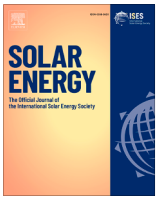
Solar Energy 313 (2026) 114622



Contents lists available at ScienceDirect

Solar Energy

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





A comprehensive study of the all-sky image and solar irradiance dataset SolarVision Almería

Yann Fabel[a,*], Niklas Blum[a], Bijan Nouri[a], Sergio Gonzalez Rodriguez[a], Stefan Wilbert[a], Thomas Schmidt[b], Ole Johannsen[c], Luis F. Zarzalejo[d], Julia Kowalski[e], Robert Pitz-Paal[a]

[a] German Aerospace Center (DLR), Institute of Solar Research, Calle Doctor Carracido 44, Almería, 04005, Andalucía, Spain
[b] German Aerospace Center (DLR), Institute of Networked Energy Systems, Carl-von-Ossietzky-Straße 15, Oldenburg, 26129, Niedersachsen, Germany
[c] Deutsches Krebsforschungszentrum (DKFZ), Applied Computer Vision Lab, Im Neuenheimer Feld 280, Heidelberg, 69120, Baden-Württemberg, Germany
[d] CIEMAT Energy Department, Renewable Energy Division, Avenida Complutense 40, Madrid, 28040, Madrid, Spain
[e] RWTH Aachen University, Chair of Methods for Model-based Development in Computational Engineering, Eilfschornsteinstraße 18, Aachen, 52062, Nordrhein-Westfalen, Germany

HIGHLIGHTS

- Two-year multi-camera all-sky images from southern Spain with high-quality irradiance (GHI, DNI, DHI) and weather data.
- Data preprocessing and quality control framework with automated anomaly detection for sky images.
- Dataset analysis, including evaluation of exposure settings and their impact on deep learning-based solar estimation.

ARTICLE INFO

Keywords:
All-sky imager
Solar irradiance
Sky image dataset
Ground-based observations
Data quality control
Sky image preprocessing
Sky image anomaly detection

ABSTRACT

Sky imagery has become an important resource in solar energy research, weather modeling and atmospheric science. The high temporal and spatial resolution enables detailed analysis of short-term, localized cloud dynamics and accurate intra-hour solar forecasting. However, publicly available datasets combining high-quality sky images with meteorological measurements remain scarce. In this work, we present a new dataset of all-sky images collected over a two-year period at a research facility in southern Spain. Sky image data are recorded in 1-minute intervals and accompanied by meteorological measurements, including global, direct, and diffuse solar irradiance (GHI, DNI, DHI) from ISO 9060:2018 Class A radiometers. Its key feature is the inclusion of multiple camera models and configurations operated in parallel, enabling direct comparisons of image hardware and exposure settings. To ensure high data quality, we implement quality control procedures, including a novel deep learning-based approach to detect anomalies in sky images that achieves over 90% precision and 78% recall. An in-depth data analysis highlights the key characteristics of this dataset, including differences in pixel saturation across camera configurations. We also provide an open-source Python package for sky image preprocessing, including geometric calibration, masking, undistortion, and merging of exposure series. In a case study for solar irradiance estimation, we demonstrate the dataset's utility by investigating the influence of exposure settings. Results indicate that lower exposures (80-160 µs) are generally more suitable for training deep learning models to infer solar irradiance from sky images. The dataset is openly accessible via the PANGAEA data publisher (DOI: 10.1594/PANGAEA.980067).

1. Introduction

In recent years, the demand for sky images and irradiance data has grown significantly, largely driven by the need for sustainable energy solutions. Such high-resolution data form a critical foundation for studying cloud dynamics on local and short-term scales, with applications in weather modeling, climate analysis, and solar energy. The latter has emerged as a cornerstone of the renewable energy transition, with

* Corresponding author.
Email address: yann.fabel@dlr.de (Y. Fabel).

https://doi.org/10.1016/j.solener.2026.114622
Received 15 June 2025; Received in revised form 8 April 2026; Accepted 9 April 2026
Available online 25 April 2026
0038-092X/© 2026 The Authors. Published by Elsevier Ltd on behalf of International Solar Energy Society. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Table 1
Comparison of selected publicly available all-sky image datasets.

Dataset	Duration	Temp. Res.	Image Res.	Irradiance	Imaging Mode	Comments
SIRTA [15]	8 yr	1 min	480×640 768×1024	GHI, DNI, DHI	LDR	Two sky imagers; long-term atmospheric observatory in France
UCSD Folsom [16]	3 yr	1 min	1536×1536	GHI, DNI, DHI	LDR	Widely used benchmark dataset for solar forecasting
Girasol [17]	244 d	15 s	80×60 (IR) 450×450 (LDR/HDR)	GHI	LDR, HDR, IR	Sun-tracking mount; infrared reveals near-sun cloud features
SkyCam [18]	1 yr	10 s	600×600	GHI	LDR, HDR	Multi-site alpine dataset; exposure series available
SKIPP'D [19]	3 yr	1 min	2048×2048 64×64	PV power only	LDR	Quality-controlled dataset paired with PV production
Eye2Sky [20]	1 yr	1 min	2112×2048	GHI, DNI, DHI	LDR	29 ASIs in NW Germany; 0.4–100 km spacing; regional forecasting

photovoltaic (PV) systems rapidly integrating into electrical grids worldwide [1]. However, the inherent variability of solar irradiance, which is primarily caused by dynamic cloud patterns, poses a significant challenge to grid stability, reliable power dispatch and efficient energy management [2–4]. To address this, intra-hour solar forecasts can anticipate changing conditions of solar irradiance, leading to optimized power plant operation and improved grid integration [5,6]. Consequently, sky image–based solar forecasting for intra-hour horizons has attracted considerable research interest in recent years [3,7,8].

While several publicly available all-sky imager (ASI) datasets exist, there remains a lack of standardized benchmark datasets that facilitate objective comparisons of model performance. In addition, the trend towards data-driven models in solar forecasting underscores the need for large amounts of high-quality training data [8]. Robust model generalization further requires diverse datasets that span multiple geographic locations and a variety of camera hardware configurations [9].

Existing publicly available datasets are typically provided by national laboratories or research organizations, often as part of continuous data services (e.g., SURFRAD [10] by the NOAA Earth System Research Laboratory (US)) or through targeted data publications by research groups. These datasets aim to support the research community, by enabling reproducibility of scientific results and model benchmarking. A recent survey on open-source datasets has identified 72 sky image datasets for different applications that are suitable for deep learning [11].

Notable examples, highlighting their key characteristics are summarized in Table 1.

Although these datasets represent a solid foundation of open-access ASI data, several limitations persist. First, the number of high-quality datasets with high temporal resolution and coverage of multiple years is limited to a few sites in the world, to our knowledge, none of them are located in the Mediterranean region. Hence, to cover a wide variety of cloud conditions under different climates and latitudes, more datasets are needed. Besides, most existing datasets only include data from one camera setup per measurement site, which makes it impossible to compare different configurations. Often, there is also little to no information provided on metadata and camera calibration. Additionally, solar irradiance measurements are not always or are just partially/indirectly available in these datasets, for example, limited to global horizontal irradiance or PV power output. A further major limitation across existing datasets is the lack of transparency regarding data quality control. For sky image data, very little information is given on how high image quality is assured. The quality however, may vary strongly due to irregular cleaning schedules and differing environmental conditions (e.g., dust or precipitation), leading to variable levels of lens soiling. Finally, preprocessing routines for sky image analysis, camera calibration or transformations are rarely provided or are only minimally addressed.

In this work, we aim to address these challenges and close existing gaps. The remainder of this paper is structured as follows: Section 2 presents the measurement site and data sources, specifying utilized instrumentation and acquisition procedures of the SolarVision Almería dataset [12]. In Section 3.1 we specify quality control measures and image processing specific to sky imagery. Next, we delve into the dataset in more detail in Section 4, presenting a comprehensive analysis of data availability, atmospheric parameters and exposure settings. Finally, in Section 5, we present a case study on the effect of exposure settings on solar irradiance estimation, before concluding our work and providing an outlook in Section 6.

2. Measurement site and data sources

The data presented in this study were collected at a measurement site located in Tabernas, southern Spain (Latitude: 37.0942°N, Longitude: −2.3547°E, 500 m.a.m.s.l.). According to the Köppen climate classification provided by [13], the site falls under the BWk category (cold desert climate). Despite predominantly clear-sky conditions typical of desert environments, the site's proximity to several mountain ranges —including the Sierra Nevada to the west, Sierra Filabres to the north, and Sierra Alhamilla to the south— frequently leads to complex and dynamic cloud formations. These include multiple cloud layers and rapid changes in cloud morphology, making this location both an interesting and challenging environment for solar forecasting, as also discussed in the ASI nowcasting benchmark conducted at PSA in 2019 [14].

The meteorological stations used for data acquisition are part of the Plataforma Solar de Almería (PSA), which is owned and operated by our Spanish partner, the Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT). The primary station, referred to as PVot, hosts all installed sky cameras. Sky imaging was performed using off-the-shelf surveillance cameras capturing RGB images in the visible spectrum. The cameras were configured to take images with fixed or variable exposure times, in the case of high dynamic range (HDR) images. A Mobotix Q26 camera was deployed throughout the entire measurement period, capturing images with a fixed exposure time of $160\,\mu s$. From September 2022 to December 2023, an AXIS M3057-PLVE Mk II camera was also installed, recording HDR images. Data collection with this camera was discontinued after December 2023 due to a decline in image quality caused by deterioration of the acrylic dome. In addition to these two cameras, a Mobotix Q71 camera was installed in September 2022. Initially configured to capture HDR images, its configuration was changed to take exposure series at the beginning of 2023. The dataset contains images that were captured every full minute during daytime, specifically when the solar elevation angle exceeded 5°.

PVot additionally served as the location for irradiance data collection, using ISO 9060:2018 Class A spectrally flat pyranometers and pyrheliometers manufactured by Kipp & Zonen. A Solys2 sun tracker was equipped with a sun sensor, a CHP1 pyrheliometer for direct normal irradiance (DNI), a shadowed CMP21 pyranometer for diffuse horizontal irradiance (DHI) and an unshaded CMP21 pyranometer for global horizontal irradiance (GHI). All radiometers sampled data at a frequency of 1 Hz, which was subsequently recorded as 1-minute means based on the preceding 60 s. The dataset is further enriched with meteorological measurements at 1-minute resolution, obtained from two nearby stations



Fig. 1. Location and layout of the research site. The aerial view shows the research facility (Plataforma Solar de Almería, owned and operated by CIEMAT) with markers indicating the positions of the meteorological stations. Image credit: CIEMAT.

Table 2
Specifications of all-sky camera models and configurations. All images have an 8-bit pixel depth and are stored in JPEG format. Images were recorded at 1-minute intervals during daytime (solar elevation > 5°). The Mobotix Q71 was operated in two distinct configurations: First in automatic HDR mode with variable exposure and subsequently as exposure series with fixed exposure times.

Camera / Config.	Exposure [µs]	Resolution	Acquisition period
Mobotix Q26	160	2112×2048	01/08/2022–31/07/2024
Mobotix Q71 (HDR)	Variable	2880×2880	10/09/2022–31/12/2022
Mobotix Q71 (Series)	80 / 160 / 320	2880×2880	03/02/2023–31/07/2024
AXIS M3057	Variable	2016×2016	10/09/2022–06/12/2023

Table 3
Sensor specifications for irradiance and meteorological measurements. All data are recorded as 1-minute averages over the period August 2022–July 2024.

Sensor	Manufacturer	Model	Parameter	Unit
Pyrheliometer	Kipp & Zonen	CHP1	DNI	$W\,m^{-2}$
Shaded pyranometer	Kipp & Zonen	CMP21	DHI	$W\,m^{-2}$
Pyranometer	Kipp & Zonen	CMP21	GHI	$W\,m^{-2}$
Hygro-thermometer	Thies Clima	1.1005.54.7xx	Temperature	°C
Hygro-thermometer	Thies Clima	1.1005.54.7xx	Rel. humidity	%
Barometer	Setra	278	Amb. pressure	Pa
Wind vane	NRG	200M	Wind direction	deg
Anemometer	Thies Clima	4.3351.00.161	Wind speed	$m\,s^{-1}$

referred to as HP and Kontas. At Kontas, ambient temperature and relative humidity were measured using a hygro-thermo transmitter by Thies Clima, while barometric pressure was recorded using a Setra pressure transducer. Wind speed and direction were measured at HP using a wind vane and transmitter setup by NRG and Thies Clima. An aerial view of the research facility with the positions of the meteorological stations is provided in Fig. 1.

A comprehensive overview of the all-sky imager configurations and specifications is provided in Table 2, while Table 3 summarizes the details of the irradiance and weather sensors. All measurement and sky image data are available at the data publisher PANGAEA [12].

3. Data quality control and preprocessing

This section outlines the procedures implemented to ensure high quality in the acquired dataset and preprocessing techniques, commonly applied in sky imaging.

3.1. Quality control

A critical requirement for data integrity is regular maintenance of sensors and imaging equipment. At our meteorological stations, all instruments, including radiometers and sky cameras, are cleaned on a daily basis during working days. This routine minimizes inaccuracies in sensor readings and helps maintain consistently high image quality. In addition to cleaning and maintenance work, irradiance and meteorological measurements undergo a validation process based on the methodology described in [21]. This approach combines semi-automated screening tests with manual inspection to identify and flag common data issues, including missing values, physically implausible readings, and internal inconsistencies. This procedure also ensures that data that may be affected by special events, like maintenance work or power shortages, is documented.

3.1.1. Anomaly detection in sky images

Despite regular cleaning, maintaining a perfectly clear camera lens is not always possible. Transient events such as condensation, raindrops, or occluding objects can also degrade image quality, regardless of lens cleanliness. To address this, automated anomaly detection methods can be used to identify different anomalies in sky images that may adversely affect downstream applications, like solar forecasting.

Definition and types of anomalies. We define an anomaly as any condition where parts of the sky dome are temporarily occluded by objects in the vicinity of the camera or when incoming light is distorted due to particles or moisture on the camera lens. These anomalies can be either permanent, such as soiling, or transient, like a bird briefly obstructing the lens. Both types are important to detect, though they differ in impact and mitigation strategy. Depending on the degree of soiling, an ASI-based system can still work fine, although performance may degrade under certain conditions that cause strong scattering effects. In contrast, the system will most likely not be able to derive anything meaningful if a bird covers large parts of the camera lens. However, the interference will disappear once the bird leaves. Therefore, anomaly detection serves in two ways. First, to inform the system operator about data quality issues that should be addressed manually (e.g., via cleaning), and second, to signal to the system itself that data may be unreliable, prompting activation of fallback mechanisms.

A list of common types of anomalies in sky imagery includes:

- Soiling (e.g., dust, grease or bird droppings)
- Water drops (from dew or rain)

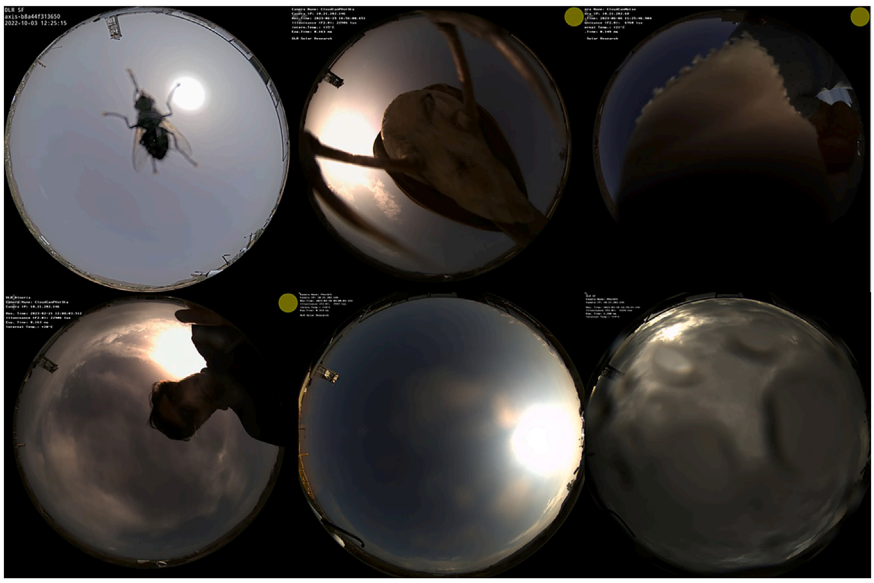


Fig. 2. Examples of common anomalies in ground-based sky imagery. Top row, left to right: arthropod, bird, partially obscured lens during cleaning. Bottom row, left to right: person near the camera, soiling from a previous rain event, water droplets from an ongoing rain event.

- Birds, insects, or spiders sitting on the camera
- Other objects obstructing the camera lens (e.g., leaves)
- People in the field of view (e.g., during maintenance activities)

Static occlusions caused by permanent environmental features, such as buildings or vegetation, are not considered anomalies. However, these should still be accounted for by creating camera masks covering the image parts that do not represent the sky (see Section 3.2.4). Fig. 2 presents examples of common anomalies observed in our dataset.

Although initial studies have explored anomaly detection in sky images, such as detecting raindrops [22] or identifying anomalous patterns via auto-encoders [23], there is no established and standardized method for detecting, categorizing, and localizing a broader range of anomalies in sky imagery. In practice, the identification of these corrupted images, if identified at all, is often performed manually by researchers. For datasets spanning multiple cameras or long time periods this quickly becomes infeasible. In these cases, as well as for automated ASI operation, a different solution is required.

Method overview. To address the problem of unidentified corrupted images, we implemented a fine-tuned object detection model based on YOLO (You Only Look Once) [24]. First, a dataset of anomalous images needs to be compiled through manual screening of camera logbooks, which are continuously maintained for the ASIs at PSA. Each identified anomaly is annotated with bounding boxes to precisely localize the affected regions within the images. Then a pretrained YOLO model is fine-tuned on this annotated dataset. In this work, 390 images were collected in total, drawn from all camera configurations to fine-tune the YOLO model. Hence, various camera models, exposure settings as well as HDR images are covered. Table 4 summarizes the number of labeled instances for each anomaly class, with efforts made to balance the dataset across categories and camera models. Afterwards, this annotated dataset is randomly split into 80% training and 20% validation subsets and model training is conducted over 50 epochs.

Evaluation and results. Evaluation of the anomaly detection model is performed on the validation set comprising 78 manually annotated anomalous images in total. Detection performance is quantified using standard object detection metrics. Precision and recall are employed to assess detection accuracy and are defined as follows:

Table 4
Composition of the anomaly detection dataset by defined classes.

Class	Number of Images
Arthropod	50
Bird	40
Lens covered	40
Person	40
Soiling	70
Water drops	80

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{1}$$

and

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{2}$$

where TP, FP, and FN denote the number of true positives, false positives, and false negatives, respectively. In addition, the *mean Average Precision at an Intersection over Union threshold of 50%* (mAP_{50}) is used to jointly evaluate detection and localization performance. While precision and recall characterize detection accuracy at the image level, the mAP_{50} additionally accounts for the spatial accuracy of the predicted bounding boxes and thus reflects localization performance. More precisely, the mAP_{50} is defined as the mean of the Average Precision (AP) across all anomaly classes C,

$$\text{mAP50} = \frac{1}{C} \sum_{i=1}^{C} AP_i^{50}, \tag{3}$$

where a detection is considered correct if the Intersection over Union (IoU) between the predicted and ground-truth bounding boxes exceeds 0.5. The Average Precision for a given class is computed as the area under the corresponding precision–recall curve,

$$\text{AP} = \sum_n \left(\text{Recall}_n - \text{Recall}_{n-1}\right) \text{Precision}_n, \tag{4}$$

Table 5
Classification and detection performance metrics for the anomaly detection in sky images using a fine-tuned YOLO model.

Class	Instances	Precision	Recall	mAP50
All	87	0.96	0.78	0.91
Arthropods	12	0.97	0.75	0.96
Birds	20	0.99	0.85	0.87
Lens cov.	6	1	0.72	1
People	21	0.9	0.84	0.96
Soiling	13	0.98	0.77	0.86
Water drops	15	0.9	0.73	0.84

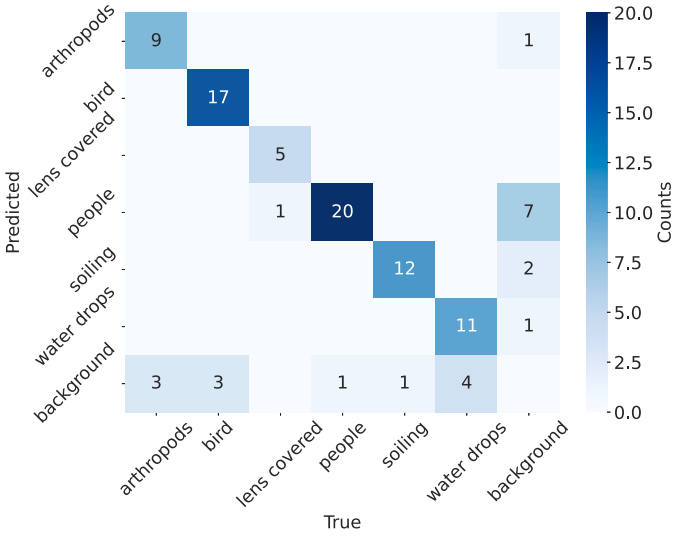


Fig. 3. Confusion matrix for anomaly detection in sky images using a fine-tuned YOLO model.

with each point $(\text{Precision}_n, \text{Recall}_n)$ corresponding to a different detection threshold. The IoU is defined as

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}. \tag{5}$$

As shown in Table 5 the majority of anomalies are correctly detected by the model. For the default prediction threshold of 50%, precision exceeds 90%, indicating a very low false positive rate. Recall values range between 72% and 85%, implying that while some anomalies are missed, most anomalous instances in the validation set are successfully identified.

The confusion matrix shown in Fig. 3 further indicates that most anomaly classes are reliably distinguished by the model, as evidenced by the strong concentration of predictions along the main diagonal. This confirms that the majority of anomalous images are correctly classified into their respective categories. The most frequent misclassifications arise from false positive detections of the people class, where no anomaly is present (background). In addition, for nearly all anomaly classes, a small number of instances remain undetected. These missed detections can be attributed to variations in anomaly prominence. While some anomalies are immediately obvious, others appear only weakly in the image (soiling) are located near the image edges (birds), or are not easily identified due to cloud structures in the background (water droplets).

Overall, these results are encouraging given the limited size of the fine-tuning dataset. Nevertheless, further improvements are expected when incorporating additional anomalous samples in future training iterations.

3.2. Sky image preprocessing

Data preparation is a crucial step in any application involving all-sky imagers. This section outlines a selection of preprocessing techniques commonly used for sky image data. Some methods are particularly relevant for processing individual images when developing and validating models to predict solar irradiance, whereas others, like geometric calibration, are essential when setting up a new all-sky imager. The corresponding Python code is available in the open-source GitHub repository [25].

3.2.1. Data fusion

All-sky images are often complemented by additional observations, such as irradiance measurements, making data fusion a key step. Ideally, all measurement data are perfectly aligned in time and image exposures are triggered and taken exactly at the requested timestamp. In practice however, this is often not the case since the devices are controlled independently. As a result, temporal mismatches can occur between sky images and other measurements, depending on the timing accuracy of each acquisition/measurement system. To address this, it can be reasonable to introduce a tolerance window when aligning data from different sources. The appropriate tolerance depends on the specific application and the temporal resolution of the data. For this dataset, which has a 1-minute resolution, we propose a tolerance of $\pm 10\,\text{s}$ when fusing sky images with meteorological measurements. This is justified for two reasons: first, the acquisition of exposure series can take several seconds; second, using 1-minute images aligns well with the 1-minute averaged meteorological data, ensuring consistent temporal resolution across modalities.

3.2.2. Data filtering

Filtering data based on specific conditions is a common step in data analysis to gain deeper insights. In the context of machine learning, filtering sky images by certain criteria can also increase the information density of a dataset. For example, when training a solar forecasting model, a dataset dominated by clear-sky conditions provides limited value for learning cloud dynamics. By filtering out a large portion of these clear-sky samples, the relative share of more informative cloudy conditions increases, reducing the amount of training data required while improving relevance. Furthermore, such filtering techniques can support model performance evaluations under different conditions. This makes an analysis and comparison across datasets more expressive and interpretable. The DNI variability classification provides a useful basis for such filtering, as it relies solely on DNI measurements and correlates well with varying cloud conditions [26]. Several recent studies have already adopted this classification method for evaluation purposes [27,28]. To illustrate, we filtered the dataset by removing

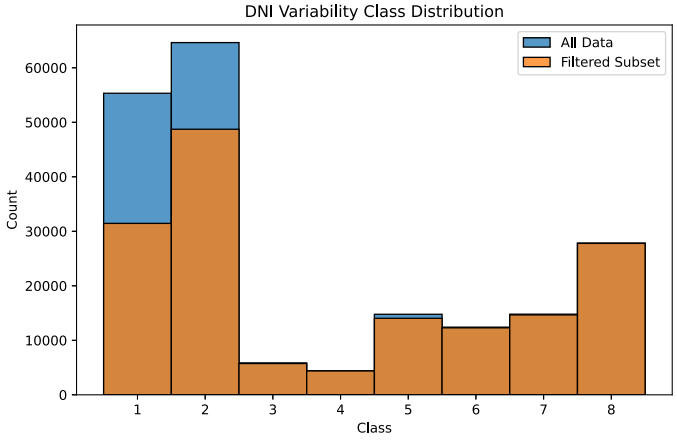


Fig. 4. Distribution of DNI variability classes after filtering out days dominated by clear sky conditions, resulting in a more balanced dataset.

days that are entirely dominated by clear-sky conditions. The resulting dataset is significantly more balanced in terms of DNI variability classes, and therefore cloud conditions, as shown in Fig. 4.

3.2.3. Geometric camera calibration

When observing the sky with all-sky imagers, the position of observed objects, like clouds, needs to be determined from the image. In geometrically calibrated all-sky images, each pixel can be assigned an azimuth and elevation angle. To achieve this, a geometric camera model can be used, which requires calibration of both the intrinsic lens distortion and the external orientation of the camera. A widely used parametric camera model for fisheye lenses was introduced by [29]. While manual calibration using checkerboards was traditionally employed, more recent (semi-)automated approaches have been developed for all-sky imagers, like SuMo and ORION [30,31]. These methods use computer vision techniques to detect orbs, like stars, sun or moon from regular sky images avoiding any manual interference on site. The SuMo method, as implemented in [25], is a recent, fully automated approach for geometrically calibrating all-sky imagers. It has been validated across different camera hardware and locations and can also be applied to sky image data in retrospect. The method detects the sun or moon in all-sky images and compares their observed positions with those calculated from astronomical models. It then iteratively adjusts the calibration parameters to minimize the angular distances between the observed and calculated positions of the orb.

3.2.4. Masking

In sky imagery, masking is another essential preprocessing step to focus analysis on specific parts of the image.

Camera masks. The wide field of view of the fish-eye lenses often captures static objects, such as buildings, vegetation or other instrumentation, which can cover parts of the sky, depending on the camera location and its surroundings. To restrict image analysis to the relevant image areas, it can thus be useful to mask out pixels that do not represent the sky. Such masks can be created manually using a simple graphical user interface or generated automatically through a masking algorithm. We propose a dedicated masking algorithm for sky images, which identifies static objects by averaging a set of images, typically spanning a full day or longer. This temporal averaging suppresses transient elements, such as clouds, while enhancing the visibility of static features near the horizon. The resulting mean image exhibits stronger contrast between the sky and static objects compared to individual frames. To refine this further, a series of contrast enhancement techniques is applied in the following order: Contrast Limited Adaptive Histogram Equalization (CLAHE), gamma correction, and Gaussian blurring to suppress noise. The approximate field of view is then identified using Canny edge detection followed by Hough circle fitting. To strengthen the detected edges, a dilation (maximum filter) is applied, followed by median blurring to smooth the contours. Subsequently, connected component analysis is performed on the inverted edge mask to identify contiguous regions. The final sky region is extracted by selecting the component that includes the center of the image, assuming it represents the unobstructed sky, and converted into a binary mask. Fig. 5 shows a mask created for the Q26 model using the proposed technique.

Elevation and azimuth masks. Camera models described in Section 3.2.3 may be computationally inefficient or not convenient for certain tasks. Therefore, elevation and azimuth angle matrices can be created for each camera instance from a calibrated camera model. For example, a limitation of all-sky imagers is the decreasing so-called horizontal image resolution towards the horizon, which reduces the reliability of extracted information at low elevations [30]. Therefore, in addition to masking static objects, elevation masks are useful for restricting the analysis to specific elevation ranges. For example, when training a deep learning model for cloud classification, the distance between camera and clouds in low-elevation regions often prevents accurate labeling, both for the model and for human annotators. An elevation mask helps avoid such ambiguous areas and can be easily generated using a calibrated camera model as described in Section 3.2.3. Exemplary elevation and azimuth masks of the Q26 camera are shown in Fig. 6.

3.2.5. Undistortion

When observing objects located in a single plane perpendicular to the fisheye camera's optical axis, spatial distortion must be considered. In such cases, the distance between two image pixels does not translate



Fig. 5. Camera mask for the Mobotix Q26 camera model generated automatically using the described procedure.

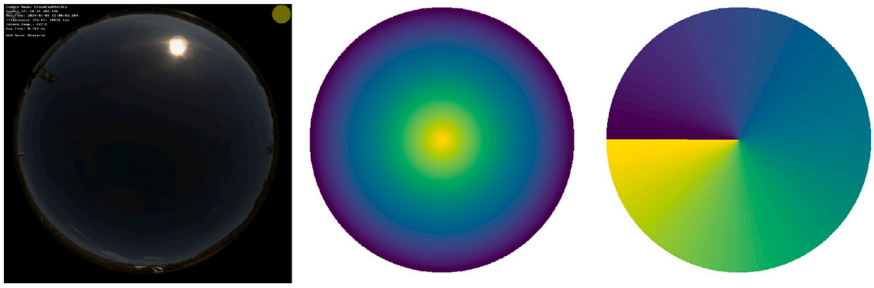


Fig. 6. Sky image of the Q26 camera (left) with corresponding solar elevation (center) and azimuth (right) masks, computed using the geometric camera calibration method from [32].

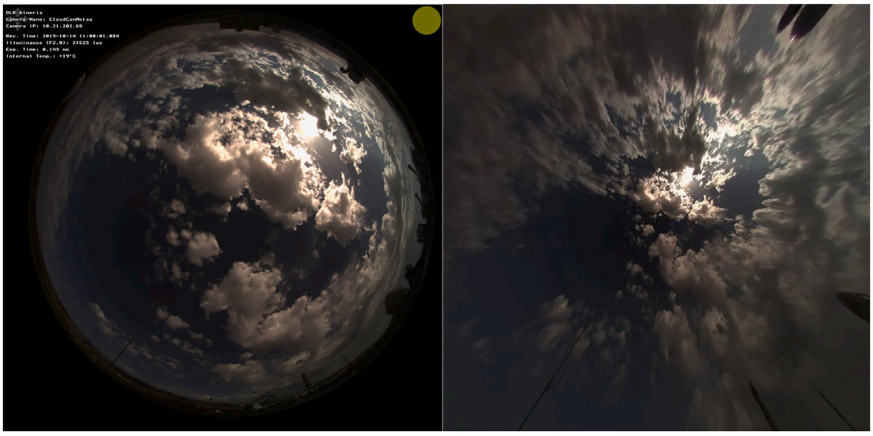


Fig. 7. Exemplary fish-eye sky image before and after applying undistortion. The resulting undistorted image covers zenith angles up to 72°.

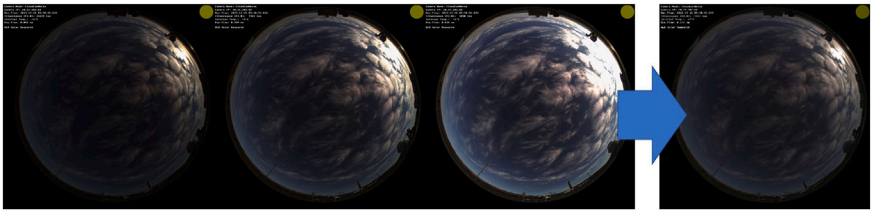


Fig. 8. Image merging of three sky images taken with different exposure times and the resulting HDR image. From left to right: 80µs, 160µs, 320µs, HDR. Note that the HDR image has been tone-mapped and converted for visualization into an 8-bit jpeg file, which reduces the original dynamic range and precision.

uniformly to real-world distances, as the distortion increases with the pixel's distance from the optical axis. As a result, objects near the zenith appear disproportionately large compared to those closer to the horizon, as illustrated in Fig. 7. Undistortion is a common image transformation applied to fisheye images to correct for this lens-induced distortion. It uses the geometric camera model to map distorted image coordinates to undistorted "world coordinates", making pixel distances more proportional to actual distances in the observed scene. This results in a more realistic and spatially consistent representation of the sky, assuming that the observed objects lie in a plane perpendicular to the camera's optical axis.

3.2.6. Merging exposure series

Capturing multiple exposures in quick succession, commonly known as an exposure series, has become increasingly popular in sky imaging, as it helps reduce image saturation caused by bright sunlight and scattering effects. These exposure series can be merged into a single image using dedicated algorithms, each with different goals and characteristics. Two widely used approaches are:

1. **Debevec** [33]: This approach combines multiple exposures into an HDR image of increased bit-depth that is physically consistent. It is especially useful for applications where accurate radiometric information is required, such as irradiance estimation or cloud optical property retrieval.
2. **Mertens** [34]: This method produces low dynamic range (LDR) images that appear visually natural and are well-suited for display or qualitative analysis. However, they are not physically accurate and should be used with caution in quantitative applications.

Depending on the use case, e.g., visualization versus physical modeling, either approach may be preferred. Fig. 8 visualizes the Debevec approach of merging three images of different exposure times.

4. Dataset characteristics

In this section, we analyze the dataset from different perspectives. We begin by examining the availability of both meteorological measurements and sky images from the different camera systems. Next, we derive key atmospheric parameters, such as the Linke turbidity factor and clear-sky irradiance. For each derived and measured atmospheric parameter, we then analyze statistical properties to gain deeper insight into the environmental characteristics captured in the dataset. Finally, we compare camera configurations, particularly the effect of exposure settings on image saturation.

4.1. Data availability

As mentioned in Section 2, sky images are provided exclusively during daytime, defined here as periods with a solar elevation angle greater than 5°. To ensure consistency between the sky imagery and the corresponding meteorological measurements, all sensor data were filtered using the same solar elevation threshold. In addition, we excluded all meteorological data that did not pass the quality control procedures described in Section 3.1, to ensure high data quality, particularly for irradiance measurements.

The availability of sky images from the three ASI models varies depending on their respective installation periods and configurations. In addition to data gaps caused by maintenance work or occasional hardware problems, a total of 1222 images containing identifiable human faces were removed. Most of these images were captured during daily cleaning routines, when personnel were present in front of the cameras. This filtering was performed using the same object detection model (YOLO) [24] as used for anomaly detection but fine-tuned specifically for face detection and applied across the complete sky image dataset.

A graphical overview of sensor and camera availability is shown in Fig. 9. Detailed statistics are provided in Table A.8, which summarizes data availability for both meteorological measurements and sky images by instrumentation and configuration. Meteorological and irradiance

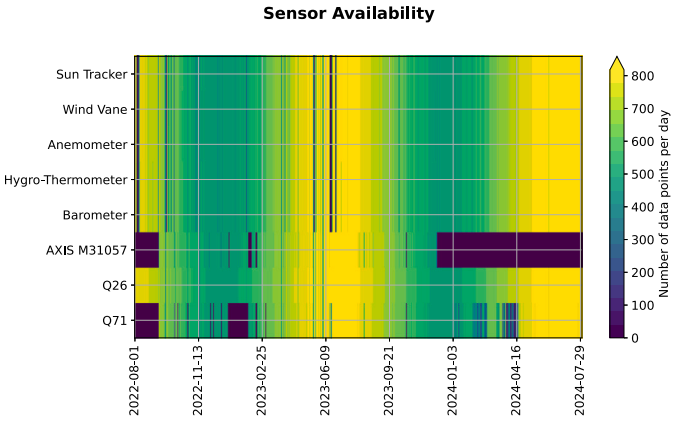


Fig. 9. Data availability across all sensors and all-sky imagers over the full dataset period.

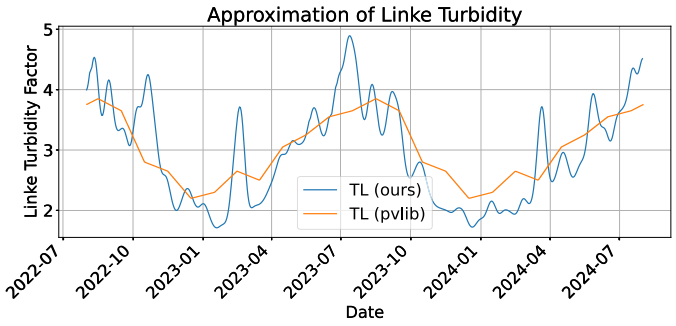


Fig. 10. Daily Linke Turbidity (TL) values compared to the corresponding values from pvlib's lookup table [36]. Both curves exhibit the expected seasonal trend of lower turbidity during winter months. However, only our TL estimates capture short-term fluctuations caused by local atmospheric conditions, such as aerosol events.

data are available for nearly the entire dataset period, totaling approximately 500,000 samples with around 5% missing data during daytime. Sky image availability varies across the different camera models:

- Mobotix Q26: This camera covers the full observation period with the highest image availability among all cameras (1% missing).
- Mobotix Q71: Initially configured to record HDR images for the first 3.5 months from September 2022, it was later reconfigured to record exposure series. Both configurations exhibit slightly reduced availability due to more frequent downtimes (7–8% missing).
- AXIS M3057: This camera provides over a year of HDR image data with approximately 5% missing entries.

Overall, data availability is high across all sensors and cameras with a median of 0 to 8 missing data points per day (see Table A.8). The difference in availability between the camera models is in part explained by the way the cameras were controlled. Only in the case of Q26, a camera-internal backup storage was used. This allowed recovery of images from periods with a network outage but operational backup power via UPS (uninterruptible power supply).

4.2. Atmospheric parameters

In this subsection, we examine both measured and derived atmospheric parameters in more detail. An overview is provided by the histograms in Fig. A.16 and the corresponding summary statistics in Table A.9.

4.2.1. Linke turbidity

Linke turbidity (TL), a measure of atmospheric opacity, is often approximated using large-scale climatological models or look-up tables, which only represent long-term averages. In contrast, our approach accounts for more localized and day-to-day variability by deriving TL values directly from measured DNI under clear-sky conditions, following the formulation by [35].

To identify clear-sky conditions, we first compute TL values for each individual measurement. A measurement is considered applicable if the temporal variability of the TL is low and DNI is sufficiently high. For increased robustness, we apply additional filters, such as a minimum solar elevation threshold, and a maximum TL cutoff, to ensure the reliability of the computed values. We then select the daily minimum TL as a representative approximation for that day. Days without valid TL estimates, typically due to persistent cloud cover, are linearly interpolated using the closest preceding and following valid values. Note, that this method is designed for retrospective analysis and is not suitable for real-time applications. A Python implementation of the procedure is available in the accompanying GitHub repository [25].

The TL histogram in Fig. A.16 reveals a prominent concentration of low TL values around 1.5–3.5, indicative of predominantly clear atmospheric conditions throughout the year. This is particularly pronounced in winter, where TL values often fall below 2, as illustrated in Fig. 10. In contrast, higher TL values during summer months are largely attributed to increased water vapor, given the site's proximity (approximately 30 km) to the Mediterranean Sea. When comparing our TL values to those obtained from the global lookup table provided in pvlib [36], we observe that while both exhibit similar seasonal trends, our method captures local and transient atmospheric variations more accurately. For instance, extreme aerosol events, such as Calima (Saharan dust intrusions), are reflected in our TL estimates but are absent in the climatological lookup table. Such an event can be seen, for example, in Fig. 10, where a high peak of Linke Turbidity can be observed in July 2023 that corresponds to a high level of aerosols as represented in the center image of Fig. 11.

4.2.2. Irradiance

Located in a desert climate, the measurement site typically receives high levels of solar irradiance. To quantify this, we compare it to the clear-sky irradiance, calculated using the same model by [35], as used for Linke turbidity estimation. This calculation was performed using the pvlib implementation and its associated correction functions. The ratio between measured and clear-sky irradiance is referred to as the *clear-sky index*:

$$\text{Clear-Sky-Index}_{\text{GHI}} = \frac{\text{GHI}}{\text{Clear-Sky GHI}} \tag{6}$$

With mean values of 471.2 W/m^2 (GHI) and 558.1 W/m^2 (DNI), compared to 525.9 W/m^2 (clear-sky GHI) and 791.7 W/m^2 (clear-sky DNI), high clear-sky indices are often observed, especially for GHI. A relatively low mean value of DHI (140.5 W/m^2) further supports a predominance of clear sky conditions.

To evaluate the quality of the irradiance measurements, we compare the measured DNI from the pyrheliometer with calculated DNI values, derived from GHI, DHI, and the solar zenith angle θ_Z, using the following closure equation:

$$\text{DNI}_{\text{calc}} = \frac{\text{GHI} - \text{DHI}}{\cos(\theta_Z)} \tag{7}$$

The difference between measured and calculated DNI, the DNI coincidence error, is computed as:

$$\text{Error} = \text{DNI}_{\text{meas}} - \text{DNI}_{\text{calc}} \tag{8}$$

Ideally, measured and calculated DNI values align perfectly. However, small deviations are expected due to sensor inaccuracies and increased

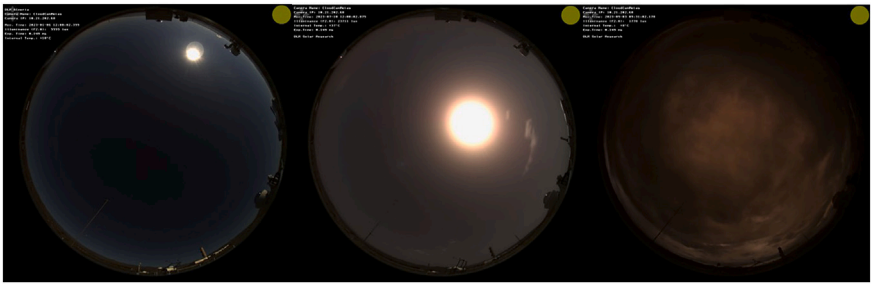


Fig. 11. Examples of sky conditions corresponding to different Linke Turbidity (TL) values. The left image shows a very clear atmosphere in January 2023, whereas the center and right images show the impact of high level of aerosols on clear (July 2023) and cloudy conditions (September 2023).

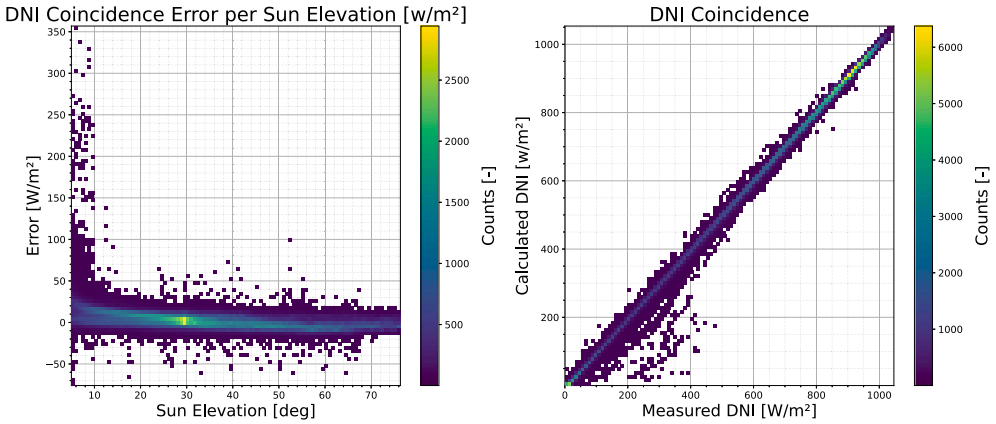


Fig. 12. Comparison of measured and calculated DNI. Left: DNI coincidence error by sun elevation. Right: Scatter plot of measured versus calculated DNI values. Please note that, apart from 35 outlier timestamps, DNI coincidence errors beyond ± 50 W/m^2 were only observed at sun elevation angles below 10° (500 timestamps).

uncertainties at low solar elevations. Fig. 12 provides two visualizations of the DNI coincidence error. The left panel displays the error as a function over solar elevation, while the right panel shows a scatter plot comparing measured and calculated DNI. It can be observed that larger errors are more likely to occur at low solar elevations in the morning and the evening, where DNI is typically low as well. During these times, the calculated DNI often underestimates the measured values, reflecting the increased uncertainty in both measurements and geometric assumptions. In total, 535 data points show deviations larger than 50 W/m^2, with 90% of these occurring at solar elevation angles below 10°.

4.2.3. DNI variability classes

The classification of DNI variability in this work follows the method proposed by [26]. It combines various meteorological indices from the literature to categorize DNI measurements based on their magnitude and short-term variability. Unlike the original study, which uses a longer time window, we adopt a 15-minute interval to assign a variability class to each timestamp, following the approach of [37], which results in a finer temporal resolution. As described in [26] the resulting classes correlate strongly with optical properties of different cloud genera. Therefore, this classification serves as a good representation of cloudiness in the dataset. The classes range from class 1 (clear-sky conditions) to class 8 (opaque overcast skies). Table 6 summarizes the classification scheme in terms of clear-sky index and irradiance variability. As shown in the histogram in Fig. A.16, clear-sky conditions (classes 1 and 2) dominate the dataset, accounting for nearly 60% of all data points. At the same time, around 15% of the dataset corresponds to class 8, indicating also a significant presence of opaque overcast conditions. The remaining 25% fall into intermediate classes (3–7), representing variable sky conditions. These variable periods make the dataset particularly valuable for research in intra-hour solar forecasting.

Table 6
Overview of DNI variability classes according to [26].

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

4.2.4. Cloud coverage

As an additional analysis, we evaluate cloud coverage using a deep learning-based cloud detection model introduced by [38]. The model was trained solely on sky images from a Mobotix Q25 camera located at a different meteorological station within PSA, which uses the same CMOS sensor as the Mobotix Q26 model. Therefore, we here applied cloud detection exclusively to the Q26 images of our dataset. It is important to note that the cloud coverage reported in this work reflects the fraction of cloudy pixels within the circular field of view of the fisheye image. Due to lens distortion effects, this does not perfectly correspond to the actual cloud fraction in the sky. Furthermore, only image regions corresponding to elevation angles greater than 20° were evaluated to reduce uncertainty in low-resolution image areas near the horizon.

The histogram in Fig. 13 supports previous findings, showing that more than half of the dataset exhibits low cloud coverage (less than 10%). On the other end of the spectrum, nearly 20% of the dataset corresponds to high cloud coverage (greater than 90%), indicative of overcast conditions. However, this category also includes overcast situations with

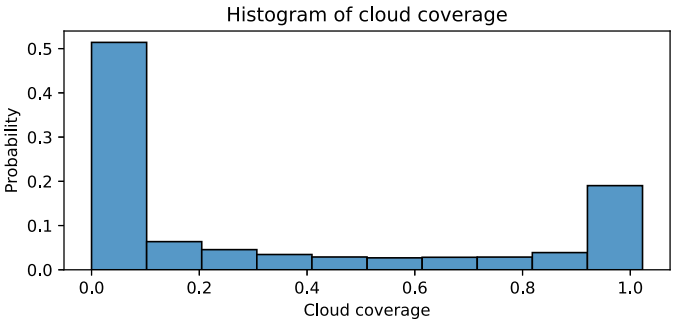


Fig. 13. Distribution of cloud coverage across the dataset. The histogram highlights that over 50% of the sky images show low cloud coverage (<10%), while nearly 20% exhibit high coverage (>90%). Since cloud types are not distinguished, optically thin high-layer clouds are also considered in the calculation of cloud coverage.

optically thin high-layer clouds, as we do not distinguish between cloud types in this analysis.

4.3. Pixel saturation in sky images

Accurate image analysis depends critically on appropriate exposure settings to avoid both over- and underexposure. However, visible-spectrum sky imaging typically employs fixed exposure settings for consistent radiometric reproduction. This makes configuring an optimal exposure particularly challenging due to the extreme brightness variations across the sky. Long exposure times often lead to saturation in regions around the solar disk and adjacent clouds or aerosols, resulting in a loss of detail in these bright areas. On the other hand, image regions further away from the sun disk can become very dark for shorter exposure times, obscuring subtle cloud textures and boundaries. Additionally, optical and electronic imperfections of the camera lens and sensor can distribute the light from a point-wise light source, such as the sun under very clear conditions, over a wider image area via glare effects. As a result, accurately determining the sun's position and the shape or location of clouds close to the sun can become difficult depending on the exposure. This impedes the estimation of solar irradiance, particularly DNI, from such images, as described by [39].

In this section, we analyze image saturation as one aspect of exposure-related information loss. Specifically, we assess the proportion of saturated pixels across various camera configurations and exposure settings. For this analysis we split the dataset into two time periods to account for the change in camera configuration of the Q71 model in early 2023. Saturated pixels are identified by first converting each image to grayscale, then counting all pixels exceeding a fixed intensity threshold. We set the threshold to 240 for all camera models (below the maximum value of 255) to ensure consistency and account for cases where saturation is visually evident but pixel values fall short of the absolute limit.

The saturation proportion is calculated as the number of saturated pixels divided by the total number of sky pixels. In Table 7(a) statistical analysis on saturated pixels corresponding to each camera configuration is listed. As expected, HDR images and images taken with low exposure times exhibit the lowest saturation levels, averaging around 0.4-0.5%. Interestingly, the statistics for HDR images from the AXIS and Q71 cameras are nearly identical, despite their differing hardware and visibly distinct image characteristics. An outlier in the Q71 HDR dataset, causing an unusually high maximum value, is the result of an overexposed image. Among all configurations, the Q71 with an 80 µs exposure time yields the lowest proportion of saturated pixels, slightly less than the HDR images from the AXIS camera. Compared to the Q26 images, the 80 µs images have three times fewer saturated pixels on average. Remarkably, even for the same exposure time of 160 µs, the Q71 shows nearly half the saturation level in comparison to the Q26. This highlights that also the underlying hardware characteristics and firmware have a significant effect on image saturation, independent of exposure settings. Fig. 14 illustrates an example of a scene with cloud enhancement captured using different camera configurations, which demonstrate varying levels of image saturation. In the extreme case of 320 µs exposure, almost half of the sky dome is saturated. But even at 160 µs, substantial information loss occurs. On the other hand, the Q71's 80 µs image exhibits strongly underexposed areas towards the horizon. As previously noted, underexposure also leads to information loss as, for instance, dark noise becomes more dominant. These results clearly emphasize the value of employing exposure series or HDR imaging to capture cloud detail under challenging lighting conditions. When such techniques are not available, relatively short fixed exposure times in the range of approximately 80–160 µs provide a practical compromise between avoiding saturation and preserving cloud structures across the sky dome. Although very short exposures may lead to underexposed regions under optically thick overcast conditions, these situations are typically associated with low irradiance, where detailed cloud structures are less critical.

Table 7
Mean and standard deviation of the percentage of saturated pixels in the sky region for each camera configuration. Due to changes in camera configuration in early 2023, for this analysis the dataset is divided into two time periods to cover all installed configurations.

	2022			2023				
	AXIS	Q26	Q71 HDR	AXIS	Q26	Q71 80	Q71 160	Q71 320
Mean [%]	0.51	1.81	0.52	0.48	1.71	0.40	0.84	3.06
Std [%]	0.28	1.52	0.31	0.26	1.53	0.29	0.87	4.03
Median [%]	0.41	1.17	0.41	0.40	1.06	0.30	0.47	1.27
Max [%]	1.91	14.16	18.47	2.09	19.70	2.84	11.01	42.31

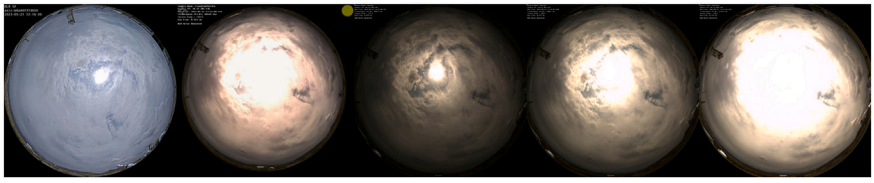


Fig. 14. Comparison of camera configurations with different exposure settings. All images were captured on 2023-05-21 at 13:16 (UTC + 01:00) at the same location. The degree of pixel saturation varies significantly depending on the exposure setting. From left to right: AXIS (HDR), Q26 (160µs), Q71 (80µs), Q71 (160µs), Q71 (320µs).

5. Sample use case: solar estimation from all-sky images

To better understand the impact of different camera configurations on data-driven models, we examine the estimation of solar irradiance from all-sky images in this section, also referred to as *solar estimation* [8]. More generally, solar estimation has been addressed using both physics-based models and data-driven approaches, with recent studies showing that machine learning methods can outperform traditional physical models when suitable input features are available [40]. Given sky imagery as input, the task can be formulated as learning a mapping $f : I_t \rightarrow y_t$, where an image I_t is mapped to the concurrent irradiance measurement y_t (e.g., GHI). Convolutional neural networks (CNNs) are particularly well suited for this task because they can directly learn the complex, nonlinear relationship between spatial cloud patterns in sky images and the resulting irradiance, without requiring explicit modeling assumptions. In contrast, physics-based approaches rely on accurate cloud property retrieval and radiative transfer parameterization, while empirical models such as Ångström–Prescott [41] or Hargreaves–Samani [42] are based on simplified climatological relationships (e.g., sunshine duration or temperature range) and cannot resolve instantaneous, spatially heterogeneous cloud effects [40]. CNNs therefore provide a flexible data-driven framework that leverages the full spatial information content of the images.

This research area has gained growing attention with advances in computer vision, largely because ASIs provide rich spatiotemporal information on cloud patterns that point sensors such as pyranometers cannot capture. In the last decade, several studies have shown that CNNs are effective approach for this task [43,44]. In addition, solar estimation can also be considered a first step towards ASI-based irradiance forecasting [45], and recent generative approaches further extend this concept by predicting future sky images and deriving irradiance from synthetic scenes [46,47].

It should be emphasized, however, that the objective of this case study is not to replace direct irradiance measurements from radiometers, which remain the reference standard for accurate solar resource assessment. Instead, it is intended to systematically analyze how camera configurations and exposure settings influence the behavior and performance of data-driven models trained on sky imagery. In this sense, the presented dataset is designed to serve as a controlled benchmark, enabling a deeper understanding of how deep learning models extract solar-relevant features from complex and highly variable sky conditions.

In this case study, we employ a simple convolutional neural network (CNN) based on the ResNet18 architecture [48] to estimate GHI from a single sky image, without incorporating additional input features. The objective is to compare estimation performance across various camera configurations. Specifically, we evaluate the following five configurations (camera - exposure setting):

- Q26-160µs
- Q71-80µs
- Q71-160µs
- Q71-320µs
- AXIS-HDR

To ensure comparability, a standardized setup is defined across all configurations. The predicted target is the clear-sky index of GHI, which normalizes the measured GHI using the clear-sky irradiance as calculated in Section 4.2.2. This approach forces the model to learn the clouds' impact on irradiance, independent of solar elevation.

We restricted the training, validation, and testing period to February through December 2023, ensuring all analyzed camera configurations were operational. Further, all images undergo the same preprocessing steps: First, camera-specific masks are applied to hide static obstructions. Second, the images are cropped to 128x128 pixels with the zenith in the image center. Third, we apply a data filtering based on the DNI variability classes as discussed in Section 3.2.2 to exclude days that are dominated by clear skies. Each model is trained for 100 epochs with random initialization, using a one-cycle learning rate scheduler [49]. To account for variability stemming from random initialization and the potential convergence to different local minima, we performed five independent training runs for each model. For each training run, the checkpoint with the lowest validation loss is used for evaluation on the test set. The full training setup is summarized in Table B.10.

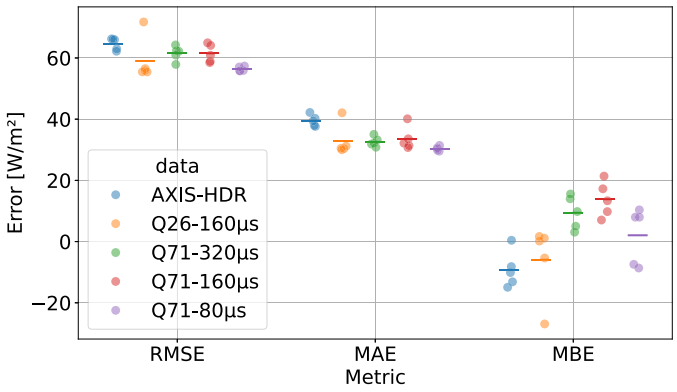


Fig. 15. Performance of solar irradiance estimation across different camera configurations. The strip plot illustrates the distribution of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Bias Error (MBE) in [W/m^2] for GHI estimation on the test dataset. Each dot corresponds to an independent training run for a specific camera configuration with its exposure setting, while the dash represents the arithmetic mean.

Fig. 15 presents the results across all five configurations in a strip plot. The metrics evaluated are Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Bias Error (MBE), all expressed in W/m^2. The dots represent the respective errors of one of the five training runs for each configuration, while the arithmetic mean is indicated by the dash.

Comparing the mean results across camera configurations, the Q71-80 µs, utilizing the lowest exposure time, yielded the best results across all evaluated metrics. However, a closer examination of the spread in error metrics across individual training runs reveals that the Q26-160 µs, configuration often achieved comparable or even slightly superior results, with the exception of one outlier run. This is particularly evident in the MBE, indicating that the Q26-160 µs configuration exhibits the lowest systematic deviation compared to the other configurations. The other two fixed exposure configurations, Q71-160 µs and Q71-320 µs, produced nearly identical error metrics and exhibited very similar spreads across their respective training runs. Conversely, the AXIS-HDR images resulted in the highest errors in this case study, though notably, they showed a relatively low spread across training runs.

These findings suggest that, for the investigated task of solar irradiance estimation, fixed exposures within the lower range (e.g., 80–160 µs) are generally more suitable. However, it is important to note that the AXIS-HDR images used in this dataset were saved as 8-bit JPEGs, implying that their high dynamic range was tone-mapped to an 8-bit resolution. This process, while creating visually enhanced images, inherently alters the natural correlation between image brightness and actual solar irradiance, which may have a negative effect on model training.

Overall, this case study shows that camera configuration significantly influences model learning and must be carefully considered for any target application. While the results presented here serve as a valuable demonstration, they cannot be directly generalized to other applications, such as cloud detection, without further investigation. Future studies could conduct a more extensive exploration of optimal camera configurations for different applications to thoroughly understand the influence of exposure settings. Specifically, incorporating true HDR images with their full range and precision, or comprehensive exposure series, will be crucial when developing new machine learning models based on ASIs.

6. Conclusion

In this work, we introduced a new dataset comprising two years of all-sky images along with ancillary meteorological measurements, including global, direct, and diffuse solar irradiance (GHI, DNI, DHI). The dataset is openly available via the PANGAEA data publishing platform [12]. To our knowledge, this is the first dataset incorporating multiple camera configurations, including fixed and variable exposure, operating simultaneously at the same site, enabling direct comparisons between camera hardware and exposure settings. High data quality was ensured through regular sensor maintenance and continuous quality checks. Further, we presented a novel anomaly detection approach for sky images using a fine-tuned object detection model. Despite being trained on a limited number of labeled samples, the model achieved a precision of over 90% and a recall above 78% on the validation set. However, the impact of these detected anomalous or low-quality images on the respective application's performance remains an open research question and should be addressed in future studies. By conducting an in-depth data analysis, we explored key characteristics of the dataset, including differences in image saturation across camera configurations, a known challenge in visible-light sky imaging. In addition, we compiled and published a comprehensive set of image preprocessing techniques that are essential for working with sky images. These methods, implemented in Python, are available as an open-source package on GitHub [25]. Finally, we demonstrated the practical relevance of the dataset through a use case in solar estimation, showing how different camera configurations affect deep learning models in their predictive accuracy. These insights can guide future decisions on camera selection and configuration for machine learning applications in solar energy.

With this work, we aim to contribute to the growing pool of high-quality, publicly available sky image datasets and encourage others in the community to do the same. Moving forward, the challenges encountered and solutions presented in this work underscore the importance of developing a standardized data format for sky images, including comprehensive metadata, to further enhance interoperability and usability within the research community.

CRediT authorship contribution statement

Yann Fabel: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Niklas Blum:** Writing – review & editing, Software, Methodology, Investigation, Data curation. **Bijan Nouri:** Writing – review & editing, Supervision, Software, Project administration, Methodology, Conceptualization. **Sergio Gonzalez Rodriguez:** Writing – review & editing, Resources, Data curation. **Stefan Wilbert:** Writing – review & editing, Supervision. **Thomas Schmidt:** Writing – review & editing, Software. **Ole Johannsen:** Writing – review & editing, Software. **Luis F. Zarzalejo:** Writing – review & editing, Resources. **Julia Kowalski:** Writing – review & editing, Supervision. **Robert Pitz-Paal:** Writing – review & editing, Supervision, Resources.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to improve readability and language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interests

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

Acknowledgement

A special thanks goes to our technician at the Tabernas site, David Muruve Tejada, for his efforts in the setup, maintenance and cleaning of the instrumentation. This research was partly funded by the German Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection through the AuSeSol-AI project (grant agreement no. 67KI21007A), based on a decision by the German Bundestag.

Appendix A. Supplementary data analysis

Table A.8
Summary of data availability for meteorological sensors and all-sky imagers. The column "Offline (days)" indicates the number of full days with no recorded data during each instrument's respective operational period.

	Instrumentation	Online (days)	Offline (days)	Samples	Missing samples	Fraction missing	Median missing samples per day
Meteo Data	Sun Tracker	724	7	466,790	23,178	0.05	3
	Wind Vane	724	7	466,706	23,262	0.05	3
	Anemometer	724	7	466,718	23,250	0.05	3
	Hygro-thermometer	724	7	466,561	23,407	0.05	3
	Barometer	724	7	466,561	23,407	0.05	3
Image Data	AXIS M3057	444	3	283,389	14,168	0.05	8
	Q26	731	0	485,998	3970	0.01	0
	Q71 (HDR)	110	3	60,804	4862	0.07	7
	Q71 (80µs)	538	7	348,685	28,245	0.07	5
	Q71 (160µs)	538	7	348,441	28,489	0.08	5
	Q71 (320µs)	538	7	348,266	28,664	0.08	6
	Q71 (complete series)	538	7	348,136	28,794	0.08	6

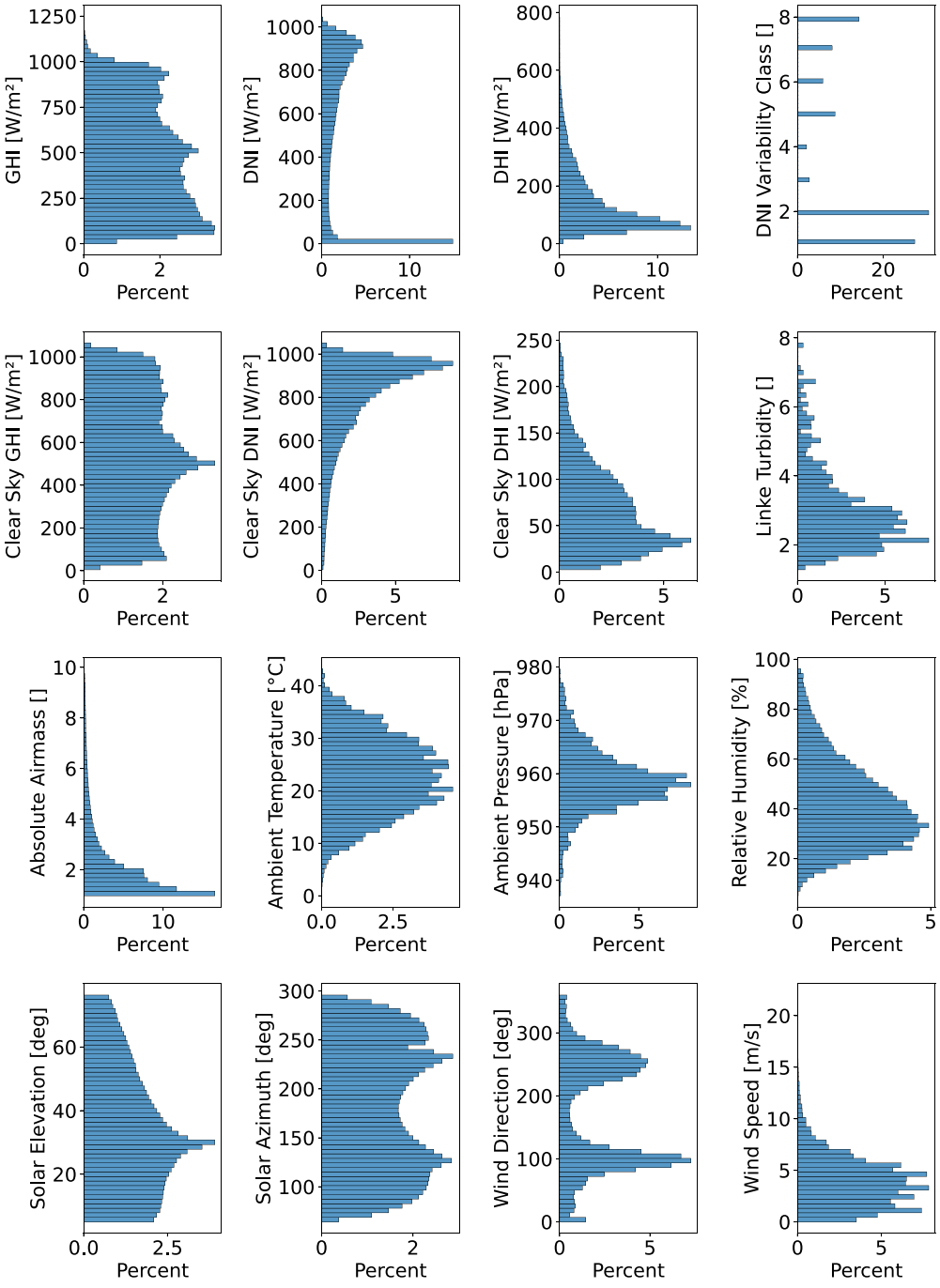


Fig. A.16. Histograms of meteorological measurements and derived quantities over the full two-year dataset period.

Table A.9
Summary statistics of the measured and derived parameters in the dataset, including minimum, maximum, mean, and standard deviation over the full 2-year period.

Parameter	Min	Max	Mean	Std
GHI [W/m^2]	0.1	1245.9	471.2	286.5
DNI [W/m^2]	0	1046.4	558.1	344.9
DHI [W/m^2]	0	784.5	140.5	104.3
Temperature [°C]	0	43.2	22.9	7.2
Amb. Pressure [hPa]	937.1	979.6	958.8	5.7
Relativ Humidity [%]	5.1	96.1	41.4	16.5
Wind Direction [deg]	0	360	169.6	52.9
Wind Speed [m/s]	0	22	3.9	2.4
Solar Elevation [deg]	5	76.3	34.7	18.3
Solar Azimuth [deg]	64.2	295.9	179.7	63.5
Abs. Airmass	1	9.9	2.4	1.7
Linke Turbidity	1.3	7.9	3	1.2
Clear Sky GHI [W/m^2]	4.7	1064.7	525.9	278.2
Clear Sky DNI [W/m^2]	2.9	1051	791.7	201.6
Clear Sky DHI [W/m^2]	2.5	247	68.6	45.8
DNI Coincidence Error [W/m^2]	−76.0	356.9	2.1	8.8

Appendix B. Hyperparameters for the solar irradiance estimation case study

Table B.10
Overview of the hyperparameters and settings used to train the solar estimation model. The training, validation, and test datasets are partitioned by full calendar days from the year 2023.

Hyperparameter	Value
Model Architecture	ResNet18 [48]
Optimizer	AdamW [50]
Loss Function	Mean-Squared-Error (MSE)
Max. Learning Rate (LR)	0.0001
LR Scheduler	One-Cycle-LR [49]
Weight Decay	0.01
Epochs	100
Batch Size	512
Image Resize	128×128
Input Normalization (mean, std)	0.5, 0.5
Data Augmentations	Horizontal Flip (p = 0.75) Rotation (0°–360°, p = 0.75)
Training split	171 days
Validation split	43 days
Test split	25 days

Data availability

The SolarVision Almería dataset is openly available via PANGAEA at https://doi.org/10.1594/PANGAEA.980067. The source code of the presented tools is available at https://github.com/DLR-SF/Sky-Imaging.

References

[1] G. Masson, A. Van Rechem, M. de l'Epine, A. Jäger-Waldau, Snapshot of Global PV Markets 2025, 2025, https://doi.org/10.69766/pbhv9141

[2] R. Perez, M. David, T.E. Hoff, M. Jamaly, S. Kivalov, J. Kleissl, P. Lauret, M. Perez, Spatial and temporal variability of solar energy, Found. Trends Renew. Energy 1 (1) (2016) 1–44, https://doi.org/10.1561/2700000006

[3] Y. Chu, M. Li, C.F. Coimbra, D. Feng, H. Wang, Intra-hour irradiance forecasting techniques for solar power integration: a review, iScience 24 (10) (2021) 103136, https://doi.org/10.1016/j.isci.2021.103136

[4] M. Shafiullah, S.D. Ahmed, F.A. Al-Sulaiman, Grid integration challenges and solution strategies for solar PV systems: a review, IEEE Access 10 (2022) 52233–52257, https://doi.org/10.1109/access.2022.3174555

[5] S.R. West, D. Rowe, S. Sayeef, A. Berry, Short-term irradiance forecasting using skycams: motivation and development, Sol. Energy 110 (2014) 188–207, https://doi.org/10.1016/j.solener.2014.08.038

[6] W. M. O. (WMO), 2022 State of Climate Services: Energy, Tech. rep.accessed on, World Meteorological Organization, Geneva, Switzerland, 2022, https://library.wmo.int/records/item/58116-2022-state-of-climate-services#.Y8lKouzMLjC (12 June 2025).

[7] F. Lin, Y. Zhang, J. Wang, Recent advances in intra-hour solar forecasting: a review of ground-based sky image methods, Int. J. Forecast. 39 (1) (2023) 244–265, https://doi.org/10.1016/j.ijforecast.2021.11.002

[8] Q. Paletta, G. Terrén-Serrano, Y. Nie, B. Li, J. Bieker, W. Zhang, L. Dubus, S. Dev, C. Feng, Advances in solar forecasting: computer vision with deep learning, Adv. Appl. Energy 11 (2023) 100150, https://doi.org/10.1016/j.adapen.2023.100150

[9] 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 heterogeneous multi-location data: dataset fusion versus transfer learning, Appl. Energy 369 (2024) 123467, https://doi.org/10.1016/j.apenergy.2024.123467

[10] NOAA Earth System Research Laboratory, Surface radiation budget (surfrad) network observations, 1995. https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc.

[11] 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, https://doi.org/10.1016/j.rser.2023.113977

[12] Y. Fabel, N. Blum, B. Nouri, S. Wilbert, S.G. Rodriguez, L. Zarzalejo, SolarVision Almería: A Ground-Based Dataset of All-Sky Images, Solar Irradiance, and Meteorological Measurements, PANGAEA, 2025, https://doi.org/10.1594/PANGAEA.980067

[13] H.E. Beck, T.R. McVicar, N. Vergopolan, A. Berg, N.J. Lutsko, A. Dufour, Z. Zeng, X. Jiang, A.I.J.M. van Dijk, D.G. Miralles, High-resolution (1 km) köppen-geiger maps for 1901–2099 based on constrained cmip6 projections, Sci. Data 10 (1) (Oct 2023), https://doi.org/10.1038/s41597-023-02549-6

[14] S.A. Logothetis, V. Salamalikis, S. Wilbert, J. Remund, L.F. Zarzalejo, Y. Xie, B. Nouri, E. Ntavelis, J. Nou, N. Hendrikx, L. Visser, M. Sengupta, M. Po, R. Chauvin, S. Grieu, N. Blum, W. van Sark, A. Kazantzidis, Benchmarking of solar irradiance nowcast performance derived from all-sky imagers, Renew. Energy 199 (2022) 246–261, https://doi.org/10.1016/j.renene.2022.08.127

[15] M. Haeffelin, L. Barthès, O. Bock, C. Boitel, S. Bony, D. Bouniol, H. Chepfer, M. Chiriaco, J. Cuesta, J. Delanoë, P. Drobinski, J.-L. Dufresne, C. Flamant, M. Grall, A. Hodzic, F. Hourdin, F. Lapouge, Y. Lemaître, A. Mathieu, Y. Morille, C. Naud, V. Noël, W. O'Hirok, J. Pelon, C. Pietras, A. Protat, B. Romand, G. Scialom, R. Vautard, Sirta, a ground-based atmospheric observatory for cloud and aerosol research, Ann. Geophys. 23 (2) (2005) 253–275, https://doi.org/10.5194/angeo-23-253-2005

[16] H.T.C. Pedro, D.P. Larson, C.F.M. Coimbra, A comprehensive dataset for the accelerated development and benchmarking of solar forecasting methods, J. Renew. Sustain. Energy 11 (3) (2019), https://doi.org/10.1063/1.5094494

[17] G. Terrén-Serrano, A. Bashir, T. Estrada, M. Martínez-Ramón, Girasol, a sky imaging and global solar irradiance dataset, Data Br. 35 (2021) 106914, https://doi.org/10.1016/j.dib.2021.106914

[18] E. Ntavelis, J. Remund, P. Schmid, Skycam: A dataset of sky images and their irradiance values, 2021, https://doi.org/10.48550/arXiv.2105.02922 (1 May 2021). arXiv:2105.02922.

[19] 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, https://doi.org/10.1016/j.solener.2023.03.043

[20] T. Schmidt, J. Stührenberg, N. Blum, J. Lezaca, A. Hammer, S. Wilbert, B. Nouri, M. Schroedter-Homscheidt, D. Heinemann, T. Vogt, Eye2sky Dataset - All-Sky Images and Meteorological Measurements, Zenodo, 2024, https://doi.org/10.5281/ZENODO.12804613

[21] N. Geuder, F. Wolfertstetter, S. Wilbert, D. Schüler, R. Affolter, B. Kraas, E. Lüpfert, B. Espinar, Screening and flagging of solar irradiation and ancillary meteorological data, Energy Procedia 69 (2015) 1989–1998, https://doi.org/10.1016/j.egypro.2015.03.205

[22] A. Kazantzidis, P. Tzoumanikas, A. Bais, S. Fotopoulos, G. Economou, Cloud detection and classification with the use of whole-sky ground-based images, Atmos. Res. 113 (2012) 80–88, https://doi.org/10.1016/j.atmosres.2012.05.005

[23] M. Krinitskiy, M. Aleksandrova, P. Verezemskaya, S. Gulev, A. Sinitsyn, N. Kovaleva, A. Gavrikov, On the generalization ability of data-driven models in the problem of total cloud cover retrieval, Remote Sens. 13 (2) (2021) 326, https://doi.org/10.3390/rs13020326

[24] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Redmon_You_Only_Look_CVPR_2016_paper.html.

[25] DLR Institute of Solar Research, Sky-Imaging: an open-source Python package for ASI processing, 2025, https://github.com/DLR-SF/sky-imaging (accessed: 2 April 2025).

[26] M. Schroedter-Homscheidt, M. Kosmale, S. Jung, J. Kleissl, Classifying ground-measured 1 minute temporal variability within hourly intervals for direct normal irradiances, Meteorol. Z. 27 (2) (2018) 161–179, https://doi.org/10.1127/metz/2018/0875

[27] B. Nouri, S. Wilbert, N. Blum, Y. Fabel, E. Lorenz, A. Hammer, T. Schmidt, L.F. Zarzalejo, R. Pitz-Paal, Probabilistic solar nowcasting based on all-sky imagers, Sol. Energy 253 (2023) 285–307, https://doi.org/10.1016/j.solener.2023.01.060

[28] Y. Fabel, B. Nouri, S. Wilbert, N. Blum, D. Schnaus, R. Triebel, L.F. Zarzalejo, E. Ugedo, J. Kowalski, R. Pitz-Paal, Combining deep learning and physical models: a benchmark study on all-sky imager-based solar nowcasting systems, Sol. RRL 8 (4) (2024) 2300808, https://doi.org/10.1002/solr.202300808

[29] D. Scaramuzza, A. Martinelli, R. Siegwart, A toolbox for easily calibrating omnidirectional cameras, in: RSJ International Conference on Intelligent Robots and Systems, 2006, pp. 5695–5701, https://doi.org/10.1109/IROS.2006.282372

[30] N.B. Blum, S. Wilbert, B. Nouri, J. Stührenberg, J.E.L. Galeano, T. Schmidt, D. Heinemann, T. Vogt, A. Kazantzidis, R. Pitz-Paal, Analyzing spatial variations of cloud attenuation by a network of all-sky imagers, Remote Sens. 14 (22) (2022), https://doi.org/10.3390/rs14225685

[31] J.C. Antuña-Sánchez, R. Román, J.L. Bosch, C. Toledano, D. Mateos, R. González, V. Cachorro, A. de Frutos, Orion software tool for the geometrical calibration of all-sky cameras, PLOS ONE 17 (3) (2022) e0265959. https://doi.org/10.1371/journal.pone.0265959

[32] N. Blum, P. Matteschk, Y. Fabel, B. Nouri, R. Román, L.F. Zarzalejo, J.C. Antuña-Sánchez, S. Wilbert, Geometric calibration of all-sky cameras using sun and moon positions: a comprehensive analysis, Sol. Energy 295 (2025) 113476, https://doi.org/10.1016/j.solener.2025.113476

[33] P.E. Debevec, J. Malik, Recovering high dynamic range radiance maps from photographs, in: Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '97, SIGGRAPH '97, ACM Press, 1997, pp. 369–378, https://doi.org/10.1145/258734.258884

[34] T. Mertens, J. Kautz, F. Van Reeth, Exposure fusion: a simple and practical alternative to high dynamic range photography, Comput. Graph. Forum 28 (1) (2009) 161–171, https://doi.org/10.1111/j.1467-8659.2008.01171.x

[35] P. Ineichen, R. Perez, A new airmass independent formulation for the linke turbidity coefficient, Sol. Energy 73 (3) (2002) 151–157, https://doi.org/10.1016/S0038-092x(02)00045-2

[36] K.S. Anderson, C.W. Hansen, W.F. Holmgren, A.R. Jensen, M.A. Mikofski, A. Driesse, Pvlib python: 2023 project update, J. Open Source Softw. 8 (92) (2023) 5994, https://doi.org/10.21105/joss.05994

[37] B. Nouri, S. Wilbert, L. Segura, P. Kuhn, N. Hanrieder, A. Kazantzidis, T. Schmidt, L. Zarzalejo, P. Blanc, R. Pitz-Paal, Determination of cloud transmittance for all sky imager based solar nowcasting, Sol. Energy 181 (2019) 251–263, https://doi.org/10.1016/j.solener.2019.02.004

[38] D. Magiera, Semi-Supervised Learning for Probabilistic Cloud Detection in Ground-Based Imagery (Master's thesis), RWTH Aachen, 2024, https://elib.dlr.de/204657/.

[39] N.B. Blum, S. Wilbert, B. Nouri, J. Lezaca, D. Huckebrink, A. Kazantzidis, D. Heinemann, L.F. Zarzalejo, M.J. Jiménez, R. Pitz-Paal, Measurement of diffuse and plane of array irradiance by a combination of a pyranometer and an all-sky imager, Sol. Energy 232 (2022) 232–247, https://doi.org/10.1016/j.solener.2021.11.064

[40] R.A. Ramadhan, Y.R. Heatubun, S.F. Tan, H.-J. Lee, Comparison of physical and machine learning models for estimating solar irradiance and photovoltaic power, Renew. Energy 178 (2021) 1006–1019, https://doi.org/10.1016/j.renene.2021.06.079

[41] M. Paulescu, N. Stefu, D. Calinoiu, E. Paulescu, N. Pop, R. Boata, O. Mares, Ångström–prescott equation: physical basis, empirical models and sensitivity analysis, Renew. Sustain. Energy Rev. 62 (2016) 495–506, https://doi.org/10.1016/j.rser.2016.04.012

[42] Z. Samani, G. Hargreaves, V. Tran, S. Bawazir, Estimating solar radiation from temperature with spatial and temporal calibration, J. Irrig. Drain. Eng. 137 (11) (2011) 692–696, https://doi.org/10.1061/(asce)ir.1943-4774.0000342

[43] 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, https://doi.org/10.1039/c7ee03420b

[44] Y. Nie, Y. Sun, Y. Chen, R. Orsini, A. Brandt, PV power output prediction from sky images using convolutional neural network: the comparison of sky-condition-specific sub-models and an end-to-end model, J. Renew. Sustain. Energy 12 (4) (Jul 2020), https://doi.org/10.1063/5.0014016

[45] S. Dev, F.M. Savoy, Y.H. Lee, S. Winkler, Estimating solar irradiance using sky imagers, Atmos. Meas. Tech. 12 (10) (2019) 5417–5429, https://doi.org/10.5194/amt-12-5417-2019

[46] Y. Nie, E. Zelikman, A. Scott, Q. Paletta, A. Brandt, Skygpt: probabilistic ultra-short-term solar forecasting using synthetic sky images from physics-constrained videogpt, Adv. Appl. Energy 14 (2024) 100172, https://doi.org/10.1016/j.adapen.2024.100172

[47] Y. Fabel, D. Schnaus, B. Nouri, S. Wilbert, N. Blum, L.F. Zarzalejo, J. Kowalski, R. Pitz-Paal, Leveraging generative models for asi-based solar nowcasting, in: EMS Annual Meeting 2024, Barcelona, Spain, 2024, https://elib.dlr.de/208085/.

[48] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015, https://doi.org/10.48550/ARXIV.1512.03385. arXiv:1512.03385.

[49] L.N. Smith, N. Topin, Super-convergence: Very fast training of neural networks using large learning rates, 2018, https://doi.org/10.48550/arXiv.1708.07120. arXiv:1708.07120.

[50] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2019, https://doi.org/10.48550/arXiv.1711.05101.arXiv:1711.05101.