Global Horizontal Irradiance (GHI) fluctuations, known as ramp events [35]. Traditional error metrics focus on instantaneous accuracy, whereas RS captures fluctuations over a specific time period, better addressing short-term fluctuation challenges as described in the introduction. The unit of RS is consistent with GHI measurements, represented in kWh/m^2.

To compute RS, the prediction and observation time-series are compressed using the Swinging Door (SD) algorithm [36]. The sensitivity parameter ϵ determines the significance of detected ramps. A ramp is flagged when the GHI deviation from a linear approximation exceeds ϵ, which is scaled by the daily maximum clear-sky GHI value. In this study, ϵ is set to 0.05.

Adapting the RS to our 20-minute prediction horizon, we average over 5-minute intervals instead of the hourly basis used in [35], scaling down by a factor of 12. The RS is calculated using the equation:

$$RS = \frac{1}{t_{max} - t_{min}} \int_{t_{max}}^{t_{min}} |SD(T(t)) - SD(R(t))| dt \tag{6}$$

where:

SD = Output swinging door compression,
t_{max} = maximum bound of the period with $GHI > 0$,
t_{min} = minimum bound of the period with $GHI > 0$,
T(t) = test time series (the forecast),
R(t) = reference time series (measurements).

Fig. 7 has been updated to clearly depict actual ramp events, with visual markers added to highlight significant deviations between the observed and forecasted GHI values, as determined by the SD algorithm's sensitivity threshold. The SD output in this figure is averaged by the hour to provide a clear example, whereas our experimental analyses utilize a 5-minute averaging interval.

Table 2
The best performing hyperparameters for different applied models (RF, ANN, LSTM).

Model	Hyperparameters	Value	Best results
RF	Number of estimators	[50, 100, 150, **200**, 300]	**200**
	Minimum samples leaf	[**1**, 2, 4, 12, 24, 64]	1
	Max depth	[**25**, 50, 75, 100, 200]	25
ANN	Nodes	[(32, 64, 32), (64, 128, 64), (128, 265, 128)]	(64, 128, 64)
	Activation functions	['Relu', 'sigmoid']	Relu
	Learning rates	[**0.001**, 0.01, 0.1]	**0.001**
	Dropout	[**0**, 0.1, 0.5]	0
LSTM	Nodes	[(50, 25, 10), (60, 30, 15), (80, 40, 20)]	(50, 25, 10)
	Activation functions	['Relu', 'Sigmoid']	Relu
	Learning rates	[**0.001**, 0.01, 0.1]	**0.001**
	Dropout	[**0**, 0.1, 0.5]	**0**

3.5. Statistical significance

To compare if some model a is significantly better than another model b, the *Diabold–Mariano test* [37,38] is proposed. In this significance test, the prediction errors of both models are studied over a particular prediction horizon. We want to test if the null hypothesis:

$$H_0 : E(d_t) = 0, \forall_t \tag{7}$$

in comparison with the alternative hypothesis:

$$H_1 : E(d_t) \neq 0, \forall_t \tag{8}$$

Where E represents equal predictive accuracy.

We define the loss differential between two forecasts as $d_t = e_{1,t} - e_{2,t}$, representing the error at time t for model 1 and model 2. In our analysis, the error for some models is quantified using the Root Mean Square Error (RMSE). A common significance level of $\alpha = 0.05$ is adopted in forecasting, as supported by literature [37,38]. Given that prediction uncertainty increases with the horizon length, model errors are evaluated against a specific forecast horizon, denoted as H. For this study, we examine forecast horizons ranging from 1 to 20 min, covering a comprehensive spectrum of short-term predictions.

This approach allows for a nuanced assessment of model performance across different temporal scales, ensuring a robust analysis of forecast accuracy.

3.6. Hardware

All experiments (unless stated otherwise) are run on a node that contains an AMD EPYC 7451 24-Core Processor and 256 GB (RAM) memory. Additionally, the node is equipped with a GTX 1080 Ti GPU.

3.7. Model selection

3.7.1. Hyper-parameters_Model (RF, ANN, LSTM)

For **Random Forest**, we chose the mean square error (see Section 3.4) as a loss function for the random forest, because we are dealing with a regression problem and we want to penalize extreme values. Our model always considers all features. For the grid, a search is chosen for a 10-fold cross-validation on the training set (all days until 15 September). The grid search is done for the number of estimators, the minimum samples per leaf, and the maximum depth of a tree. The search comprised parameters as presented in Table 2. The best results came with Estimators of 200, a minimum samples leaf of 1, and a max depth of 25.

In the context of **ANN**, a 3-layer architecture outperformed others. Hyperparameter optimization involved a grid search for node count, activation functions, optimizers, learning rate, and dropout rate. The optimal number of epochs minimized validation loss, yielding nodes (50, 25, 10), ReLU activation, learning rate = 0.001, and dropout = 0.

In the case of Long Short-Term Memory (LSTM), a 3-layer architecture is utilized, comprising two LSTM layers followed by a dense layer. Hyperparameter optimization involves searching for the ideal node counts, activation functions, optimizers, and learning rate. The optimal configuration, as displayed in Table 2, includes nodes (64, 128, 64), ReLU activation, a learning rate of 0.001, and a dropout rate of 0.

The dataset includes October 5th–8th and October 20th. Four models (RF, ANN, LSTM, and Optical Flow) were applied, with RMSE displayed in the plots; a detailed summary is in Appendix. Satisfactory models will predict the test set, detailed in the following subsections, where the selected model is presented.

3.7.2. Model selection-random forest

Fig. 8 and Table A.1 demonstrate that when the input is the same, 'multi' outperforms 'single'. The most accurate predictions are made with a sequence length of 5, but the difference in accuracy between different sequence lengths is not significant enough to draw any firm conclusions.

The feature subsets that perform the best are 'all data' and 'onsite', while external data ('meteor') does not perform as well as the baseline. Therefore, we will select the feature subsets 'all data' and 'onsite' as they perform the best. In addition, we will consider sequence lengths of 5, 10, 20, 30, and 60.

3.7.3. Model selection-artificial neural network

Single models can be unstable, occasionally predicting 0's, leading to high error levels, which is not an issue in multi-models.

According to Table A.2, and Fig. 9 models with shorter sequence lengths typically perform better on RMSE, while longer sequences perform better on MAE. We have chosen feature lengths of 10, 20, 40, and 60, along with the feature subsets "all data" and "onsite", as our selection criteria.

3.7.4. Model selection-long short term memory

LSTM shares the issue of ANN in predicting 0's at times using a single model, but it is more reliable than ANN. However, LSTM appears to be more affected by sequence length, and experiments indicate that anything over 10 does not work well.

Using a sequence length of 3 is too brief and leads to slightly worse prediction results compared to lengths 5 and 10 (as illustrated in Fig. 10 and Table A.1). As a result, we choose sequence lengths 5 and 10, as well as the feature subsets "all data" and "onsite", for our selection.

3.7.5. Perez-model

In this article, the performance of the Perez ('prz') model in predicting Clear Sky Index (CSI) was examined. GHI values were derived from the ratio (CSI = DNI/GHI) and then compared to Global Horizontal Irradiance (GHI) measurements. The Perez model, known for its sophisticated approach in estimating solar radiation under clear sky conditions, incorporates factors such as the sun's position and atmospheric conditions into its calculations. Despite its comprehensive nature, the results showed that the 'prz' model performed much worse than other baseline models at all prediction horizons, with an RMSE of 255.28 at prediction horizon 20. However, it is worth noting that this

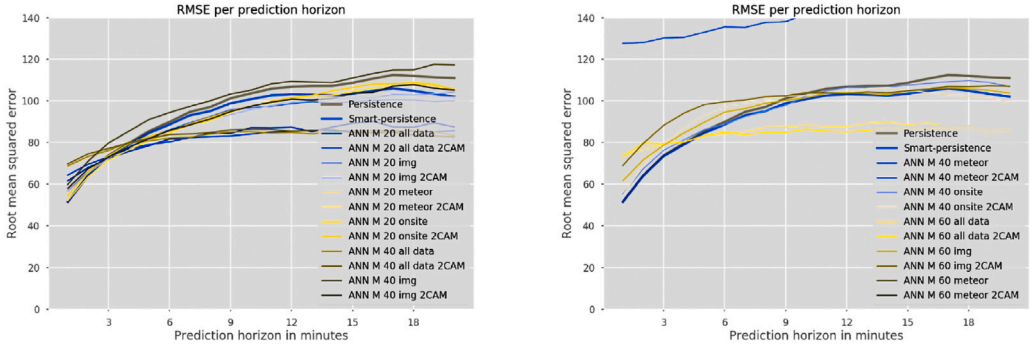


Fig. 9. RMSE per prediction horizon for preliminary data-set-ANN.

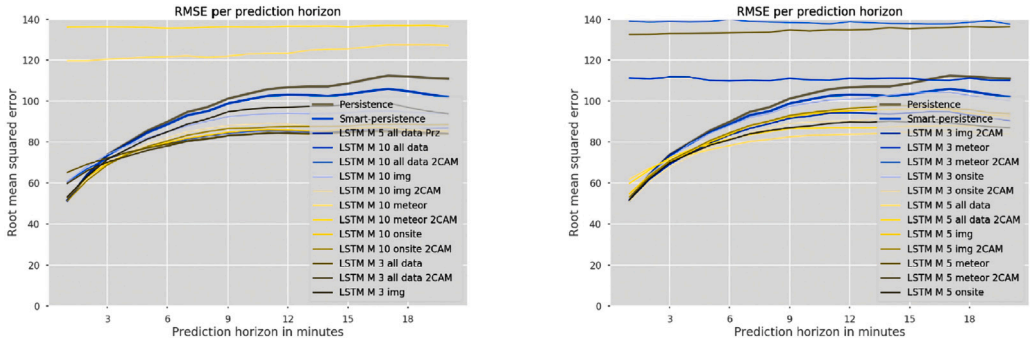


Fig. 10. RMSE per prediction horizon for preliminary data-set-LSTM.

poor performance is not apparent in Fig. 10, as the error for 'prz' is above 120.

This lower performance of the Perez model in our study could be attributed to its specialized focus on clear sky conditions, posing challenges when adapting to varying CSI conditions. These results suggest that predicting CSI might be more challenging than predicting GHI, potentially due to errors in the CSI-to-GHI conversion model. These findings raise questions about the suitability of the 'prz' model for predicting CSI, and whether a different approach may be necessary.

Model selection reveals a common difficulty in predicting cloudy weather, primarily due to limited data: only three cloudy days in the training set and two in the test set. This data scarcity hinders the models' ability to predict such conditions, as evident from their initial attempts where cloudy weather was barely or not encountered. The persistence model, in particular, struggles in these scenarios. A thorough analysis of each model's performance on the test set is detailed in Section 4.

4. Results

This section assesses the test set results using the selected model. We begin with an overview of each model's performance, including their strengths, weaknesses, and comparisons to the baseline. Subsequently, a detailed performance analysis is presented for all models, covering: model selection, valuable features, different weather circumstances, times of the day, computational time, additional training data, and statistical significance. We also tested models on the limited cloudy days in the data set.

4.1. Random forests

Results from our study presented in Table 3 demonstrate that RF is the fastest training model among all the proposed models. However, one detail that should be noted is that the model is trained to minimize Mean Squared Error (MSE) instead of Mean Absolute Error (MAE).

Table 3
Training execution times (in seconds) per model on day 27 September 2019. Training-set contains all available data before 27 September 2019 (excluding 3 days of validation set). The prediction times are 1 prediction for the next 1 .. 20 min in milliseconds. For the neural network the number of parameters per model.

	RF	ANN	LSTM
Training (s)	69.20	72.59	722.65
Prediction(ms)	0.014	0.0038	15.6
Parameters	n/a	71 452	19 280

In the study, RF was found to perform well on sunny days, with 'RF M 5 onsite' beating the baseline on average. However, on average, RF predicts slightly worse than smart-persistence. Despite this, when predicting weather conditions for a longer horizon of 20 min, RF's prediction performance was found to be slightly better than smart-persistence, as shown in Table A.1. For partially cloudy weather, 'RF M 60 all data' was found to perform best on average, outperforming persistence and smart-persistence, as indicated in Table 4. However, on cloudy days, RF performed poorly compared to persistence. Fig. 11 shows that the RMSE was plotted per prediction horizon, providing a detailed analysis of the performance of different models.

The study noted that RF performance varies with prediction sequence length. Short sequences excel on sunny days, while longer ones perform better on partially cloudy days. However, shorter sequences outperform longer ones beyond a 13-minute prediction horizon for partially cloudy weather.

4.2. ANN

The results presented in Table 5 and in Fig. 12 indicate that ANN models that take a sequence with length 10 perform the best in predicting weather conditions. This suggests that the optimal sequence length for weather prediction using ANN models may vary based on specific

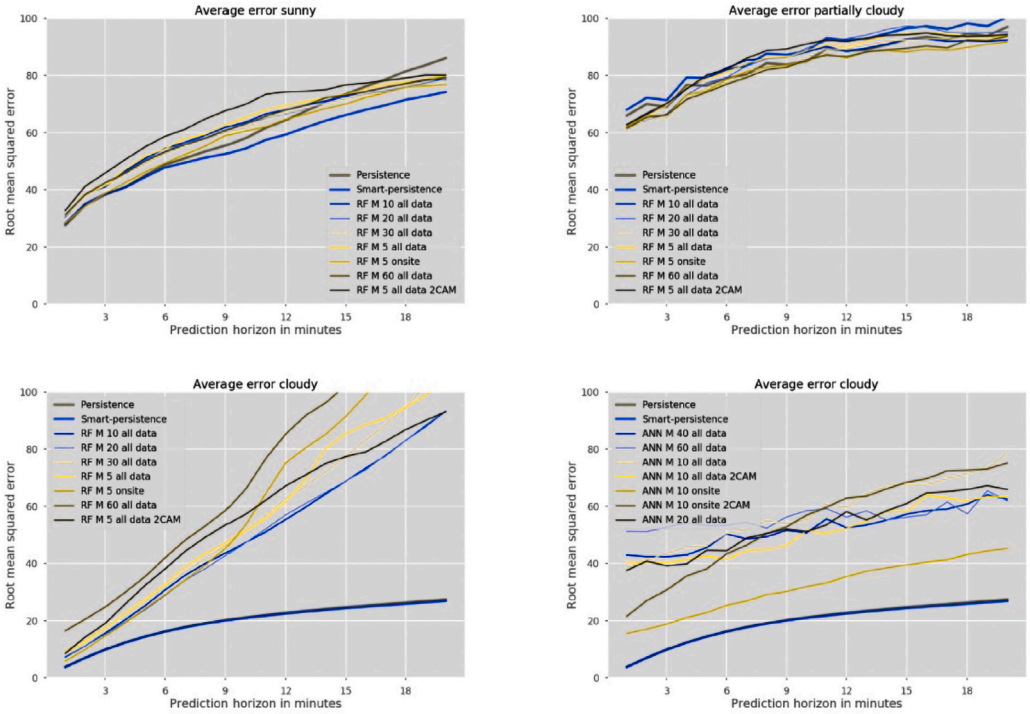


Fig. 11. Average RMSE on test-set for models RF and ANN.

Table 4
Average RF performance on test-set with weather circumstance: sunny (top) and partially cloudy (bottom).

Model	RMSE ↓	MAE ↓	MAPE ↓	Ramp-score ↓	FS-RMSE ↑	FS-MAE ↑	FS-MAPE ↑	FS-RAMP ↑
Persistence	59.83	33.65	28.58	26.68	NA	NA	NA	NA
Smart-persistence	55.37	19.77	84.26	21.3	NA	NA	NA	NA
RF M 10 all data	59.31	29.53	19.2	26.73	0.01	0.12	0.33	-0.0
RF M 20 all data	57.88	26.85	20.8	**25.24**	0.03	0.2	0.27	**0.05**
RF M 30 all data	57.5	26.78	20.14	25.56	0.04	0.2	0.3	0.04
RF M 5 all data	60.12	30.77	19.85	28.16	-0.0	0.09	0.31	-0.06
RF M 5 onsite	**56.73**	27.95	**18.36**	26.17	**0.05**	0.17	**0.36**	0.02
RF M 60 all data	58.41	**26.75**	19.58	26.2	0.02	**0.21**	0.31	0.02
RF M 5 all data 2CAM	62.25	32.37	22.38	28.9	-0.04	0.04	0.22	-0.08

Model	RMSE ↓	MAE ↓	MAPE ↓	Ramp-score ↓	FS-RMSE ↑	FS-MAE ↑	FS-MAPE ↑	FS-RAMP ↑
Persistence	84.59	48.08	37.12	29.31	NA	NA	NA	NA
Smart-persistence	87.85	47.82	74.22	28.66	NA	NA	NA	NA
RF M 10 all data	79.98	47.87	**44.82**	31.17	0.05	0.0	**-0.21**	-0.06
RF M 20 all data	80.41	47.95	50.05	31.21	0.05	0.0	-0.35	-0.06
RF M 30 all data	78.3	47.26	45.05	30.58	0.07	0.02	-0.21	-0.04
RF M 5 all data	81.11	48.66	46.5	31.59	0.04	-0.01	-0.25	-0.08
RF M 5 onsite	80.26	49.58	45.45	31.36	0.05	-0.03	-0.22	-0.07
RF M 60 all data	**77.4**	**46.43**	48.11	**29.61**	**0.08**	**0.03**	-0.3	**-0.01**
RF M 5 all data 2CAM	81.05	48.16	63.14	31.36	0.04	-0.0	-0.7	-0.07

circumstances and conditions. As the sequence length increases, the accuracy of the ANN models drops, indicating that longer sequences may not necessarily lead to better predictions. In sunny weather circumstances, the 'ANN M 10 all-data' model is able to predict well with a small difference compared to 'ANN M 10 onsite,' but it has a bad ramp-score with respect to the baselines. On the other hand, for partially cloudy weather, 'ANN M 10 onsite' outperforms persistence on all prediction horizons, but the 'all data' subset performs worse in this weather circumstance.

Choosing the appropriate error metric is crucial for accurate ANN model evaluation. For example, ramp-score evaluation may yield different results than MSE evaluation, impacting model performance. Additionally, the study reveals that shorter prediction horizons pose more challenges for ANN models, with the performance gap widening as the prediction horizon increases.

4.3. LSTM

The experiments conducted demonstrate that the performance of LSTM models is **better with shorter sequences**. In fact, any sequence length greater than 10 was **unreliable and resulted in some very off results**. The best performance was achieved with a sequence length of 5. On sunny weather, the '**LSTM M 5 all data**' model had the **best performance on average**, as shown in Fig. 13. However, none of the LSTM models were able to beat the baselines on very short prediction horizons, with the **break-even point being after 4 minutes** for the 'LSTM M 5 all data' model. The 'LSTM M 5 all data' model's ability to predict GHI under sunny conditions is depicted in Fig. 14, where it captures the general pattern of the observed data, yet exhibits discrepancies during periods of rapid irradiance change, highlighting the room for further refinement of the model. For partially cloudy weather, the '**LSTM M 5 onsite**' model performed slightly better with an **average**

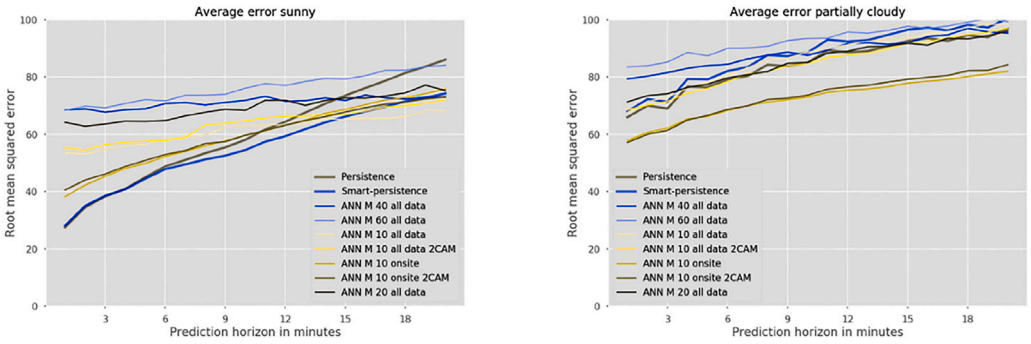


Fig. 12. Average RMSE on test-set for model ANN.

Table 5
Average ANN performance on test-set with weather circumstance: sunny (top) and partially cloudy (bottom).

Model	RMSE ↓	MAE ↓	MAPE ↓	Ramp-score ↓	FS-RMSE ↑	FS-MAE ↑	FS-MAPE ↑	FS-RAMP ↑
Persistence	59.83	33.65	28.58	26.68	NA	NA	NA	NA
Smart-persistence	55.37	19.77	84.26	21.3	NA	NA	NA	NA
ANN M 40 all data	66.52	41.71	41.63	33.16	-0.11	-0.24	-0.46	-0.24
ANN M 60 all data	71.21	42.79	34.63	34.77	-0.19	-0.27	**-0.21**	-0.3
ANN M 10 all data	57.2	35.53	38.79	28.4	0.04	-0.06	-0.36	-0.06
ANN M 10 all data 2CAM	59.13	37.26	36.97	30.83	0.01	-0.11	-0.29	-0.16
ANN M 10 onsite	55.51	**31.17**	59.92	**26.93**	**0.07**	**0.07**	-1.1	**-0.01**
ANN M 10 onsite 2CAM	55.37	33.53	103.71	27.84	**0.07**	0.0	-2.63	-0.04
ANN M 20 all data	64.63	40.28	52.96	32.29	-0.08	-0.2	-0.85	-0.21

Model	RMSE ↓	MAE ↓	MAPE ↓	Ramp-score ↓	FS-RMSE ↑	FS-MAE ↑	FS-MAPE ↑	FS-RAMP ↑
Persistence	84.59	48.08	37.12	29.31	NA	NA	NA	NA
Smart-persistence	87.85	47.82	74.22	28.66	NA	NA	NA	NA
ANN M 40 all data	81.75	51.35	79.3	36.1	0.03	-0.07	-1.14	-0.23
ANN M 60 all data	85.61	55.55	126.4	39.72	-0.01	-0.16	-2.4	-0.36
ANN M 10 all data	79.29	46.38	73.98	32.21	0.06	0.04	-0.99	-0.1
ANN M 10 all data 2CAM	77.08	47.3	70.14	32.88	0.09	0.02	-0.89	-0.12
ANN M 10 onsite	66.13	**39.46**	63.26	26.29	**0.22**	**0.18**	-0.7	0.1
ANN M 10 onsite 2CAM	66.83	39.56	**45.4**	**26.07**	0.21	**0.18**	**-0.22**	**0.11**
ANN M 20 all data	78.08	48.72	72.56	33.93	0.08	-0.01	-0.95	-0.16

Table 6
Average LSTM performance on test-set with weather circumstance: sunny (top) and partially cloudy (bottom).

Model	RMSE ↓	MAE ↓	MAPE ↓	Ramp-score ↓	FS-RMSE ↑	FS-MAE ↑	FS-MAPE ↑	FS-RAMP ↑
Persistence	**59.83**	33.65	28.58	**26.68**	NA	NA	NA	NA
Smart-persistence	55.37	19.77	84.26	21.3	NA	NA	NA	NA
LSTM M5 PXL	54.07	30.23	88.99	24.52	0.1	0.1	-2.11	0.08
LSTM M10 all data	52.25	28.23	**38.6**	24.3	0.13	0.16	**-0.35**	0.09
LSTM M5 all data	48.87	24.51	69.58	**21.33**	**0.18**	0.27	-1.43	0.2
LSTM M5 all data 2CAM	49.29	**23.47**	48.97	**20.91**	**0.18**	**0.3**	-0.71	**0.22**
LSTM M5 onsite	51.84	24.87	38.84	21.21	0.13	0.26	-0.36	0.21

Model	RMSE ↓	MAE ↓	MAPE ↓	Ramp-score ↓	FS-RMSE ↑	FS-MAE ↑	FS-MAPE ↑	FS-RAMP ↑
Persistence	**84.59**	48.08	37.12	**29.31**	NA	NA	NA	NA
Smart-persistence	87.85	47.82	74.22	28.66	NA	NA	NA	NA
LSTM M5 PXL	68.69	38.76	**45.62**	24.01	0.19	0.19	**-0.23**	0.18
LSTM M10 all data	76.13	49.54	622.03	32.88	0.1	-0.03	-15.76	-0.12
LSTM M5 all data	71.35	41.04	53.63	26.34	0.16	0.15	-0.44	0.1
LSTM M5 all data 2CAM	76.28	43.58	146.07	28.44	0.1	0.09	-2.93	0.03
LSTM M5 onsite	67.43	**37.34**	55.45	**23.43**	**0.2**	**0.22**	-0.49	**0.2**

RMSE of 67.43, as demonstrated in Table 6. **LSTM outperformed the baselines across all prediction horizons**. For short prediction horizons, '**LSTM M 5 onsite**' also performed better in sunny weather, as shown in Fig. 13. Additionally, **LSTM was the only tested model in this study that had a better ramp-score than (smart-)persistence** for both sunny and partially cloudy weather.

Table 7 summarizes model performance across weather conditions and prediction horizons. This summary highlights the models that perform well for specific weather conditions and prediction horizons and suggests the most appropriate error metric for evaluating the accuracy of the model. Most likely, this is due to the fact that the variable to predict GHI is a feature onsite (and all features), making it much easier to predict.

Onsite data outperforms other sources in the implemented models. Model selection favored the subsets "all-data" and "onsite". In the test set, "onsite" performs better for RF and ANN models. LSTM "all data" excels in sunny weather, while "onsite" is better for partially cloudy conditions (averaging over all prediction horizons).

4.4. Valuable features

In our analysis, models with complete data access performed best in partially cloudy weather (Fig. 15). Conversely, for sunny conditions,

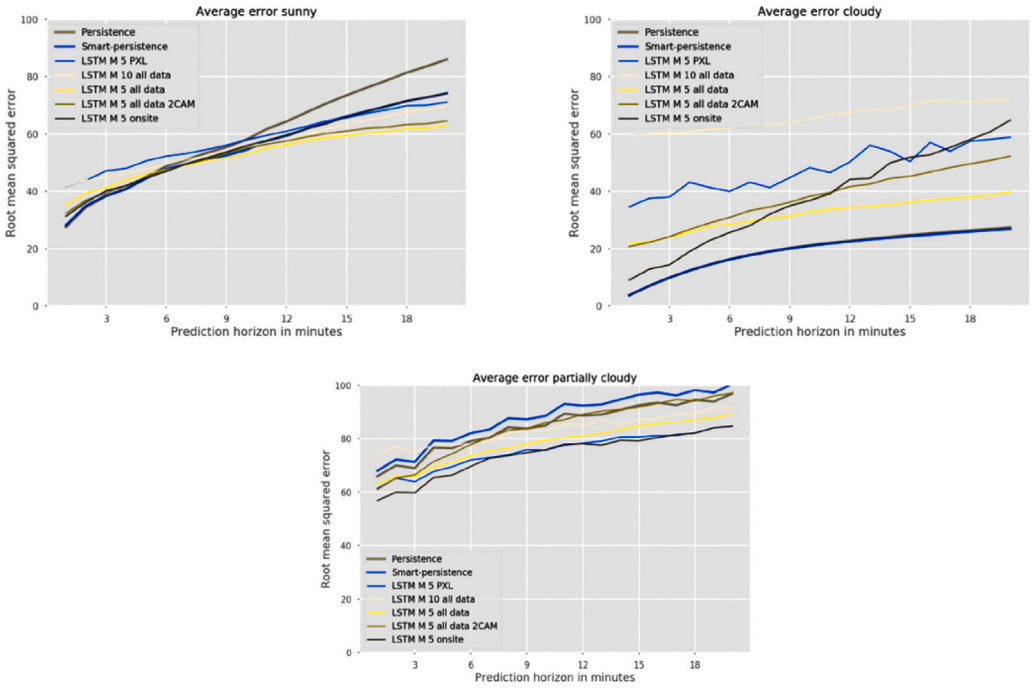


Fig. 13. Average RMSE on test-set for model LSTM.

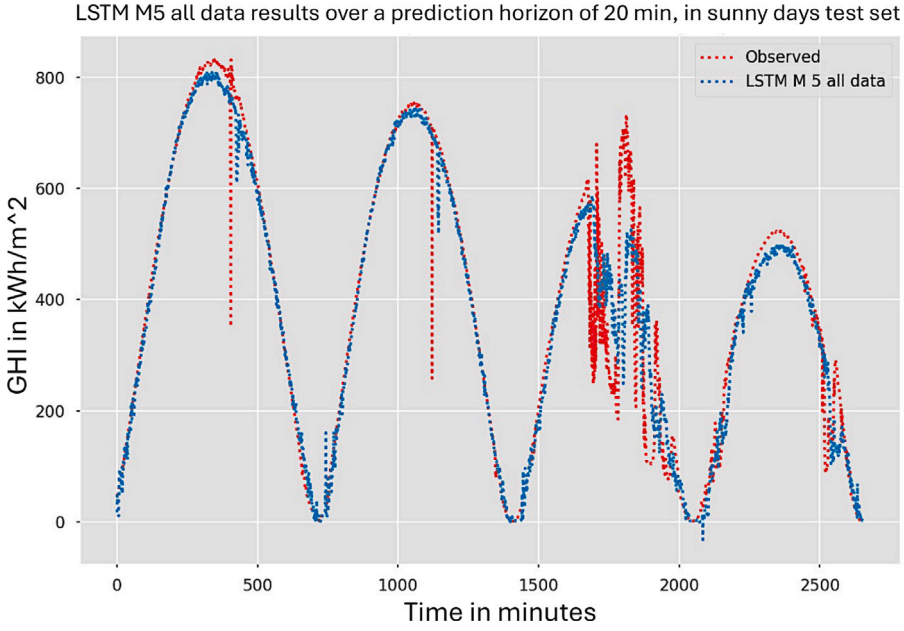


Fig. 14. Result LSTM M 5 all data, prediction horizon 20, sunny days test set.

on-site sensor data-based models excelled in shorter predictions (up to 7 min, Fig. 15 for analysis-based MAE).

4.5. Different weather circumstances

As discussed, the dataset is unbalanced due to the location in the south of Spain. Therefore, we only discuss sunny and partially cloudy weather.

A - Sunny Weather Conditions

Table 8 displays the average performance on the test-set under "sunny" weather conditions. The 'LSTM M 5 all data' model stands out with the lowest RMSE of 58.43, indicative of its accurate predictions. Notably, it also achieves high fairness metrics, with FS-RMSE of 0.19 and FS-MAE of 0.27, underscoring its equitable performance across diverse groups. Furthermore, the model demonstrates a low MAPE of 61.81%, indicating its capability for precise percentage predictions. These findings establish 'LSTM M 5 all data' as a strong candidate for forecasting in "sunny" weather circumstances. Remarkably, it outperforms the baseline substantially, with $FS-RMSE = 18\%$ and $FS-MAE/FS-Ramp = 22\%$, while comparative analysis against ANN ($FS-RMSE = 8\%$) and RF ($FS-RMSE = 5\%$, $FS-MAE = 21\%$, and $FS-Ramp = 5\%$) further reinforces LSTM's superiority in predicting accurately under "sunny" weather conditions.

B - Partially Cloudy Weather Conditions

In Table 9 is shown that 'LSTM M 5 all data' outperforms all other models. On average (overall prediction horizons) this model has RMSE

Table 7
Summary of Test-set Results for Method Selection based on Weather Conditions and Time Horizon: Comparing 'On Site' and 'All Data' Approaches with RF, ANN, and LSTM Methods for a Specific Weather Condition and Horizon, along with Baseline Evaluation Metrics.

	Weather conditions	*Short_Horizons<10 min*	*Long_Horizons>= 10 min*	*Horizon of 20 min*	*Error metrics*
RF	***Sunny***	*FS-RMSE 5%, FS-MAE 21% And FS-Ramp 5%.*			**MAE**
RF	***Partially Cloudy***		*FS-RMSE 8%, MAE 3%, FS-RAMP 20%*		
RF	***cloudy***	—		—	
ANN	***Sunny***	*<8% for all the error metrics*			**MSE**
ANN	***Partially Cloudy***		*FS-RMSE 22% FS-MAE 18%, FS-RAMP 11%*		
ANN	***Cloudy***	—	—		
LSTM	***Sunny***	*FS-RMSE 18% FS-MAE/FS-Ramp of 22%.*			**RMSE**
LSTM	***Partially Cloudy***		*FS-RMSE 20%, FS-MAE 22%, FS-RAMP 0%*		
LSTM	***cloudy***	—			

Table 8
Average performance on test-set, with weather circumstance sunny.

Model	RMSE ↓	MAE ↓	MAPE ↓	Ramp-score ↓	FS-RMSE ↑	FS-MAE ↑	FS-MAPE ↑	FS-RAMP ↑
Persistence	72.52	39.16	27.84	27.69	NA	NA	NA	NA
Smart-persistence	70.27	29.38	150.44	25.57	NA	NA	NA	NA
ANN M 5 all data	60.34	32.63	48.53	26.48	0.17	0.17	−0.74	0.04
LSTM M 5 all data	**58.43**	28.77	61.81	23.98	**0.19**	**0.27**	−1.22	0.13
LSTM M 5 onsite	59.36	**28.76**	54.77	**23.85**	0.18	**0.27**	−0.97	**0.14**
LSTM M 5 onsite+img	62.43	31.13	56.5	25.09	0.14	0.21	−1.03	0.09
RF M 5 onsite	63.65	30.5	16.54	26.31	0.12	0.22	0.41	0.05
RF M 60 onsite	63.72	30.71	**13.18**	27.63	0.12	0.22	**0.53**	0.0

Table 9
Average performance on test-set, with weather circumstance partially cloudy.

Model	RMSE ↓	MAE ↓	MAPE ↓	Ramp-score ↓	FS-RMSE ↑	FS-MAE ↑	FS-MAPE ↑	FS-RAMP ↑
Persistence	96.3	53.0	41.81	36.19	NA	NA	NA	NA
Smart-persistence	97.26	53.06	63.93	36.14	NA	NA	NA	NA
ANN M 5 all data	85.58	52.27	71.81	38.15	0.11	0.01	−0.72	−0.05
LSTM M 5 all data	**76.83**	**43.12**	62.85	**29.04**	**0.2**	**0.19**	−0.5	**0.2**
LSTM M 5 onsite	81.33	44.42	56.19	30.98	0.16	0.16	−0.34	0.14
LSTM M 5 onsite+img	85.42	45.53	69.27	32.04	0.11	0.14	−0.66	0.11
RF M 5 onsite	99.09	61.1	41.09	43.14	−0.03	−0.15	0.02	−0.19
RF M 60 onsite	89.39	57.26	**36.49**	39.58	0.07	−0.08	**0.13**	−0.09

76.83, MAE 43.12, MAPE 62.85, and a Ramp-score of 29.04. Relative to other models and baseline 'LSTM M 5 all data' performs better with relatively greater prediction horizons. Additionally, for a prediction horizon of 20 min 'LSTM M 5 all data' has $FS_{rmse} = 22\%$ and $FS_{mae} = 24\%$.

4.6. Predicting fluctuations

The ramp-score represents the ability to predict fluctuations. For sunny weather circumstances, we observe that until 11 min, the baselines perform better. Afterward, model 'LSTM M 5 all data' performs best (visible at the bottom of Fig. 15, an analysis-based Ramp-score).

We observe at the top of Fig. 15 and Table 9 that 'LSTM M 5 all data' performs best in partially cloudy weather on average. However, on a prediction horizon before 5 min, the baselines perform better.

Additionally, for a prediction horizon of 20 min 'LSTM M 5 all data' has $FS_{ramp} = 39\%$ for sunny weather and $FS_{ramp} = 25\%$ for partially cloudy weather.

4.7. Computation time

An application of this study would be to predict GHI for an actual solar farm to optimize energy generation. When a model is more computationally intensive it will use more energy. In this section, we highlight the resources needed to perform predictions on a single day. Usually, neural networks train on a GPU, but this would not be a fair comparison to RF. To straighten it out training and prediction will be done on only the CPU (see Section 3.6). 'Single' models would use 20 times more time, as shown in Table 3.

4.8. Statistical significance

To determine if a model performs significantly better (see Section 3.5) than the baseline, the 'Diebold–Mariano' test is applied. The null hypothesis 'competing model performs equal to the baseline persistence' is rejected when $p < 0.05$. In the given models below we define 'n/a' if that particular model does not predict better than the baseline for a certain prediction horizon.

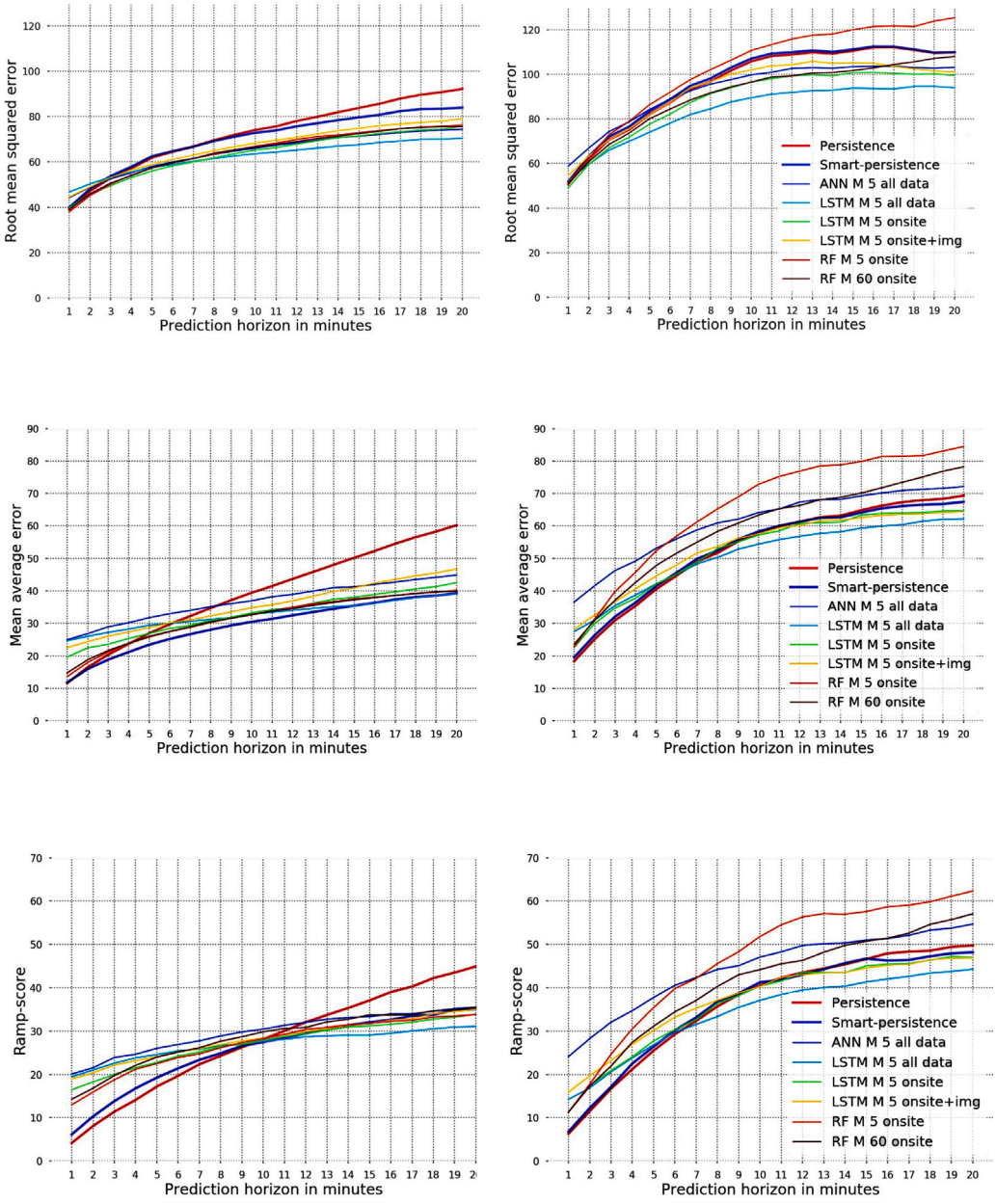


Fig. 15. The top two graphs display the test-set results for RMSE in both sunny (left) and partially cloudy (right) weather conditions. The middle graphs show the test-set results for MAE in sunny (left) and partially cloudy (right) weather conditions. The bottom graphs present the test-set results for ramp rate in both sunny (left) and partially cloudy (right) weather conditions.

In sunny weather circumstances, the only model that is significantly better on short horizons is RF. On greater prediction horizons LSTM models perform (much) better (see Table 10).

In partially cloudy weather we observe (see Table 11) that 'LSTM M 5 all data' is most statistically significant over (excluding $p = 1..7$ for model 'LSTM 5 onsite') all prediction horizon 7..20.

4.9. Model comparison

The RF model is a good choice due to its simplicity, low training time, and relatively low need for hyperparameter tuning and normalization. Additionally, it performs reasonably well across most sequence lengths. When averaging the results' overall prediction horizons, the best RF model achieves an RMSE of 56.73 for sunny weather conditions and 77.4 for partially cloudy conditions (see Table 4). However, its performance is worse than other implemented models across all weather circumstances and prediction horizons (see Figs. 16 and 17).

The ANN model improves on average in sunny weather, achieving an RMSE of 55.51, but performs worse on short horizons compared to RF. On the other hand, it performs better on average for larger horizons and partially cloudy weather with an RMSE of 66.13. The LSTM model outperforms other models on sunny weather with an RMSE of 48.87 and performs significantly better than persistence starting from a prediction horizon of 11 min compared to ANN's 15 min. For partially cloudy weather, LSTM performs slightly worse than ANN with an RMSE of 67.43. LSTM also has the best ramp-score among all tested models and baselines, with 21.91 for sunny weather and 23.43 for partially cloudy weather (see Figs. 16 and . 17).

All models perform relatively well in partially cloudy weather compared to the baseline, likely because the baseline assumes some continuity. However, in partially cloudy weather, there is more fluctuation, and the models perform differently. When considering the furthest prediction horizon of 20 min, the difference with (smart-)persistence is more significant, and the best model achieved an improvement of

Table 10
Diebold–Mariano p value per horizon, compared with baseline model Persistence on Sunny Weather circumstance. (n/a implies the competing model is not performing better than the baseline).

Prediction horizon	ANN M 5 all data	LSTM M 5 all data	LSTM M 5 onsite	LSTM M 5 onsite+img	RF M 5 onsite	RF M 60 onsite
1	n/a	n/a	n/a	n/a	**0.044**	**0.014**
2	n/a	n/a	8.816e−02	n/a	3.164e−03	5.318e−05
3	6.521e−01	n/a	8.132e−05	n/a	1.553e−03	3.597e−06
4	3.226e−02	2.121e−01	3.549e−06	4.786e−01	2.593e−03	6.255e−07
5	2.289e−04	4.911e−03	1.341e−08	1.138e−02	1.238e−04	2.650e−07
6	3.499e−05	4.326e−04	3.491e−08	3.289e−03	1.170e−04	4.033e−07
7	2.056e−06	8.493e−05	2.164e−09	7.422e−04	4.717e−05	5.121e−07
8	7.769e−07	3.335e−06	6.609e−09	4.542e−04	4.961e−05	1.534e−06
9	4.226e−08	8.979e−08	6.113e−10	9.425e−06	1.166e−05	7.440e−07
10	3.163e−09	9.578e−09	7.175e−11	4.188e−07	9.431e−06	4.243e−07
11	4.255e−09	2.945e−09	1.233e−10	4.286e−08	1.366e−05	8.134e−07
12	3.277e−10	7.508e−10	2.247e−10	5.840e−09	8.901e−06	5.246e−07
13	7.870e−11	1.266e−10	4.877e−10	4.270e−09	5.095e−06	8.214e−07
14	4.237e−12	1.018e−11	5.353e−11	4.417e−11	6.378e−07	3.136e−07
15	4.888e−14	7.009e−13	2.940e−12	1.916e−12	4.500e−07	1.877e−07
16	5.713e−15	8.494e−14	1.177e−13	1.655e−13	8.886e−08	8.725e−08
17	9.590e−15	8.275e−14	1.694e−13	6.031e−13	3.420e−08	2.921e−08
18	5.488e−15	2.494e−14	7.025e−13	8.552e−13	2.268e−08	5.279e−09
19	1.018e−16	1.526e−15	1.251e−14	2.065e−13	2.628e−09	1.821e−10
20	1.185e−18	9.127e−18	4.575e−16	2.118e−15	5.153e−10	1.956e−12

Table 11
Diebold–Mariano p value per horizon, compared with baseline model Persistence. Weather circumstance partially cloudy. (n/a implies competing model is not performing better than the baseline).

Prediction horizon	ANN M 5 all data	LSTM M 5 all data	LSTM M 5 onsite	LSTM M 5 onsite+img	RF M 5 onsite	RF M 60 onsite
1	n/a	n/a	0.359	n/a	n/a	0.454
2	n/a	0.339	**0.146**	n/a	n/a	**0.121**
3	n/a	**0.121**	**0.104**	0.987	n/a	**0.123**
4	n/a	**0.089**	**0.127**	n/a	n/a	0.204
5	n/a	**0.012**	**0.063**	0.950	n/a	0.179
6	n/a	5.655e−03	2.653e−02	n/a	n/a	7.502e−02
7	7.782e−01	5.809e−03	3.301e−02	n/a	n/a	3.916e−02
8	6.696e−01	6.436e−03	6.454e−02	9.466e−01	n/a	4.891e−02
9	2.034e−01	6.415e−03	1.482e−02	5.030e−01	n/a	2.954e−02
10	1.273e−01	1.216e−02	1.979e−02	1.285e−01	n/a	3.809e−02
11	9.787e−02	1.332e−02	1.884e−02	8.979e−02	n/a	4.896e−02
12	1.853e−01	1.872e−02	4.282e−02	1.218e−01	n/a	6.229e−02
13	1.388e−01	1.560e−02	2.170e−02	1.053e−01	n/a	6.579e−02
14	1.601e−01	1.654e−02	2.443e−02	5.642e−02	n/a	8.510e−02
15	1.491e−01	1.370e−02	4.154e−02	2.197e−02	n/a	9.319e−02
16	1.158e−01	8.457e−03	2.248e−02	8.387e−03	n/a	1.029e−01
17	1.213e−01	8.335e−03	1.757e−02	5.048e−03	n/a	1.436e−01
18	1.252e−01	6.013e−03	1.539e−02	3.294e−03	n/a	2.145e−01
19	1.718e−01	6.316e−03	2.091e−02	2.241e−03	n/a	3.536e−01
20	0.183	0.006	0.015	0.001	n/a	0.403

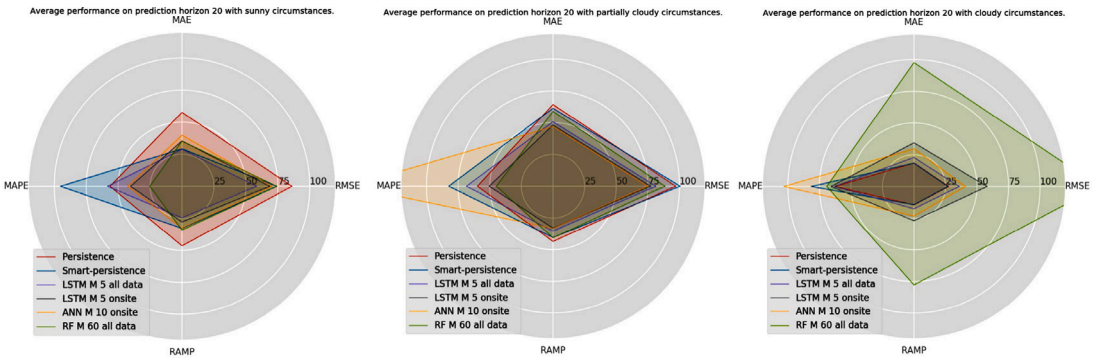


Fig. 16. Average performance on prediction horizon of 20min with cloudy, partially cloudy and sunny weather circumstances.
Note: smaller error in GHI equals better performing.

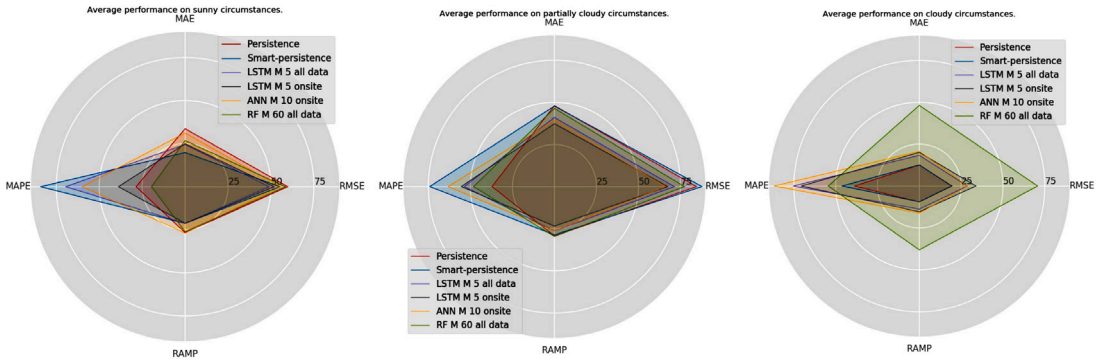


Fig. 17. Average performance overall prediction horizon of 20min with cloudy, partially cloudy, and sunny weather circumstances.
Note: smaller error in GHI equals better performance.

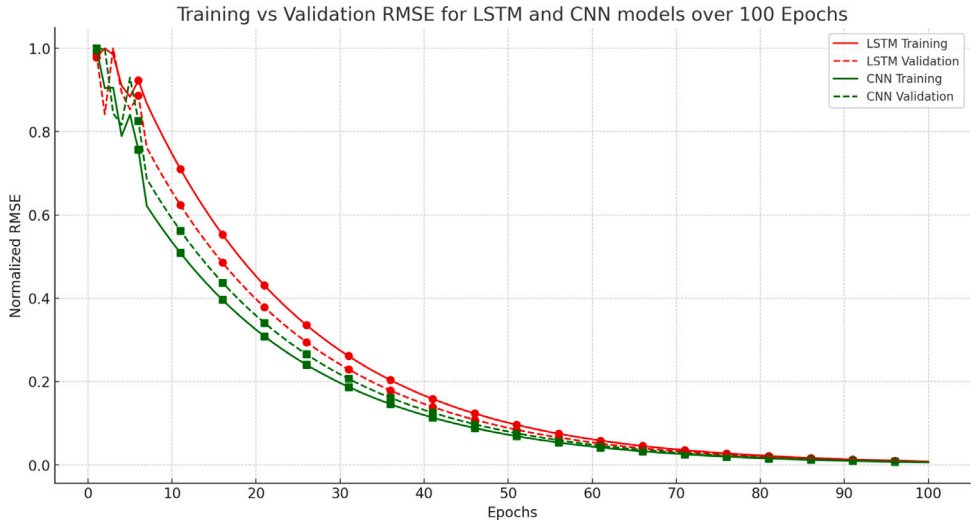


Fig. 18. Training vs Validation RMSE for LSTM and CNN models over several epochs.

32% in RMSE, 50% in MAE, and 48% in ramp-score (with respect to the baseline) (see Tables A.1–A.3).

Overall, considering that the full dataset consists of 121 sunny days, 29 partially cloudy days, and 3 cloudy days, the LSTM model performs best on average.

4.10. Model benchmark

We conducted a comparative analysis of our deep learning model in relation to the SKIPP'D method, a cutting-edge technique for short-term solar forecasting using sky imagery introduced by [3]. Our experimental approach involved two different strategies: initially, we directly employed the pre-trained SKIPP'D model on our test data, followed by a fine-tuning process using the training partition of our dataset.

As depicted in Fig. 18, the LSTM model demonstrated a remarkable performance over the CNN model, particularly in validation scenarios, thus highlighting its robustness in capturing temporal dynamics which are crucial for solar forecasting. The LSTM model showed a more significant reduction in RMSE across epochs for both training and validation, which indicates a stronger generalization capability compared to the CNN model. The performance disparity is largely attributable to the different meteorological conditions embedded in the datasets used for developing the SKIPP'D model compared to our own data. However, the fine-tuning of the SKIPP'D model on our dataset led to enhanced forecasting accuracy, signifying the versatility and adaptability of our approach across varied environmental contexts. This benchmarking exercise not only validates the effectiveness of our model but also establishes a comprehensive comparative framework within the solar forecasting domain, showcasing our commitment to continuous improvement and adaptability in the face of changing data dynamics.

5. Discussion

In this section, we discuss the results of all tested models. The results give insight into what models work well for what particular prediction horizons and weather circumstances. Additionally, a notion is given of what data is suitable as input for these models. However, these are sets of multiple features (in the example set 'meteor'). The elements in a set have a similar data source. But, for every subset of features, there is a possibility that it includes a feature that is not relevant for good predictions. It may be that dropping some of these features or partially merging subsets will increase forecast performance.

The RF and ANN models achieve their best performance when using only the "onsite" feature subset, indicating that they fail to capture valuable information from features extracted from images. However, LSTM outperforms them when given access to this subset. By grasping the complexity of these features, LSTM delivers the best predictions among all models tested. Furthermore, the "all-data" feature set performs best for LSTM on average, under all weather conditions.

In sunny conditions, our models underperform compared to baselines for short-term predictions (under 5 min). The performance gap narrows in partially cloudy weather or beyond the initial fiveminute period. Baselines rely on continuous GHI and CSI values, with sunny

weather showing less GHI fluctuation, explaining the performance differential in varying weather conditions.

Among all the models we tested, the superior performers outperform smart-persistence under partially cloudy and sunny weather conditions. However, none of the models were able to outperform smart-persistence when predicting cloudy weather. This may be due to a shortage of cloudy days in our dataset and may not occur under more balanced weather conditions. In partially cloudy weather circumstances, the absolute difference between LSTM and the baseline is greater than in sunny weather conditions. Nevertheless, when examining the relative difference, this disparity is not present.

In this study, we computed the ramp-score by averaging the values every 5 min. In prior literature, it was suggested to normalize over an hour when forecasting 5 h into the future. The 5-minute averages capture fluctuations, but it is an estimation.

Our benchmarking activities entailed a comparative analysis between our deep learning model and the SKIPP'D method, known for its use of sky imagery in short-term solar forecasting [3]. The process involved two phases: first, deploying the pre-trained SKIPP'D model on our test data, and then fine-tuning it with our dataset. This approach led to notable improvements in accuracy. As shown in Fig. 18, the LSTM model outperformed in validation, demonstrating superior temporal dynamics capture, crucial for solar forecasting. This model showed a consistent reduction in RMSE (Root Mean Square Error) in training and validation, highlighting its robust generalization, unlike the CNN model. The performance disparity is largely due to the different meteorological conditions in the datasets. However, the successful fine-tuning of the SKIPP'D model on our data highlights its adaptability in various environments. This benchmarking not only confirms our model's effectiveness but also establishes a comprehensive comparative framework in solar forecasting, underscoring our commitment to ongoing improvement and adaptability in a dynamic data landscape. Moreover, this exercise exemplifies our method's flexibility across diverse environmental conditions.

It is important to note that our research data comes exclusively from Almería, Spain. Recent approaches have shown the effectiveness of transfer learning and dataset fusion in solar forecasting across various locations [39]. Similarly, [40] explored the performance of deep learning models using sky image datasets collected from diverse global locations with different climate patterns. They compared local models, global models trained on fused datasets, and transfer learning models. Their findings suggest that local models work well within their original context, but exhibit significant errors when applied offsite. In contrast, global models adapt well to individual locations, and transfer learning models, trained on a large diversified source dataset, generally achieve superior performance over other strategies.

These studies highlight the potential of DL models in solar forecasting across different geographical and climatic conditions, especially when utilizing transfer learning and dataset fusion approaches. They underscore the importance of considering data diversity and model adaptability, and propose resampling and data augmentation as further ways to increase robustness.

6. Conclusions

This study describes the application and implementation of Long Short-Term Memory (LSTM) networks to forecast Global Horizontal Irradiance (GHI) using different subsets of data input. We compared this with more traditional methods like Random Forest (RF), Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and LSTM. In our pursuit of identifying a good set of hyper-parameters, we investigated the importance of feature subsets. Our results indicated that onsite measurement equipment was the most crucial feature subset. However, certain features demonstrated varied performance depending on weather circumstances. For instance, the subset 'meteor' was valuable in sunny weather, but its usefulness was limited in partially cloudy weather. Additionally, to validate the accuracy of our method with the advancing technologies in the state-of-the-art, a benchmarking study was conducted with [3]. The benchmarking results further highlighted the LSTM model's proficiency, especially when fine-tuned with additional data, demonstrating its adaptability and indicating potential for future application.

We initially used Clear Sky Index (CSI) to estimate GHI, but this approach did not enhance our models. Therefore, we realized that directly predicting GHI would be more effective. Additionally, we experimented with incorporating data from a different location into our models, but this did not result in any improvements.

Regarding the imbalance in our dataset due to the specific weather circumstances in southern Spain, we acknowledge that this is a limitation that we cannot change. However, following the suggestion of [41], we recognize that resampling and data augmentation methods could be employed to address this issue. Such methods, as also suggested by [42], might include techniques like oversampling the minority class or generating synthetic examples, which could potentially improve the model's performance in less-represented weather conditions.

Our findings contribute to the current literature by presenting an approach that combines All Sky Images (ASI) and numerical data in machine learning. Furthermore, our study highlights the importance of using short sequences of features for forecasting GHI instead of long sequences.

7. Future work

The data set's imbalance, caused by the predominant weather conditions in southern Spain, presents a challenge. A multi-year approach may offer a viable solution, as our models currently underperform in cloudy conditions due to the scarcity of such days. Implementing data augmentation could be an effective strategy not only to address this imbalance but also to enhance overall prediction accuracy.

Exploring advanced DL methods that process sky imagery, like convolutional or transfer-based models, could significantly refine our visual data analysis capabilities.

We are currently considering the three days before the prediction moment for early stopping. This duration was chosen through manual tuning, and further research could explore better time frames. Moreover, there is potential in treating cloud pixels differently based on their proximity to the sun, rather than treating all cloud pixels uniformly.

In this study, we had access to an additional site with similar equipment. Expanding the network of sensor sites around the prediction location and incorporating wind speed and direction data could significantly improve the accuracy of predictions by leveraging real-time data from multiple locations.

Acronym

GHIclr	Clear sky GHI
ANN	Artificial Neural Network
ASI	All Sky Imager
CNN	Convolutional neural network
CSI	Clearness index
DHI	Diffuse Horizontal Irradiance
DNI	Direct Normal Irradiance
GHI	Global horizon irradiance
LSTM	Long short-term memory
MAE	Mean Absolute Error

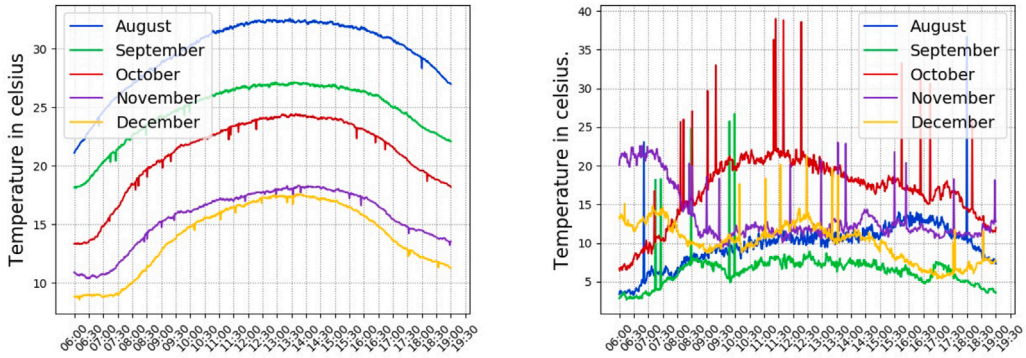


Fig. A.1. Average (left) and variance (right) in temperature (in Celsius) at camera site 1.

Table A.1
Average performance on test-set with prediction horizon 20 min, with weather circumstance sunny.

Model	RMSE ↓	MAE ↓	MAPE ↓	Ramp-score ↓	FS-RMSE ↑	FS-MAE ↑	FS-MAPE ↑	FS-RAMP ↑
Persistence	92.17	60.21	51.94	44.93	NA	NA	NA	NA
Smart-persistence	83.9	39.23	159.26	35.03	NA	NA	NA	NA
ANN M 5 all data	69.34	39.54	32.31	31.39	0.25	0.34	0.38	0.3
LSTM M 5 all data	**65.66**	**34.23**	47.03	**27.53**	**0.29**	**0.43**	0.09	**0.39**
LSTM M 5 onsite	**70.48**	37.05	**40.51**	29.58	0.24	0.38	0.22	**0.34**
LSTM M 5 onsite+img	73.75	41.07	40.44	31.22	0.2	0.32	0.22	0.31
RF M 5 onsite	75.09	39.01	20.09	33.05	0.19	0.35	0.61	0.26
RF M 60 onsite	74.01	38.28	**18.93**	33.88	0.2	0.36	**0.64**	0.25

MAPE Mean Absolute Percentage Error

MLP Multilayer perceptron

MSE Mean Square Error

NWP Numerical weather prediction

RF Random Forests

RMSE Root Mean Square Error

RNN Recurrent neural network

SS Skill Score

SVM Support vector machine

SVR Support vector regression

SZA Solar zenith angle

TSI Total solar irradiance

S Single Model

M Multi-Model

CRediT authorship contribution statement

N.Y. Hendrikx: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Visualization. **K. Barhmi:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Visualization. **L.R. Visser:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision, Project administration. **T.A. de Bruin:** Software, Coding, Investigation. **A.A. Salah:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision, Project administration. **W.G.J.H.M. van Sark:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Wilfried Van Sark reports financial support was provided by Netherlands Enterprise Agency and the Dutch Research Council.

Data availability

Data and models are available from the authors per reasonable request. Additionally, the models can be found on GitHub at https://github.com/nielsyh/ASI_playground.

Acknowledgments

We acknowledge support from the Dutch Research Council NWO, The Netherlands in the framework of the Energy Systems Integration & Big Data program, project Energy intrAneTs (NEAT) and from the Ministry of Economic Affairs and Climate, The Netherlands in the framework of the Solar Forecasting with All-Sky Imagers (SolFaSi) project. This work was part of an ASI benchmark study in the framework of IEA-PVPS Task 16, and we are very grateful for many fruitful discussions and the local support at the Plataforma Solar de Almería.

Appendix

Figs. A.1 and A.2 show monthly averages of temperature and humidity at the location of camera 1.

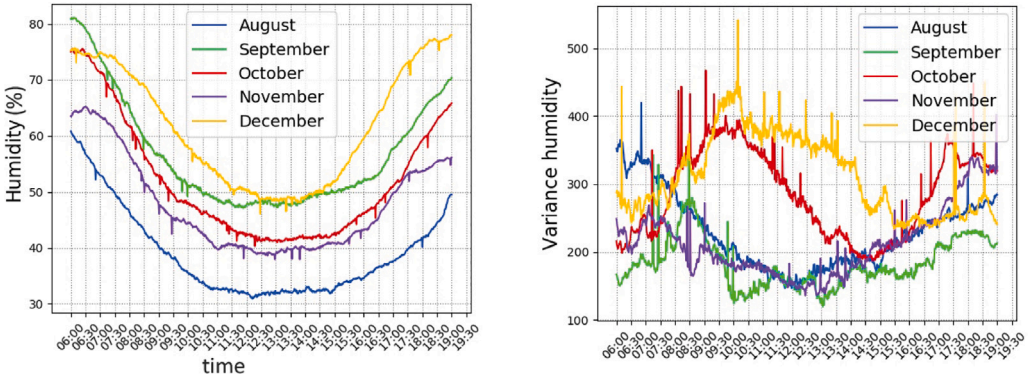


Fig. A.2. Average (left) and variance (right) humidity at camera site 1.

Table A.2
Average performance on test-set with prediction horizon 20 min, with weather circumstance partially cloudy.

Model	RMSE ↓	MAE ↓	MAPE ↓	Ramp-score ↓	FS-RMSE ↑	FS-MAE ↑	FS-MAPE ↑	FS-RAMP ↑
Persistence	109.83	69.36	66.33	49.76	NA	NA	NA	NA
Smart-persistence	109.87	67.43	79.26	48.18	NA	NA	NA	NA
ANN M 5 all data	94.47	61.61	**50.28**	46.49	0.14	0.11	**0.24**	0.07
LSTM M 5 all data	**86.03**	**52.42**	97.7	**37.1**	**0.22**	**0.24**	−0.47	**0.25**
LSTM M 5 onsite	91.07	54.38	53.95	39.57	0.17	0.22	0.19	0.2
LSTM M 5 onsite+img	92.54	54.58	52.09	40.08	0.16	0.21	0.21	0.19
RF M 5 onsite	119.71	77.08	64.5	56.73	−0.09	−0.11	0.03	−0.14
RF M 60 onsite	105.42	74.57	70.16	54.65	0.04	−0.08	−0.06	−0.1

Table A.3
Average performance on test-set with prediction horizon 20, with weather circumstance cloudy.

Model	RMSE ↓	MAE ↓	MAPE ↓	Ramp-score ↓	FS-RMSE ↑	FS-MAE ↑	FS-MAPE ↑	FS-RAMP ↑
Persistence	27.39	18.51	62.57	14.52	NA	NA	NA	NA
Smart-persistence	26.87	18.36	81.59	14.36	NA	NA	NA	NA
LSTM M5 PXL	52.72	37.61	66.37	29.47	−0.92	−1.03	−0.06	−1.03
LSTM M10 all data	64.78	49.61	66.32	39.55	−1.37	−1.68	−0.06	−1.72
LSTM M5 all data	**35.85**	**22.83**	63.07	**17.59**	**−0.31**	**−0.23**	−0.01	**−0.21**
LSTM M5 all data 2CAM	46.78	32.53	**60.27**	25.47	−0.71	−0.76	**0.04**	−0.75
LSTM M5 onsite	57.98	34.35	65.76	26.94	−1.12	−0.86	−0.05	−0.86

References

[1] P. Singla, M. Duhan, S. Saroha, A comprehensive review and analysis of solar forecasting techniques, Front. Energy (2021) 1–37.

[2] S.-A. Logothetis, V. Salamalikis, B. Nouri, J. Remund, L.F. Zarzalejo, Y. Xie, S. Wilbert, E. Ntavelis, J. Nou, N. Hendrikx, et al., Solar irradiance ramp forecasting based on all-sky imagers, Energies 15 (17) (2022) 6191.

[3] Y. Nie, X. Li, A. Scott, Y. Sun, V. Venugopal, A. Brandt, SKIPP'D: A SKy images and photovoltaic power generation dataset for short-term solar forecasting, Sol. Energy 255 (2023) 171–179.

[4] B. Elsinga, Chasing the Clouds: Irradiance Variability and Forecasting for Photovoltaics (Ph.D. thesis), University Utrecht, 2017.

[5] TenneT, Imbalance pricing system, 2020, URL https://www.tennet.eu/fileadmin/user_upload/SO_NL/Imbalance_pricing_system.pdf.

[6] H. Yang, B. Kurtz, D. Nguyen, B. Urquhart, C.W. Chow, M. Ghonima, J. Kleissl, Solar irradiance forecasting using a ground-based sky imager developed at UC San Diego, Sol. Energy 103 (2014) 502–524, http://dx.doi.org/10.1016/j.solener.2014.02.044, URL https://www.sciencedirect.com/science/article/pii/S0038092X14001327.

[7] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F.M. de Pison, F. Antonanzas-Torres, Review of photovoltaic power forecasting, Sol. Energy 136 (2016) 78–111, http://dx.doi.org/10.1016/j.solener.2016.06.069, URL https://www.sciencedirect.com/science/article/pii/S0038092X1630250X.

[8] EKO Instruments, ASI-16 all sky imager, 2022, URL https://www.eko-instruments.com/eu/categories/products/all-sky-imagers/asi-16-all-sky-imager.

[9] Y. Ai, Y. Peng, W. Wei, A model of very short-term solar irradiance forecasting based on low-cost sky images, AIP Conf. Proc. 1839 (2017) 020022, http://dx.doi.org/10.1063/1.4982387.

[10] S. Tiwari, R. Sabzehgar, M. Rasouli, Short term solar irradiance forecast based on image processing and cloud motion detection, 2019, pp. 1–6, http://dx.doi.org/10.1109/TPEC.2019.8662134.

[11] R. Chauvin, J. Nou, S. Thil, A. Traore, S. Grieu, Cloud detection methodology based on a sky-imaging system, Energy Procedia 69 (2015) 1970–1980, http://dx.doi.org/10.1016/j.egypro.2015.03.198.

[12] Y. Sun, G. Szűcs, A.R. Brandt, Solar PV output prediction from video streams using convolutional neural networks, Energy Environ. Sci. 11 (7) (2018) 1811–1818.

[13] J. Tang, Z. Lv, Y. Zhang, M. Yu, W. Wei, An improved cloud recognition and classification method for photovoltaic power prediction based on total-sky-images, J. Eng. 2019 (2019) http://dx.doi.org/10.1049/joe.2018.9249.

[14] E. Kassianov, C. Long, M. Ovchinnikov, Cloud sky cover versus cloud fraction: Whole-sky simulations and observations, J. Appl. Meteorol. 44 (2005) http://dx.doi.org/10.1175/JAM-2184.1.

[15] Y. Sun, V. Venugopal, A.R. Brandt, Short-term solar power forecast with deep learning: Exploring optimal input and output configuration, Sol. Energy 188 (2019) 730–741.

[16] Q. Li, W. Lyu, J. Yang, A hybrid thresholding algorithm for cloud detection on ground-based color images, J. Atmos. Ocean. Technol. 28 (2011) 1286–1296, http://dx.doi.org/10.1175/JTECH-D-11-00009.1.

[17] J. Canny, A computational approach to edge detection, IEEE Trans. Pattern Anal. Mach. Intell. 8 (1986) 679–698, http://dx.doi.org/10.1109/TPAMI.1986.4767851.

[18] B.Y. Liu, R.C. Jordan, The interrelationship and characteristic distribution of direct, diffuse and total solar radiation, Sol. Energy 4 (3) (1960) 1–19, http://dx.doi.org/10.1016/0038-092X(60)90062-1, URL https://www.sciencedirect.com/science/article/pii/0038092X60900621.

[19] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, http://dx.doi.org/10.1023/A:1010933404324.

[20] A. Liaw, M. Wiener, Classification and regression by randomforest, R News 2 (3) (2002) 18–22.

[21] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016, http://www.deeplearningbook.org.

[22] W. Feng, N. Guan, Y. Li, X. Zhang, Z. Luo, Audio visual speech recognition with multimodal recurrent neural networks, 2017, pp. 681–688, http://dx.doi.org/10.1109/IJCNN.2017.7965918.

[23] S. Tiwari, R. Sabzehgar, M. Rasouli, Short term solar irradiance forecast based on image processing and cloud motion detection, 2019, pp. 1–6, http://dx.doi.org/10.1109/TPEC.2019.8662134.

[24] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv:1412.6980.
[25] C. Crisosto, M. Hofmann, R. Mubarak, G. Seckmeyer, One-hour prediction of the global solar irradiance from all-sky images using artificial neural networks, Energies 11 (2018) 2906, http://dx.doi.org/10.3390/en11112906.
[26] A. Kumler, A Physics-based Smart Persistence model for Intra-hour forecasting of solar radiation (PSPI) using GHI measurements and a cloud retrieval technique, Sol. Energy 177 (2019) 494–500, http://dx.doi.org/10.1016/j.solener.2018.11.046.
[27] J. Zhang, R. Verschae, S. Nobuhara, J.-F. Lalonde, Deep photovoltaic nowcasting, Sol. Energy 176 (2018) 267–276.
[28] K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, Q. Tian, Accurate medium-range global weather forecasting with 3D neural networks, Nature 619 (7970) (2023) 533–538.
[29] J. Lian, T. Liu, Y. Zhou, Aurora classification in all-sky images via CNN–transformer, Universe 9 (5) (2023) 230.
[30] Y. Nie, X. Li, Q. Paletta, M. Aragon, A. Scott, A. Brandt, Open-source sky image datasets for solar forecasting with deep learning: A comprehensive survey, Renew. Sustain. Energy Rev. 189 (2024) 113977.
[31] S. L., C.S. George, Computer Vision, Prentice Books, Upper Saddle River, 2001.
[32] G. Farnebäck, Two-frame motion estimation based on polynomial expansion, in: Proc. 13th Scandinavian Conference on Image Analysis, Springer, 2003, pp. 363–370.
[33] S. Chow, E. Lee, D. Li, Short-term prediction of photovoltaic energy generation by intelligent approach, Energy Build. 55 (2012) 660–667, http://dx.doi.org/10.1016/j.enbuild.2012.08.011.
[34] C.-L. Fu, H.-Y. Cheng, Predicting solar irradiance with all-sky image features via regression, Sol. Energy 97 (2013) 537–550, http://dx.doi.org/10.1016/j.solener.2013.09.016.
[35] L. Vallance, B. Charbonnier, N. Paul, S. Dubost, P. Blanc, Towards a standardized procedure to assess solar forecast accuracy: A new ramp and time alignment metric, Sol. Energy 150 (2017) 408–422, http://dx.doi.org/10.1016/j.solener.2017.04.064.
[36] E. Bristol, Swinging door trending: Adaptive trend recording? in: ISA National Conference Proceedings, 1990.
[37] F. Diebold, R. Mariano, Comparing predictive accuracy, J. Bus. Econom. Statist. 20 (2002) 134–144, http://dx.doi.org/10.1080/07350015.1995.10524599.
[38] D. Harvey, S. Leybourne, P. Newbold, Testing the equality of prediction mean squared errors, Int. J. Forecast. 13 (2) (1997) 281–291, http://dx.doi.org/10.1016/S0169-2070(96)00719-4, URL https://www.sciencedirect.com/science/article/pii/S0169207096007194.
[39] L. Zhang, Deep Learning-Based Hybrid Short-Term Solar Forecast Using Sky Images and Meteorological Data (Ph.D. thesis), University of Nottingham, 2023.
[40] Y. Nie, Q. Paletta, A. Scott, L.M. Pomares, G. Arbod, S. Sgouridis, J. Lasenby, A. Brandt, Sky-image-based solar forecasting using deep learning with multi-location data: training models locally, globally or via transfer learning? 2022, arXiv preprint arXiv:2211.02108.
[41] Y. Nie, A.S. Zamzam, A. Brandt, Resampling and data augmentation for short-term PV output prediction based on an imbalanced sky images dataset using convolutional neural networks, Sol. Energy 224 (2021) 341–354.
[42] Q. Paletta, A. Hu, G. Arbod, P. Blanc, J. Lasenby, SPIN: Simplifying polar invariance for neural networks application to vision-based irradiance forecasting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5182–5191.