# Solar Irradiance Anticipative Transformer

Thomas M. Mercier
Bournemouth University
tmercier2@gmail.com

Tasmiat Rahman
University of Southampton
t.rahman@soton.ac.uk

Amin Sabet
EscherCloud AI
a.sabet@eschercloud.ai

## Abstract

*This paper proposes an anticipative transformer-based model for short-term solar irradiance forecasting. Given a sequence of sky images, our proposed vision transformer encodes features of consecutive images, feeding into a transformer decoder to predict irradiance values associated with future unseen sky images. We show that our model effectively learns to attend only to relevant features in images in order to forecast irradiance. Moreover, the proposed anticipative transformer captures long-range dependencies between sky images to achieve a forecasting skill of 21.45 % on a 15 minute ahead prediction for a newly introduced dataset of all-sky images when compared to a smart persistence model.*

## 1. Introduction

Solar energy has emerged as one of the most promising alternatives to non-renewable energy sources. As the photovoltaic (PV) industry grows at pace from gigawatt to terawatt scale, the need for more accurate and efficient forecasting of PV output becomes ever more critical. Grid scale solar based power generation poses challenges for grid operators due to the intermittent nature of the supply [2, 23]. Since solar irradiance is a key predictor of PV output, irradiance forecasting on a sub-hour level can greatly support stable and economical power generation. Even forecasting 5 minutes into the future is critical in PV systems to balance storage and load for intermittency as well as having benefits in energy minute by minute trading. The level of solar irradiance seen in a particular location varies based on the cyclical changes of the season, the sun position throughout the day and the weather conditions. While the first two factors are consistently predictable, weather conditions, especially the level of cloud cover make purely time based predictions inaccurate [30].

Two common approaches for short term irradiance forecasting are the use of statistical methods derived from past irradiance measurements and image based forecasts using either ground based sky images or satellite imagery [6].

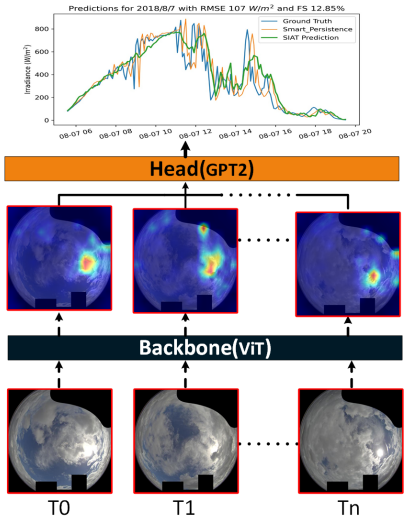



Figure 1. High level overview of model operation. The backbone encodes features from each sky image and the head predicts future irradiance.

Common deep learning (DL) based approaches make use of convolutional neural networks (CNNs) to extract features from images that can then be used to give an associated irradiance value [23]. Ordinarily to predict irradiance, a series of consecutive images are used in either a 3 dimensional CNN or a combination of a CNN and a long short term memory (LSTM) based architecture [26]. This is ultimately based on the temporal information contained in the series of images. In contrast to LSTM based models, transformers offer both the ability to process sequences in parallel as well as excellent modeling of long-term dependencies [21]. The recent application of the self-attention based transformer architecture to computer vision tasks combined with the high performance of transformer-based networks for tasks where long term dependencies are crucial, makes this type of network attractive for solar irradiance forecasting [23, 36]. We propose utilizing a self-attention based backbone network that creates feature representations for each frame in a sequence of all-sky images and then using a Generative Pre-

trained Transformer 2 (GPT-2) based decoder on the resulting sequence of encoded feature vectors to produce solar irradiance predictions [7, 29]. We refer to our model as Solar Irradiance Anticipative Transformer (SIAT). A high level overview of our approach is depicted in Fig. 1.

Our contributions are 1) introduction of a purely attention based forecasting framework that only uses images without any auxiliary data and outperforms previous models on three timestep prediction task, 2) evaluation of our model on three datasets and 3) introduction of three stage training procedure and multiple loss components supervision scheme for strong supervision signal.

## 2. Related Work

Due to the importance of solar irradiance forecasting, many different approaches have been reported in the literature with classical irradiance modelling being based on meteorological input data such as humidity, rainfall and temperature [22]. The general success and the increasingly low barrier of entry to machine learning, it has seen broad adoption in the physical sciences [4, 24]. DL based approaches have become increasingly popular for tackling previously intractable or poorly addressed problems.

For computer vision based irradiance forecasting approaches a range of different datasets are used in literature. Irradiance predictions are commonly made based on a sequence of past sky images, often combined with auxiliary data or input from a classical prediction model. Typically, a dataset consists of a large collection of all-sky images that can be temporally aligned with irradiance values collected at the same site. The National Renewable Energy Research (NREL) dataset was collected in the state of Colorado in the USA and it is publicly available [34]. The newly introduced Chilbolton dataset was collected in a south England based location and is available upon request [31, 33]. The SIRTA dataset was collected by the SIRTA Atmospheric Research Observatory, a meteorological institute near Paris in France and the institute makes the dataset available upon request [15]. The EDF dataset was collected on La Reunion Island and is not publicly available as it was collected by a private company [14].

Le Guen et al. have shown that a time series of all-sky image data in combination with past irradiance data can be used to predict 5 minutes of irradiance data given 5 minutes worth of past data [14]. The images in their dataset were spaced only one minute apart, offering dense temporal information about changes in sky condition. Their dataset was collected in-house at an EDF test site on La Reunion Island. Their model consists of two sub-models, a convolutional LSTM and PhyDNet, which uses partial differential equations for video prediction tasks. The output of the sub-models is combined to produce an irradiance prediction as well as a sky image prediction. They utilise a very large dataset of 6 million images at a size of 80 by 80 pixels and achieve a nRMSE of 23.5 % for a 5 minute ahead irradiance forecast.

Wen et al. show that solar forecasting can be achieved without using a sequential model [37]. They utilise a ResNet18 architecture with the red channel of the past images stacked as input to their network. On the NREL dataset and another California based dataset they report a forecasting skill (FS) up to 17.7 % for a 10 minute ahead prediction compared to a smart persistence (SP) model. Please see Sec. 3.2 for details on how the FS is calculated from the SP model.

Gao and Liu utilise a vision transformer (ViT) to encode the information contained in sky images from two NREL datasets as well as auxiliary meteorological data [11]. A sequence of encoded images and auxiliary information is then fed into another transformer encoder together with a learnable embedding. The output of this encoder is then concatenated with the prediction from a clear sky model and fed into an MLP with residual connections to produce the irradiance forecast. Using one hour worth of past images they report a normalized absolute percentage error of 22.6 % for a one hour forecast.

Paletta et al. present a benchmarking study of different DL based models with the convolutional LSTM giving the best results [26]. All presented models take as an input either a sequence of images or a single image pair, the latter consisting of all-sky images taken at the same time but with different exposure settings. They report a RMSE based FS of 20.4 % for their best model with a SP model used as a comparison. The same group further improved their predictions by implementing the ECLIPSE, a model that has both irradiance and image segmentation as an output [27]. For both studies they utilise a dataset collected and provided by SIRTA laboratory in France which contains all-sky images captured every 2 minutes at 2 different exposure levels [15]. They achieve RMSE of 83.8, 98.5, 109.1 $W/m^2$ which corresponded to a RMSE based FS of 8.7, 23.7 and 24.8 % for 2, 6 and 10 minute ahead irradiance prediction respectively. Since the authors of the ECLIPSE model report that they outperform previous studies we compare our model's performance to this method.

From the presented reports in the literature it is clear that a large variety of prediction approaches exist but that there is a lack of standardisation in the reporting of prediction results as well as in the chosen prediction time horizon. This makes direct performance comparisons difficult. Comparisons are further complicated by the fact that the locations of data collection vary significantly and thus differences in local weather patterns result in datasets of varying difficulty. We will therefore present multiple different performance metrics and evaluate both our model as well as the competing ECLIPSE model on three datasets. Most of the reported

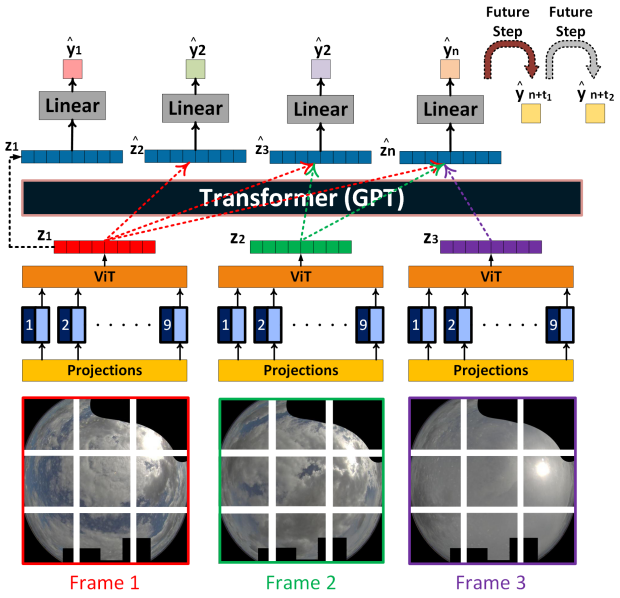



Figure 2. Model flow for SIAT. The ViT backbone encodes the projected flattened image patches into feature vectors $z$ for each image in the input sequence. Together with temporal positional embeddings the feature vectors $z$ are fed into the GPT-2 based decoder. The decoder produces future feature vectors $\hat{z}$ from which irradiance values $\hat{y}$ are produced through a linear layer. For illustrative purposes only 9 image patches and only 3 images are shown as the past context for the model. For visualisation purposes time steps are unfolded; otherwise the same set of weights are used for the ViT backbone and projections to process frames.

works utilize LSTM based networks, which struggle with longer term dependencies. We address this issue by relying on a self-attention mechanism.

## 3. Proposed Framework

Fig. 2 depicts the proposed model architecture. Inspired by anticipative video transformer [13], our model utilises a ViT as a backbone $\mathcal{B}$ which operates on linearly projected flattened image patches $x_t$ to produce an encoding $z_t$ for each of the $s$ input images in the sequence [35]. The input images are split into 16 by 16 patches which are flattened and linearly projected. A class token is prepended to the patch features and a spatial position embedding is added. The output associated with this class token is then used as the image feature representation $z_t$.

$$z_t, z_{t+1}, ... = \mathcal{B}(x_t), \mathcal{B}(x_{t+1}), ... \tag{1}$$

From each input image in the sequence, a feature representation is extracted. Together with temporal positional embeddings, this sequence of features is then used by the GPT-2 based decoder $\mathcal{D}$ [29]. The decoder consists of four layers of masked multi-head attention, a layer norm and a multi layer perceptron (MLP). The decoder produces one $\hat{z}$ for each timestep in the sequence of $s$ images, which are then put through a linear layer $\mathcal{L}$ to produce an irradiance value $\hat{y}_{t+1}$. The linear layer is a fully connected layer with the number of input neurons depending on the dimensionality of the $\hat{z}$ and the output being a single value, representing the predicted irradiance.

$$\hat{z}_{t+1}, ..., \hat{z}_{t+s+1} = \mathcal{D}(z_t, ..., z_{t+s}) \tag{2}$$

$$\hat{y}_{t+1} = \mathcal{L}(\hat{z}_{t+1}) \tag{3}$$

$\hat{z}_{t+1}$ here represents the predicted image features one timestep ahead of the past image feature $z_t$. The masked attention of the GPT-2 decoder ensures that the model can only attend to past features to make the prediction. To predict multiple timesteps into the future, the predicted feature vector is appended to the past context and this is then fed into the head decoder network to predict another step into the future. The proposed framework is both purely attentional in nature and purely image based with no auxiliary data such as past irradiance values, cloud cover or sun location being utilized to make the irradiance predictions. This significantly reduces the requirements for deployment as the equipment needed to collect such auxillary data can present a significant expense.

### 3.1. Training

As Fig. 2 shows, the GPT-2 decoder produces predicted image features which are then fed through a linear layer to give an irradiance prediction. During training of the full model the presented architecture allows for optimization using two loss metrics.

$$L_{irr} = \frac{1}{n} \sum (y_t - \hat{y}_t)^2 \tag{4}$$

$$L_{enc} = \frac{1}{n} \sum (z_t - \hat{z}_t)^2 \tag{5}$$

$L_{irr}$ represents the difference between the predicted irradiance $\hat{y}_t$ and ground truth irradiance $y_t$, and $L_{enc}$ the difference between the encodings $z_t$ and $\hat{z}_t$. $L_{irr}$ can be further separated into the loss associated with the intermediate irradiance predictions and the final prediction, the latter of which is ultimately what is of interest.

The GPT-2 based decoder utilizes masked multi-head attention and hence only attends to encoded features before the time of the prediction. This allows the model to simultaneously predict irradiance values for all input timesteps. Thus an input sequence of four images will produce five irradiance values with all but the first irradiance resulting from the decoder output. If more than one future timestep is to be predicted, the model can be unrolled to predict future image encodings by iteratively adding the intermediate predictions to the past context. Supervising the difference

between $\hat{z}_t$ and $\hat{z}_{t+1}$ ensures that the decoder is able to predict future encoded features. During training the model supervision is based on a weighted sum of both loss $L_{irr}$ and $L_{enc}$ as follows.

$$L_{total} = \alpha L_{irr,f} + \beta L_{irr,i} + \gamma L_{enc} \tag{6}$$

Here, $\alpha$, $\beta$ and $\gamma$ represent the weight of the loss associated with the final ($L_{irr,f}$) and intermediate ($L_{irr,i}$) irradiance predictions and the encoding ($L_{enc}$) prediction, respectively. We train the model in three stages. The first stage consists of training the backbone ViT to map a single image to a single irradiance value. During this training stage the ViT is only supervised by $L_{irr}$. During the second stage of training, the mapping trained ViT model has its regression head removed and the remaining model is frozen and used as the image encoding backbone of the overall architecture. With this frozen backbone, the head GPT-2 based decoder is then trained to predict future encodings which are turned into irradiance value predictions via a linear head. During this and the following stage all loss components are used to supervise the model. In the third stage the backbone model is unfrozen and the model is fine-tuned back to back.

### 3.2. Model evaluation

Since every solar irradiance model ultimately aims to give an accurate prediction of a continuous value, error metrics commonly used for regression tasks such as mean absolute error (MAE), mean squared error (MSE) and RMSE can be employed. However, since irradiance values are strongly weather dependent the mean and variance of a given dataset can vary substantially for different measurement locations. A region with largely clear skies will produce irradiance values that vary smoothly over time and are therefore much easier to predict. A simple difference based error metric would not take the differences in prediction difficulty into account. To improve comparability of model performance on different datasets, evaluation metrics can be normalised by dividing them by the mean of the training irradiance values [14]. A normalised RMSE will be abbreviated by nRMSE. While this improves comparability, it is generally recommended to use a FS metric that compares the error achieved by the presented model to the error achieved by a reference model [39]. An overview of the loss metrics is given below.

$$MAE = \frac{1}{n}\sum |y - \hat{y}| \tag{7}$$

$$RMSE = \sqrt{\frac{1}{n}\sum (y - \hat{y})^2} \tag{8}$$

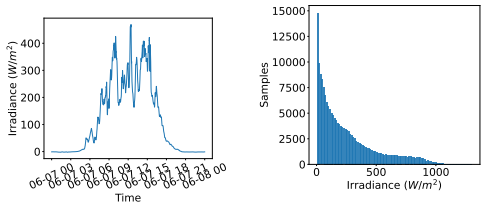
$$FS = 1 - \frac{RMSE_{model}}{RMSE_{reference}} \tag{9}$$

A FS above 0 indicates that the model in question outperforms the reference model. The FS can be calculated based on any loss metric that can be computed for both models. However, the RMSE based FS is the most commonly used metric. Clear sky irradiance and SP are the most commonly used reference models [19, 20, 39]. Clear sky irradiance models use meteorological data such as aerosol optical density and air pressure in combination with the location and time of year to model what the irradiance would be without cloud cover. SP models use the most recent observation in the data as a prediction with the value being adjusted by a clear sky index, as shown in Eq. (10). This index can either be derived from a clear sky irradiance model or be based on measured data.

$$\hat{y}_{t+T} = \frac{y_t}{y_{clear,t}} y_{clear,t+T} \tag{10}$$

Here, $\hat{y}_{t+T}$ represents the predicted irradiance at time $t+T$, $y_t$ the real irradiance value at time $t$ and $y_{clear}$ the clear sky model prediction for time $t+T$. We use the simplified Solis model to calculate the clear sky index needed for the SP reference model [17, 18]. The Solis model requires meteorological data such as air pressure, aerosol optical depth and precipitable water as input, which is sourced from [8–10, 12, 32]. Since the SP model's predictions simply shift the ground truth irradiance by a multiple of the timestep (5 minutes in the case of the Chilbolton dataset) with a small adjustment based on a clear sky index, the SP model's prediction appear to follow the ground truth relatively well for very short term predictions. However, this method still results in a large average error as is illustrated in Fig. 5c. The FS expresses how much a model outperforms this approach.

## 4. Datasets




(a) Unprocessed irradiance. (b) Histogram for irradiance values.

Figure 3. Raw irradiance data for an example day as well as the distributions of values after pre-processing and filtering for the Chilbolton dataset.

In addition to evaluating our SIAT model's performance on two datasets previously used in literature, we introduce the new Chilbolton dataset. In contrast to previously used

datasets the Chilbolton dataset was collected in the challenging weather patterns of the south of the UK. The sky images were provided by the National Centre for Atmospheric Science (NCAS). Both images and irradiance measurements were taken at Chilbolton UK Facility for Atmospheric and Radio Research [31, 33]. The Chilbolton dataset consists of the cloud images and radiometer measurements. A pyranometer collected total global solar irradiance in $W/m^2$ with a temporal resolution of 1 second and about 8000 measurement points per day. Since the images were taken roughly every 5 minutes, the data were preprocessed such that the radiometer data was averaged over a time window of 30 seconds with the resulting value being assigned to one image. The data were aligned based on the timestamps so that the time window for averaging the radiometer data always started at the time stamp of the image. Fig. 3a shows the raw measurement data that were available for a single day. As can be seen, the data varies with time of day but shows strong drops in irradiance related to change in cloud conditions. To exclude very dark images data points taken between midnight and 3 am data points or with irradiance values below 2 $W/m^2$ were removed from the dataset. Additionally images were removed where objects or animals were blocking the view of the camera and where excessive frost or rain blocked the view. The target data distribution is depicted in Fig. 3b. The data was split into a training and evaluation dataset as well as a separate testing dataset. This split was done by using days 5 to 9 of each month as the fixed testing dataset while using days 15 to 19 for evaluation during training. This left 125000 images from the Chilbolton dataset for training. To facilitate comparison to other works we also train and test our model on the the NREL-TSI and SIRTA datasets. The NREL-TSI dataset consists of all-sky images taken every 10 minutes [34]. From the NREL-TSI dataset 106000 images taken between 2015 and 2022 were used for training. The SIRTA dataset contains all-sky images captured every 2 minutes at 2 different exposure levels [15]. To allow for a direct comparison the data from the SIRTA dataset was filtered and split into train, test and evaluation sets as described in [27]. This resulted in 180000 samples being used for model training. The images from all datasets were preprocessed by cropping and resizing them to a size of 224 by 224 pixels. Furthermore, due to the presence of fixed objects in the camera's field of view, a mask of black pixels was applied to the Chilbolton images. All target data were normalised to have a mean of 0 and a standard deviation of 1 using the mean and standard deviation of the training set.

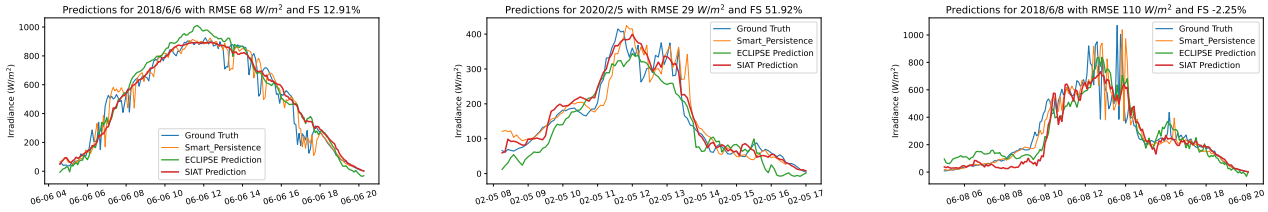
## 5. Implementation

The backbone of the proposed model consisted of a ViT, that had its weights initialized from a model trained on the ImageNet dataset [5, 35, 38]. The backbone was configured to split the 224 by 224 pixel images into 16 by 16 pixel patches which then get flattened and projected to the embedding dimension of 768. The model has a depth of 12 and uses 12 attention heads. To have the model learn to extract task-relevant features, it was trained separately from the full model by having it map single images to the associated irradiance values. To use the backbone in the full model, the final fully connected layer was removed so that the output of the backbone would be the extracted feature vector. To use a transfer learning based approach for the backbone on the SIRTA dataset, it was necessary to add an additional 2d-convolutional layer with a kernel size of 3 and a stride of 1 to the model. Since the SIRTA dataset offers two images with different exposures for each irradiance value and the ViT backbone expects images to have only 3 channels, this convolutional layer takes in the channel-concatenated images and projects them to the required channel number.

The head model was configured to use 8 attention heads, to have a depth of 4 and use an internal encoding dimension of 512. Since the time between images varied between 2 and 10 minutes depending on the dataset in question, the prediction horizon varied accordingly. Randomized image augmentation was applied during all training by varying the brightness, contrast, saturation and hue of the images by 1 % and rotating the images up to 15 degrees. For training the backbone to map images to irradiance values, we use a batch size of 64 and the Adam optimizer with a weight decay of $1 \cdot 10^{-6}$ and a learning rate of $1 \cdot 10^{-4}$ and a scheduled cosine anneal being applied to the learning rate every step. The backbone is trained for 11 epochs. For training and testing of the full model, we use a sequence of 5 sequential images to predict 3 timesteps into the future with the reported RMSE being based on the future prediction for the timestep in question. We use a batch size of 16 and the Adam optimizer with a weight decay of $1 \cdot 10^{-6}$ and a learning rate of $5 \cdot 10^{-5}$ with an exponential learning rate warm up and a scheduled cosine anneal being applied to the learning rate every step. The full network is trained for 11 epochs with the backbone staying frozen for 10 epochs. The network is supervised by using intermediate as well as final $L_{irr}$ and $L_{enc}$ loss components. Empirically we found that an equal weighting of all loss components shown in Eq. (6) gives the best performance. All models were implemented using PyTorch [28] and the code is available at [25]. All training was carried out on a machine equipped with a Nvidia RTX 3090 with 24 GB of memory and a i7-7700K with 64 GB of memory.

## 6. Results and Discussion

We report RMSE, nRMSE, MAE and FS for two different prediction tasks, a single timestep ahead and a three timesteps ahead prediction both using 5 images as the past context. For the newly introduced Chilbolton dataset our




(a) Day with low variability. (b) Day with overall low irradiance values. (c) Day with negative FS.

Figure 4. Comparisons of ground truth irradiance with 15 minute ahead predictions by our SIAT model, the competing ECLIPSE model and the SP approach. The example days are taken from the unseen test set of the Chilbolton dataset.

Table 1. Comparison of our SIAT model to the competing models for all three datasets. All training scenarios use 5 images as past context. The time between images is 2 minutes for SIRTA, 5 minutes for Chilbolton and 10 minutes for NREL-TSI. The future steps indicates how many timesteps into the future the model predicts. While the results for ECLIPSE model are based on unofficial model implementation, due to code unavailability, with a reported RMSE of 83.8 and 98.5 for 1 and 3 timesteps ahead prediction the results here closely match what the authors report in their publication. In addition to the results we computed ourselves we pull further comparisons for the SIRTA dataset directly from the publication [27].

| | | **Future Steps** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1 | | | | 3 | | | |
| **Dataset** | **Model** | **MAE (W/m)** | **RMSE (W/m)** | **nRMSE (%)** | **FS (%)** | **MAE (W/m)** | **RMSE (W/m)** | **nRMSE (%)** | **FS (%)** |
| SIRTA | Smart Persistence | **39.01** | 93.33 | 25.02 | - | 62.10 | 129.77 | 34.78 | - |
| SIRTA | SIAT(ours) | 42.05 | **76.94** | **20.62** | **17.57** | **54.26** | **97.60** | **26.16** | **24.79** |
| SIRTA | ECLIPSE [27] | 48.94 | 78.90 | 21.15 | 15.46 | 57.61 | 98.64 | 26.44 | 23.99 |
| SIRTA | PhyDNet [14, 27] | - | 87.70 | 23.51 | 6.00 | - | 102.00 | 27.34 | 21.10 |
| SIRTA | TimeSFormer [3, 27] | - | 93.10 | 24.95 | 0.20 | - | 105.00 | 28.14 | 18.80 |
| SIRTA | ConvLSTM [26] | - | 95.60 | 25.62 | -2.40 | - | 107.20 | 28.73 | 17.10 |
| Chilbolton | Smart Persistence | **51.96** | 116.09 | 46.28 | - | 73.45 | 142.58 | 56.84 | - |
| Chilbolton | SIAT(ours) | 57.51 | **98.12** | **39.11** | **15.48** | **68.15** | **112.00** | **44.65** | **21.45** |
| Chilbolton | ECLIPSE [27] | 68 | 103.69 | 41.34 | 10.68 | 76.33 | 117.35 | 46.78 | 17.69 |
| NREL-TSI | Smart Persistence | 88.87 | 169.02 | 43.81 | - | 151.72 | 241.04 | 62.48 | - |
| NREL-TSI | SIAT(ours) | **66.20** | 113.71 | 29.47 | 32.73 | **82.08** | **139.71** | **36.21** | **42.04** |
| NREL-TSI | ECLIPSE [27] | 67.17 | **112.91** | **29.27** | **33.20** | 86.63 | 142.60 | 36.96 | 40.84 |

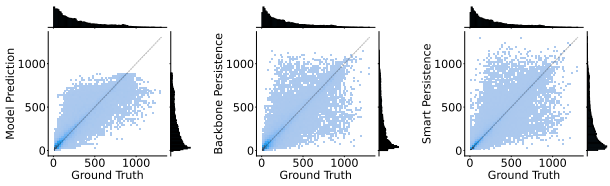
SIAT model achieves an RMSE of 112 $W/m^2$ for the 15 minute or three timesteps ahead prediction. For the same data the SP reference model achieves an RMSE of 142.58 $W/m^2$, this corresponds to an FS of 21.45 %. Tab. 1 gives on overview on how our model performs on different datasets and compared with competing models. As can be seen our model outperforms the competing model on all datasets for the three timestep ahead prediction setting while for the single timestep ahead prediction, we outperform the competing model only on the SIRTA and Chilbolton datasets. While ECLIPSE shows slightly higher FS for this dataset and scenario, the MAE of our model is still better. It is notable that for the very short term prediction of a single timestep for the Chilbolton and SIRTA datasets, the MAE for the SP approach is lower than that of both the ECLIPSE and our SIAT model. The much higher MAE for the SP model on the single timestep ahead prediction for the NREL-TSI can be expected since the time between images is double that of the Chilbolton dataset. Due to the code not being officially available for the competing ECLIPSE model, the presented results are based on an unofficial implementation. For the SIRTA dataset the authors of the ECLIPSE model report an RMSE of 83.8 and 98.5 $W/m^2$ for the one and three timestep predictions, respectively. Using the unofficial implementation we find the RMSE to be 78.9 and 98.64 $W/m^2$ for the same forecasting scenarios. Since these results either beat or match the ones the authors report themselves the implementation can be considered faithful and the computed results for the other two datasets valid. Notably, we also outperform the models for which the results in Tab. 1 were pulled directly from literature. The authors did not report MAE. We attribute the high performance of our model to the transformers ability to use the temporal and spatial information contained in the

series of images used for the prediction.

Comparisons of predictions for the 15 minute ahead case to ground truth for three randomly selected days from the unseen test set of the Chilbolton dataset are shown in Fig. 4. On the example data shown in Fig. 4a, the model outperforms the SP reference model with a FS of 12.91 % for a day with relatively low variability. Here it can also be seen that unlike the ECLIPSE model, our SIAT model avoids the overestimation of the peak irradiance. However, both models fail to predict the sharp rise between 6 and 8am and the dip between 4 and 6 pm.

Fig. 4b shows an example day where the model does very well, reaching a FS of 51.92 %. It successfully predicts the rise in irradiance around 11 am as well as the sharp drop between 1 and 2 pm. While the ECLIPSE model also anticipates the rise, it overall suffers from underestimating the irradiance values as well as anticipating changes that do not occur, as seen in the anticipated dip in irradiance between 3 and 4 pm.

Fig. 4c shows an example day with low irradiance values for most of the day with sharp peaks between noon and 2 pm. While the SIAT model follows the overall shape of the curve it is still results in a negative FS. In this particular case the SIAT model underestimates the irradiance up to 10 am while the ECLIPSE model initially overestimates them. Both models are unable to predict the rise in irradiance between 9 and 10am. As a general observation both the ECLIPSE model and our SIAT model give predictions that result in an overall smoother curve than the ground truth irradiance, however the SIAT model does a better job of predicting changes and of avoiding large over- and under-predictions.




(a) Model predictions. (b) BP predictions. (c) SP predictions.

Figure 5. Density plots comparing predicted irradiance values to ground truth with the grey dashed line representing the ideal case. We show predictions for the SIAT model as well as for both persistence approaches based predictions for 15 minute ahead forecasting task on the Chilbolton dataset. Darker blue indicates higher density. Both persistence approaches yield a much broader distribution compared to our SIAT model.

A comparison of predicted irradiance values and ground truth for all samples in the Chilbolton test set is shown in Fig. 5a. The two-dimensional histogram shows that there is no significant bias in the model's predictions. However, the model avoids predictions of very high irradiance values above 900 $W/m^2$, since these are rare in the data and as can be seen in the examples days in Fig. 4, our SIAT model does sometimes underestimate the irradiance values, especially for very brief peaks in the irradiance curves.

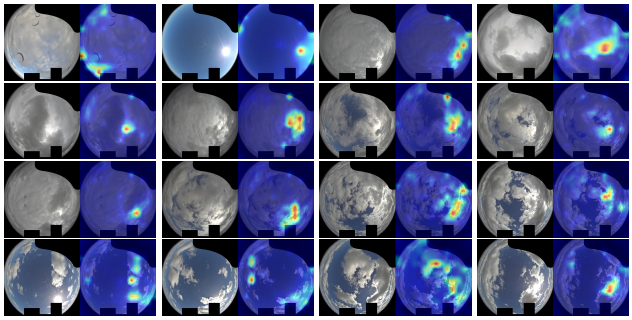


Figure 6. Attention maps for the fine-tuned backbone model for a variety of sky conditions. The overlaid heatmap visualises the areas of the image that the model learns to pay attention to with red signifying higher attention activation and blue signifying low activation. It can be seen that the fine-tuned backbone learns to attend to the areas surrounding the sun and nearby cloud formations. All attention maps were produced using attention rollout [1].

We show that the feature encoder model successfully learns to attend to irradiance relevant features of the sky images as can be seen in the example attention maps shown in Fig. 6. The backbone's attention predominantly falls on the sun as well as cloud formations near it as these are the most important features for making irradiance predictions. Furthermore, the attention maps show that the model is able to extract relevant features under a variety of sky conditions from clear skies with the sun clearly visible to strongly overcast days with the sun barely shining through. The attention maps were produced using attention rollout [1].

In the following paragraphs we show the results of a range of ablation studies analyzing the SIAT architecture using the 15 minute ahead forecasting task on the Chilbolton dataset. To evaluate by how much the GPT-2 based decoder outperforms a simple persistence prediction, we use the ViT backbone with a densely connected layer on top to map each image in the Chilbolton testset to an irradiance value (as is done in the backbone training stage) and shift this value by three timesteps. We refer to this as the backbone persistence (BP) approach. This results in an overall RMSE of 139.58 $W/m^2$ and an MAE of 79.98 $W/m^2$. Since our full model achieves an RMSE of 112 $W/m^2$ and an MAE of 72.35 $W/m^2$, this clearly demonstrates that the GPT-2 based decoder performs much better than a simple persistence model. This is also borne out in Fig. 5b as the BP approach results in a large spread around the ideal.

To gauge the effect that the attention based backbone has on the overall performance of the model, we replaced the

Table 2. Comparison of training the model with a convolution based backbone, a ResNet152, and a transformer based backbone [16]. Evaluation metrics are reported for 15 minute ahead irradiance prediction using the Chilbolton dataset.

| **Backbone** | **RMSE** ($W/m^2$) | **MAE** ($W/m^2$) | **FS (%)** |
|---|---|---|---|
| ResNet152 | 114.28 | 73.03 | 19.85 |
| ViT | **112** | **68.15** | **21.45** |

ViT backbone with a ResNet152 [16, 38], with the comparison being shown in Tab. 2. The three stage training procedure was kept the same with the backbone being trained separately. For the 15 minute ahead forecasting task the model using the ViT based backbone performed better than the ResNet152 on all evaluation metrics, with the FS dropping from 21.45 to 19.85 %.

Table 3. Comparison of model performance with and without the first stage of training where the backbone gets trained to map an irradiance value to a single image. Evaluation metrics are reported for 15 minute ahead irradiance prediction using the Chilbolton dataset.

| **Training stages** | **RMSE** ($W/m^2$) | **MAE** ($W/m^2$) | **FS (%)** |
|---|---|---|---|
| 2 | 113 | 71.75 | 20.75 |
| 3 | **112** | **68.15** | **21.45** |

Since we utilize a three stage training process where the backbone is first trained to map an all-sky image to an irradiance value with the linear head network then being removed to allow the backbone to act as a feature extractor for the all-sky images, we also ran the training of the full model without this first stage of training with the results shown in Tab. 3. As can be seen utilizing a three stage training procedure rather than a two stage procedure boosts performance on all evaluation metrics with the model's FS increasing from 20.75 to 21.45 %.

Table 4. Comparison of results of training the model using either MAE or MSE as the supervision loss function. All results shown are for the 15 minute ahead forecasting task for the Chilbolton dataset.

| **Training Loss** | **RMSE** ($W/m^2$) | **MAE** ($W/m^2$) | **FS (%)** |
|---|---|---|---|
| MAE | 116.02 | **66.41** | 18.63 |
| MSE | **112** | 68.15 | **21.45** |

Tab. 4 shows how the model performs when the MAE loss function is used to supervise the model during training. As can be seen the performance as measured by the overall RMSE and FS suffers with the FS falling from 21.45 % to 18.63 %, however the overall MAE sees some improvement. As Tab. 5 shows, including the encoding loss component in the supervision of the model brings the largest improvement in FS. Notably, including the encoding loss component results in worsening of the MAE metric. While only using the intermediate and final irradiance loss components results in the lowest MAE, the FS is significantly worse. Fig. 7 shows how the FS changes when different number of images are used as past context. For the 15 minute ahead prediction task on the Chilbolton dataset a past context of 5 images is found to be ideal.

Table 5. Evaluation results of supervising the SIAT model using different loss components. All results shown are for the 15 minute ahead forecasting task for the Chilbolton dataset.

| **Loss components** | **RMSE** ($W/m^2$) | **MAE** ($W/m^2$) | **FS (%)** |
|---|---|---|---|
| $L_{irr,f}$ | 113.8 | 70.1 | 20.19 |
| $L_{irr,f} + L_{enc}$ | 113.16 | 72.35 | 20.64 |
| $L_{irr,f} + L_{irr,i}$ | 113.65 | **69.68** | 20.29 |
| $L_{irr,f} + L_{irr,i} + L_{enc}$ | **112** | 72.35 | **21.45** |

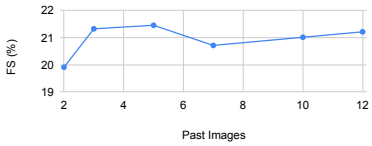



Figure 7. Evaluation results for varying past context lengths while keeping the number of predicted timesteps fixed at 3. Results shown for the Chilbolton dataset.

## 7. Conclusion

We present SIAT, a transformer based framework for the task of forecasting solar irradiance using a sequence of all-sky images without the use of auxiliary data. A ViT backbone serves as a feature extractor to create a feature vector for each frame in the sequence. Our approach then utilizes the temporal relationship contained in the extracted features via a GPT-2 based decoder network. Our training scheme first has the backbone learn to map images to irradiance values to ensure the backbone learns to extract task relevant features. This backbone remains frozen for the first part of the training of the full model. We supervise the model by both its ability to predict future features as well as irradiance values. In the last stage of training the backbone is unfrozen to allow for further fine-tuning of the full architecture. We show that the model successfully learns to attend to important features in the sky images. For the 15 minute ahead forecasting task achieve an RMSE of 112 $W/m^2$ on the Chilbolton dataset, which corresponded to an FS of 21.45 %. For the three timestep prediction we demonstrate that SIAT outperforms competing models on all datasets.

## References


[1] Samira Abnar and Willem Zuidema. Quantifying Attention Flow in Transformers. *arXiv:2005.00928 [cs]*, May 2020. 7

[2] Arthur K. Barnes, Juan C. Balda, and Jonathan K. Hayes. Modelling PV Clouding Effects Using a Semi-Markov Process with Application to Energy Storage. *IFAC Proceedings Volumes*, 47(3):9444–9449, Jan. 2014. 1

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is Space-Time Attention All You Need for Video Understanding? *arXiv:2102.05095 [cs]*, June 2021. 6

[4] Peng Dai, Yasi Wang, Yueqiang Hu, C. H. de Groot, Otto Muskens, Huigao Duan, Huigao Duan, Ruomeng Huang, and Ruomeng Huang. Accurate inverse design of Fabry–Perot-cavity-based color filters far beyond sRGB via a bidirectional artificial neural network. *Photonics Research*, 9(5):B236–B246, May 2021. 2

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. 5

[6] Maimouna Diagne, Mathieu David, Philippe Lauret, John Boland, and Nicolas Schmutz. Review of solar irradiance forecasting methods and a proposition for small-scale insular grids. *Renewable and Sustainable Energy Reviews*, 27:65–76, Nov. 2013. 1

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv:2010.11929 [cs]*, June 2021. 2

[8] T. F. Eck, B. N. Holben, O. Dubovik, A. Smirnov, P. Goloub, H. B. Chen, B. Chatenet, L. Gomes, X.-Y. Zhang, S.-C. Tsay, Q. Ji, D. Giles, and I. Slutsker. Columnar aerosol optical properties at AERONET sites in central eastern Asia and aerosol transport to the tropical mid-Pacific: AEROSOL IN ASIA AND THE MID-PACIFIC. *Journal of Geophysical Research: Atmospheres*, 110(D6):n/a–n/a, Mar. 2005. 4

[9] T. F. Eck, B. N. Holben, J. S. Reid, D. M. Giles, M. A. Rivas, R. P. Singh, S. N. Tripathi, C. J. Bruegge, S. Platnick, G. T. Arnold, N. A. Krotkov, S. A. Carn, A. Sinyuk, O. Dubovik, A. Arola, J. S. Schafer, P. Artaxo, A. Smirnov, H. Chen, and P. Goloub. Fog- and cloud-induced aerosol modification observed by the Aerosol Robotic Network (AERONET): CLOUD-INDUCED AEROSOL MODIFICATION. *Journal of Geophysical Research: Atmospheres*, 117(D7):n/a–n/a, Apr. 2012. 4

[10] T. F. Eck, B. N. Holben, A. Sinyuk, R. T. Pinker, P. Goloub, H. Chen, B. Chatenet, Z. Li, R. P. Singh, S. N. Tripathi, J. S. Reid, D. M. Giles, O. Dubovik, N. T. O'Neill, A. Smirnov, P. Wang, and X. Xia. Climatological aspects of the optical properties of fine/coarse mode aerosol mixtures. *Journal of Geophysical Research*, 115(D19):D19205, Oct. 2010. 4

[11] Huiyu Gao and Miaomiao Liu. Short-term Solar Irradiance Prediction from Sky Images with a Clear Sky Model. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3074–3082, Jan. 2022. 2

[12] David M. Giles, Alexander Sinyuk, Mikhail G. Sorokin, Joel S. Schafer, Alexander Smirnov, Ilya Slutsker, Thomas F. Eck, Brent N. Holben, Jasper R. Lewis, James R. Campbell, Ellsworth J. Welton, Sergey V. Korkin, and Alexei I. Lyapustin. Advancements in the Aerosol Robotic Network (AERONET) Version 3 database – automated near-real-time quality control algorithm with improved cloud screening for Sun photometer aerosol optical depth (AOD) measurements. *Atmospheric Measurement Techniques*, 12(1):169–209, Jan. 2019. 4

[13] Rohit Girdhar and Kristen Grauman. Anticipative Video Transformer. *arXiv:2106.02036 [cs]*, Sept. 2021. 3

[14] Vincent Le Guen and Nicolas Thome. A Deep Physical Model for Solar Irradiance Forecasting With Fisheye Images. *CVPR2020 workshop*, page 4, 2020. 2, 4, 6

[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, and R. Vautard. SIRTA, a ground-based atmospheric observatory for cloud and aerosol research. *Annales Geophysicae*, 23(2):253–275, Feb. 2005. 2, 5

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]*, Dec. 2015. 8

[17] Pierre Ineichen. A broadband simplified version of the Solis clear sky model. *Solar Energy*, 82(8):758–762, Aug. 2008. 4

[18] Pierre Ineichen. Validation of models that estimate the clear sky global and beam solar irradiance. *Solar Energy*, 132:332–344, July 2016. 4

[19] Rich H. Inman, James G. Edson, and Carlos F. M. Coimbra. Impact of local broadband turbidity estimation on forecasting of clear sky direct normal irradiance. *Solar Energy*, 117:125–138, July 2015. 4

[20] Rich H. Inman, Hugo T. C. Pedro, and Carlos F. M. Coimbra. Solar forecasting methods for renewable energy integration. *Progress in Energy and Combustion Science*, 39(6):535–576, Dec. 2013. 4

[21] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in Vision: A Survey. *arXiv:2101.01169 [cs]*, Oct. 2021. 1

[22] Pratima Kumari and Durga Toshniwal. Deep learning models for solar irradiance forecasting: A comprehensive review. *Journal of Cleaner Production*, 318:128566, Oct. 2021. 2

[23] Fan Lin, Yao Zhang, and Jianxue Wang. Recent advances in intra-hour solar forecasting: A review of ground-based sky image methods. *International Journal of Forecasting*, Jan. 2022. 1

[24] Zhaocheng Liu, Dayu Zhu, Sean P. Rodrigues, Kyu-Tae Lee, and Wenshan Cai. Generative Model for the Inverse Design of Metasurfaces. *Nano Letters*, 18(10):6570–6576, Oct. 2018. 2

[25] Thomas M. Mercier. SIAT, https://github.com/Gittingthehubbing/SIAT, Apr. 2023. 5

[26] Quentin Paletta, Guillaume Arbod, and Joan Lasenby. Benchmarking of Deep Learning Irradiance Forecasting Models from Sky Images – an in-depth Analysis. *arXiv:2102.00721 [cs, eess]*, May 2021. 1, 2, 6

[27] Quentin Paletta, Anthony Hu, Guillaume Arbod, and Joan Lasenby. ECLIPSE: Envisioning CLoud Induced Perturbations in Solar Energy. *Applied Energy*, 326:119924, Nov. 2022. 2, 5, 6

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv:1912.01703 [cs, stat]*, Dec. 2019. 5

[29] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. page 24, 2018. 2, 3

[30] Rial A. Rajagukguk, Raden A. A. Ramadhan, and Hyun-Jin Lee. A Review on Deep Learning Models for Forecasting Time Series Data of Solar Irradiance and Photovoltaic Power. *Energies*, 13(24):6623, Jan. 2020. 1

[31] Science and Technology Facilities Council, Chilbolton Facility for Atmospheric and Radio Research, Natural Environment Research Council, and D. Ladd. Chilbolton Facility for Atmospheric and Radio Research (CFARR): Cloud camera 2 imagery from Chilbolton, Hampshire (2016-present). NCAS British Atmospheric Data Centre, 2023/02/24. https://catalogue.ceda.ac.uk/uuid/f55f5649110b4b98b3d5177d8ff2eac9, Oct. 2016. 2, 5

[32] Science and Technology Facilities Council, Chilbolton Facility for Atmospheric and Radio Research, Natural Environment Research Council, and C.L. Wrench. Chilbolton Facility for Atmospheric and Radio Research (CFARR) Meteorological Sensor Data, Chilbolton Site. NCAS British Atmospheric Data Centre, 2023/02/24, https://catalogue.ceda.ac.uk/uuid/45b25a7c531563f4422afcaeea0f07a7, Sept. 2003. 4

[33] Science and Technology Facilities Council, Chilbolton Facility for Atmospheric and Radio Research, and C.L. Wrench. Chilbolton Facility for Atmospheric and Radio Research (CFARR) Visible Radiometer Data. NCAS British Atmospheric Data Centre, 2023/02/24, https://catalogue.ceda.ac.uk/uuid/bf70daf01b6257b2475b057029325869, Sept. 2003. 2, 5

[34] T. Stoffel and A. Andreas. NREL Solar Radiation Research Laboratory (SRRL): Baseline Measurement System (BMS); Golden, Colorado (Data). Technical Report NREL/DA-5500-56488, National Renewable Energy Lab. (NREL), Golden, CO (United States), July 1981. 2, 5

[35] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv:2012.12877 [cs]*, Jan. 2021. 3, 5

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *arXiv:1706.03762 [cs]*, Dec. 2017. 1

[37] Haoran Wen, Yang Du, Xiaoyang Chen, Enggee Lim, Huiqing Wen, Lin Jiang, and Wei Xiang. Deep Learning Based Multistep Solar Forecasting for PV Ramp-Rate Control Using Sky Images. *IEEE Transactions on Industrial Informatics*, 17(2):1397–1406, Feb. 2021. 2

[38] Ross Wightman. PyTorch image models, https://github.com/rwightman/pytorch-image-models. 2019. 5, 8

[39] Dazhi Yang, Stefano Alessandrini, Javier Antonanzas, Fernando Antonanzas-Torres, Viorel Badescu, Hans Georg Beyer, Robert Blaga, John Boland, Jamie M. Bright, Carlos F. M. Coimbra, Mathieu David, Âzeddine Frimane, Christian A. Gueymard, Tao Hong, Merlinde J. Kay, Sven Killinger, Jan Kleissl, Philippe Lauret, Elke Lorenz, Dennis van der Meer, Marius Paulescu, Richard Perez, Oscar Perpiñán-Lamigueiro, Ian Marius Peters, Gordon Reikard, David Renné, Yves-Marie Saint-Drenan, Yong Shuai, Ruben Urraca, Hadrien Verbois, Frank Vignola, Cyril Voyant, and Jie Zhang. Verification of deterministic solar forecasts. *Solar Energy*, 210:20–37, Nov. 2020. 4