



**HAL**  
open science

## Probabilistic solar forecasts of cloud presence as a binary event using a sky camera

Mathieu David, Joaquín Alonso-Montesinos, Josselin Le Gal La Salle,  
Philippe Lauret

### ► To cite this version:

Mathieu David, Joaquín Alonso-Montesinos, Josselin Le Gal La Salle, Philippe Lauret. Probabilistic solar forecasts of cloud presence as a binary event using a sky camera. 2023. hal-04140102

**HAL Id: hal-04140102**

**<https://hal.univ-reunion.fr/hal-04140102>**

Preprint submitted on 24 Jun 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Probabilistic solar forecasts of cloud presence as a binary event using a sky camera

Mathieu David<sup>a,\*</sup>, Joaquín Alonso-Montesinos<sup>b,c</sup>, Josselin Le Gal La Salle<sup>a</sup>,  
Philippe Lauret<sup>a</sup>

<sup>a</sup>*PIMENT, University of La Réunion, 97715 Saint-Denis, Reunion*

<sup>b</sup>*Department of Chemistry and Physics, University of Almería, 04120 Almería, Spain*

<sup>c</sup>*CIESOL, Joint Centre of the University of Almería-CIEMAT, 04120 Almería, Spain*

---

## Abstract

With the fast increase of solar energy plants, high quality short-term forecast is required to smoothly integrate their production in the electricity grids. Usually, forecasting systems predict the future solar energy as a continuous variable. But for particular applications, such as concentrated solar plants with tracking devices, the operator needs to anticipate the achievement of a solar irradiance threshold to start or to stop their system. In this case, binary forecasts are more relevant. Moreover, while most forecasting systems are deterministic, the probabilistic approach provides additional information about their inherent uncertainty that is essential for decision making. The objective of this work is to propose a methodology to generate probabilistic solar forecasts, and more specifically the presence of clouds, as a binary event for very short-term horizons between 1 and 30 minutes.

Among the various techniques developed to predict the solar potential for the next few minutes, sky imagery is one of the most promising. Therefore, we propose in this work to combine a state-of-the-art model based on a sky camera and a discrete choice model to predict the probability of cloud presence. Two well-known parametric discrete choice models, logit and probit models, and a machine learning technique, random forest, were tested to post-process the deterministic forecast derived from sky images. All three models significantly improve the quality of the original deterministic forecast. However, random forest gives the best results and especially provides reliable

---

\*corresponding author

*Email address:* mathieu.david@univ-reunion.fr (Mathieu David)

probability predictions.

*Keywords:* Solar energy, Concentrated Solar Plant (CSP), binary probabilistic forecasts, all sky imager (ASI), Photovoltaic (PV), Brier Score

---

## 1 1. Introduction

2 Solving the challenges posed by the massive integration of solar energy  
3 into electricity grids is a key issue for reducing the carbon footprint of power  
4 generation. Indeed, due to the inherent variability and lack of predictability  
5 of solar energy, a high share of solar energy in the electricity mix makes it  
6 more complicated to manage the supply-demand balance and increases the  
7 vulnerability of the grid. One of the strategies to reduce the effect of solar  
8 variability is to predict future solar irradiance and corresponding solar power  
9 for short-term horizons ranging from 1 minute to several days in advance.  
10 Many techniques have been developed to predict solar irradiance [1, 2, 3].  
11 Numerical weather prediction (NWP) is suitable for horizons longer than 6  
12 hours. Forecasts derived from geostationary meteorological satellite images  
13 are effective for a horizon ranging from about 1 hour to 6 hours. Finally, for  
14 a very short-term horizon of less than 1 hour, the approach based on All Sky  
15 Imagers (ASI) is the most promising technique.

16 Regarding very short-term horizon, Ajith and Martínez-Ramón [4] com-  
17 pared three categories of solar irradiance forecasting methods: time series,  
18 sky camera images and hybrid models combining infrared images with ra-  
19 diation time series. The authors shown that the normalized Root Mean  
20 Square Error (nRMSE) varied from 30 to 53% in terms of forecasting error.  
21 In the literature on ASI, most of the works dealing with the prediction of  
22 solar irradiance or cloudiness propose deterministic forecasts [5, 6, 7]. To  
23 improve forecasts derived from ASI, different approaches have been carried  
24 out. For instance, Paletta et al. [8] evaluated deterministic and probabilistic  
25 predictions based on ASI for different weather conditions (clear, cloudy and  
26 overcast skies). In their work, probabilistic approach demonstrates a richer  
27 operational forecasting framework by facilitating uncertainty quantification  
28 in cloudy conditions and for long-term horizons. However, very few methods  
29 were developed to generate probabilistic forecasts from sky camera images.

30 It is well-known that weather forecasts are uncertain because the evolu-  
31 tion of the weather and consequently solar irradiance are chaotic processes.  
32 Thus, in decision-making operations that use solar forecasts, such as the man-

33 agement of power plants, probabilistic forecasts are crucial. Indeed, prob-  
34 abilistic forecasts assign probability levels to future events and allow their  
35 users to assess the associated risks. Research works dealing with probabilistic  
36 solar forecast are relatively recent but numerous have been released on the  
37 topic in the last 10 years [9, 10, 3]. However, as mentioned previously, very  
38 few works concerning ASI proposed a method to generate probabilistic solar  
39 forecast. This work will contribute to fill this gap in the literature.

40 Usually, solar forecasting systems provide the future level of solar irra-  
41 diance or PV generation as a continuous variable [3]. But for particular  
42 applications, such as the management of Concentrated Solar Plants (CSP)  
43 with tracking devices, the operator needs to anticipate the achievement of  
44 a solar irradiance threshold to start or to stop their system [11]. In this  
45 case, an accurate binary forecast is more relevant. In the wide domains of  
46 meteorology or economy, numerous works propose discrete choice models to  
47 generate binary forecasts [12, 13]. However, in the field of solar energy very  
48 few works propose binary forecasts. One of the rare work on the topic is  
49 proposed by Alonso and Batlles [14] that developed a method to forecast the  
50 cloudiness from a sequence of images given by the MeteoSat Second Gener-  
51 ation (MSG) satellite MSG or an ASI. Their model generates deterministic  
52 binary forecasts of cloud presence (i.e. 0 = cloudy and 1 = clear sky). The  
53 forecasts were tested over 2 years for the city of Almería in the South-East  
54 of Spain. The success rate of the forecasts derived from the ASI is 83% for  
55 the first 15 minutes and drops to 60% for a 3 hours horizon. However, no  
56 works proposes a probabilistic approach to generate discrete solar forecast.

57 In the light of the two main lacks of the literature underlined above, the  
58 main objective of this work is to propose a novel methodology to generate  
59 probabilistic solar forecast as a binary event for horizons ranging from 1 to 30  
60 minutes using an all sky imager (ASI). The developed approach will combine  
61 a state of the art ASI method and discrete choice models proposed in other  
62 domains, such as economy or meteorology. In a first step, a model based on  
63 the detection of cloud motions will use sequences of images from an ASI to  
64 generate binary deterministic forecasts of the cloudiness. Then, in a second  
65 step, binary choice models will be used to convert the deterministic discrete  
66 forecasts into probability levels of cloud presence. Finally, we will assess the  
67 quality of the generated forecast of cloud presence on a case study to evaluate  
68 the added value of the proposed method.

69 The remainder of the paper is organized as follow. Section 2 presents the  
70 methodology used to develop and to evaluate the proposed model. Section

71 3 gives a brief overview of the state of the art model used to generate the  
72 deterministic forecasts. Then, section 4 details the discrete choice models  
73 used to forecast the probability of cloud presence. Section 5 depicts the case  
74 study and the corresponding data. Results are presented and discussed in  
75 section 6. Finally, section 7 gives our concluding remarks.

## 76 2. Overall methodology and forecasts evaluation

77 The probabilistic forecasts of cloud presence are generated in two steps  
78 as presented in figure 3. First, we generated deterministic forecasts of cloud  
79 presence using the method proposed by [14] and briefly presented in section  
80 3. The results are discrete forecasts (1 = no cloud and 0 = presence of  
81 clouds) with a time resolution of 1 minute and horizons of forecast up to 30  
82 minutes. The second step is a post-processing of the deterministic forecasts  
83 with a probabilistic model. In this work, we compared three different mod-  
84 els, described in section 4, to post-process the deterministic forecasts. The  
85 final probabilistic forecasts have the same temporal resolution and the same  
86 horizons as the deterministic forecasts. After these two steps, the generated  
87 forecast are probabilities, in the interval  $[0; 1]$ , that give a level of confidence  
88 or risk associated to the future presence of clouds. Compared to deterministic  
89 forecast, this additional information may help he user for decision-making.

90 Cloud detection was carried out following the methodology presented in  
91 [15] obtaining a cloud identification (clouds which attenuate the DNI below  
92  $400 \text{ Wm}^{-2}$ ) based on the optimal operating value for CSP plants, as the case  
93 of Gemasolar plant, which used this irradiance level, like the appropriate for  
94 producing electricity [14].

95 In this work, both deterministic and probabilistic forecasts will be evalu-  
96 ated. If comprehensive frameworks have been proposed to evaluate forecast  
97 quality of the solar irradiance as a continuous variable [16, 17, 18], no previ-  
98 ous work details the evaluation of discrete solar forecasts. However, specific  
99 error metrics have been designed in the field of meteorology to assess the  
100 quality of binary forecasts. Let us recall that the quality of a forecasting  
101 system evaluates the agreement between the forecasts and the corresponding  
102 observations [19]. Interested readers may refers to the web page published  
103 by the Joint Working Group on Forecast Verification Research to have a  
104 extended overview of weather forecast verification [20].

105 Regarding binary deterministic forecasts, the most common metrics are  
106 derived from the contingency table presented in figure 1. In our case a "yes"

107 event corresponds to a clear sky (no clouds) and  $N$  is the total number of  
 108 observation/forecast pairs used for the verification. The contingency table  
 109 is a useful tool to classify the types of errors. A perfect forecast system  
 110 would generate only hits and correct negatives, and no misses or false alarms.  
 111 Numerous metrics are derived from the four cells in the contingency table,  
 112 such as fraction correct, probability of detection (POD), success rate (SR)  
 113 or false alarm ratio (FAR) [21]. Each metric describes a different aspect of  
 114 forecast performance. In this work we will focus on the accuracy, defined  
 115 in equations 1. Accuracy ranges from 0 to 1, with 1 the perfect score. The  
 116 accuracy also called fraction correct gives the fraction of correct forecasts. It  
 117 is simple and intuitive but, in case of very rare events, this indicator may  
 118 lead to confusion [20].

		Observation	
		yes	no
Forecast	yes	hits	false alarms
	no	misses	correct negatives

Figure 1: Contingency table for a binary forecast

$$Accuracy = \frac{hits + correct\ negatives}{N} \quad (1)$$

119 Regarding the verification of probabilistic forecasts, two main attributes  
 120 define the quality: the reliability and the resolution. Reliability refers to the  
 121 statistical consistency between the forecasts and the observations. In other  
 122 words, the forecast probability should be equal to the observed probability of  
 123 the event (e.g. 20% of the events should happens for a forecast probability of  
 124 20%). The reliability is a crucial prerequisite as non reliable forecasts would  
 125 lead to a systematic bias in subsequent decision-making processes [22]. The  
 126 most used visual tools to assess the reliability is the reliability diagram [23].  
 127 It plots the correspondence between the forecast probability (x axis) and the  
 128 observed frequency of the event (y axis). Perfectly reliable forecasts should  
 129 be as close to the diagonal as possible. Figure 6 shows the reliability diagram

130 for the different models tested in this work. Resolution refers to the ability of  
 131 a forecasting system to generate case-dependent forecasts. For example, the  
 132 climatology model, which predicts the average probability of the event (i.e.  
 133 always the same probability regardless the horizon or the weather conditions)  
 134 has no resolution. Unfortunately, no graphical tool exists to evaluate the  
 135 resolution.

136 Only few metrics, also called scores, exist to quantitatively evaluate the  
 137 quality of probabilistic forecasts of binary events. For this work, we propose  
 138 to use the Brier Score (BS) [19], formulated as follow:

$$BS = \frac{1}{N} \sum_1^N (\hat{p}_i - o_i)^2, \quad (2)$$

139 where  $N$  is number of observation/forecast pairs,  $\hat{p}_i$  the forecast probability  
 140 and  $o_i$  the observation. If the event did occur  $o_i = 1$ , and if it did not  
 141 occur  $o_i = 0$ . The BS measures the mean square probability error. This  
 142 global proper score is appealing because it includes the two basic skills of a  
 143 probabilistic forecast (i.e. reliability and resolution) and it corresponds to  
 144  $1 - Accuracy$  for a deterministic forecast. The BS ranges between 0 and 1  
 145 with 0 the perfect score.

146 Skill scores, derived from the above mentioned metrics are also commonly  
 147 proposed to evaluate forecast quality [16, 17]. Skill scores quantify the im-  
 148 provement of a proposed method compared to a reference model. They are  
 149 relevant for comparing forecasts generated for different sites or time periods.  
 150 We will not provide skill scores in this work because the evaluation will be  
 151 for a unique site and time period. However, the interested reader can use the  
 152 numerical results, given table 1 at the end of this paper, to compute them.

### 153 3. Deterministic forecasts of cloud presence with a sky camera

154 To issue a forecast, a sequence of three consecutive sky camera images,  
 155 spanning about 3 minutes, is used. The correlation between these three im-  
 156 ages makes it possible to establish the behavioural pattern of cloud movement  
 157 at a given time. In order to study cloud movement, the following steps are  
 158 taken [14]:

- 159 • The picture taken with the sky camera is divided into different sectors,  
 160 since the movement of the clouds will depend on the sector covered by  
 161 the sky camera.

- 162 • The cloud motion vector (CMV) is calculated for each sector by apply-  
163 ing the maximum cross-correlation method.
- 164 • Different quality tests are applied to ensure the correct determination  
165 of the cloud motion.

166 The CMV is applied to the last image received and re-applied to the  
167 result obtained. This process is repeated up to 30 times (prediction for 30  
168 minutes ahead), obtaining the movement of the pixels from the minute in  
169 which the image was taken to the 30th minute in the future. Therefore, each  
170 application of the CMV is 1 minute of forecasting. Finally, the prediction of  
171 clouds presence consists in checking if the new position of the clouds masks  
172 the future Sun path.

#### 173 4. Post-processing with binary probabilistic models

174 In the literature, three main categories of statistical models are proposed  
175 to generate probabilistic forecasts of binary events [24]. The first family of  
176 models, called parametric, assumes that the probability of the event follows a  
177 known distribution law, such as Gaussian or logistic. Conversely, the second  
178 type of models, called non-parametric, is not based on underlying distribu-  
179 tions. Predictions are learned from a sample of data and obviously machine  
180 learning techniques dominate this second family of models. The last cate-  
181 gory, called semi-parametric, is a mix of the two previous ones. In this work,  
182 we proposed to test two parametric models and one non-parametric model  
183 to post-process the deterministic forecasts.

##### 184 4.1. Parametric approach

185 The first approach proposed here is based on the very well-known sta-  
186 tistical models logit and probit used for decision-making problems involving  
187 binary or categorical choices in various domains such as economy [12] or  
188 meteorology [25]. These two parametric models belongs to the Generalized  
189 Linear Models (GLM) [26]. Their aim is to model the probabilities of a ran-  
190 dom response variable  $Y$  as a function of some explanatory variables. The  
191 model combines two functions. First, a function of independent explanatory  
192 variables. This function, called index function or systematic component, may  
193 be linear or not. Second, a link function that links the systematic component  
194 with the random response variable.



195 For the logit and probit models, the index function  $Z$  is a linear combi-  
 196 nation of independent explanatory variables ( $x_1, \dots, x_k$ ) and corresponding  
 197 regression coefficients ( $\beta_0, \dots, \beta_k$ ), written as follow:

$$Z = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k. \quad (3)$$

198 Used alone, this linear function is not able to provide suitable probability  
 199 levels. Indeed, linear functions are not bounded in the range  $[0; 1]$ . To  
 200 overcome this issue, link functions have been proposed to transform the result  
 201 of the index function  $Z$  into probabilities ranging between 0 (i.e. cloud) and  
 202 1 (i.e. no cloud). In our case, only the link function differentiates the logit  
 203 and probit models. For the logit model, the link function is the following  
 204 logistic function (also called sigmoid function):

$$Pr(Y = 1|X) = \frac{1}{1 + e^{-Z}}. \quad (4)$$

205 For the probit model, the link function is the cumulative standard normal  
 206 distribution function given below:

$$Pr(Y = 1|X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^Z e^{-\frac{u^2}{2}} du. \quad (5)$$

207 The input variables of these models can be either continuous, binary  
 208 or categorical. This feature is very important in our case because the ex-  
 209 planatory variables available to generate probabilistic forecasts of the cloud  
 210 presence are binary (i.e. the deterministic forecasts) and continuous (i.e.  
 211 measured irradiance, solar zenith angle, hour of the day, etc.). The main  
 212 difference between these two parametric models is the shape of the link func-  
 213 tion. The logistic function produces heavier tails than the standard normal  
 214 distribution function. To implement the logit and probit models, we used the  
 215 "glm" function of the package "stats" that is part of R [27], which is based  
 216 on the maximum likelihood approach to estimate the coefficients.

#### 217 4.2. Non-parametric approach

218 As most of real-life phenomena do not follow a known distribution law,  
 219 non-parametric models have been developed. Non-parametric binary choice  
 220 models have been initially developed for economic applications [28]. A set of  
 221 non-parametric regressions, designed for continuous variables and also suit-  
 222 able for binary events probability, are available in the literature [29, 30].

223 The main challenge for these regression models is to combine continuous,  
 224 categorical and binary data as input [31] like in our work.

225 Decision Trees (DT) and by extension Random Forests (RF), which be-  
 226 long to the supervised machine learning methods are appealing non-parametric  
 227 models to predict discrete choice. Indeed, they can predict either a nu-  
 228 merical value (regression tree), a class or a discrete choice (classification  
 229 tree). They can use either continuous, categorical or binary variable as in-  
 230 put. They require less computational effort than classical non-parametric  
 231 regression methods. Indeed, the computation time of regression methods in-  
 232 creases exponentially with the number of variables which is not the case for  
 233 DT and RF.

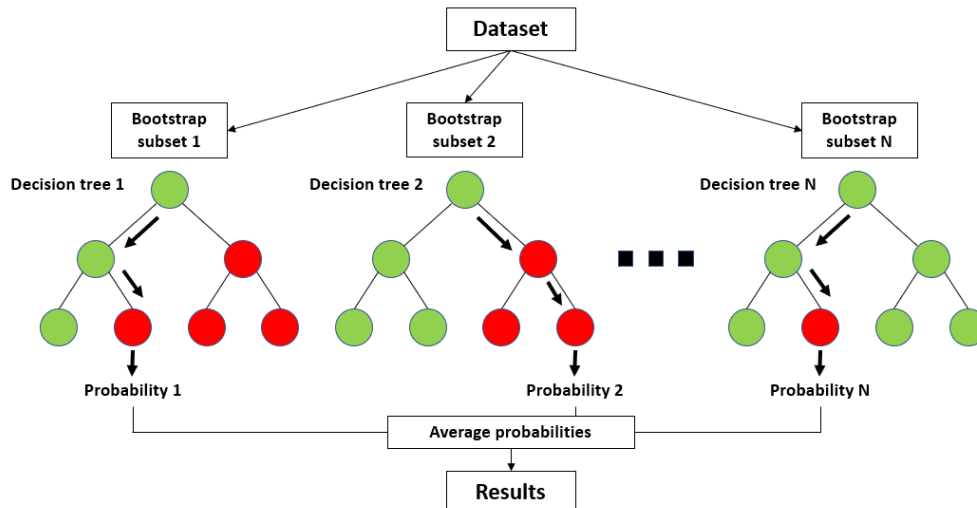


Figure 2: Simplified illustration of a Random Forest classifier used to predict class probability

234 The characteristics of the RF used in this work are introduced by [32].  
 235 The readers may refer to [33] for a general presentation. A classification tree  
 236 is a decision tool that estimates the most likely class of a categorical or a  
 237 binary variable to predict, when the input variables are known. Decision  
 238 trees are simple models that partition the features (or inputs) space into  
 239 subsets [33]. An iterative algorithm is used to split the input space. At each  
 240 step or node, the data are divided into two subsets, applying an “If, Then”  
 241 rule to one of the input variables. At each step, the selected input is chosen  
 242 to provide the best possible separation of the classes to predict. The aim is

243 to generate the optimal sequences of rules to predict the different possible  
244 classes. [32].

245 A RF is a set trees that are built on bootstrapped training subsets. Sev-  
246 eral decision trees are therefore trained. When RF are used as classifier, the  
247 probabilities of the predicted classes are averaged from the answers of the  
248 individual trees, as illustrated in figure 2. In the RF, the strengths (and  
249 weaknesses) of each tree are aggregated. A cross-validation was done on the  
250 number of trees and a good trade off was obtained for 500 trees. In this  
251 work, we used the RF classifier algorithm implemented in the R package  
252 "randomForest" based on [34].

### 253 4.3. Implementation

254 As previously introduced and presented in figure 3, the forecasting process  
255 has two main steps: generation of deterministic forecast from the sky imager  
256 and post-processing with the probabilistic models. The first step is briefly  
257 detailed in section 3. Here, we will focus on the implementation of the  
258 probabilistic models. The simplest approach is to use only the discrete cloud  
259 forecast  $\hat{y}_{t+h}$  as input of the probabilistic model. However, numerous works  
260 show that the addition of inputs, such as past observations or solar path  
261 variables, can significantly improve the quality of solar forecasts generated by  
262 time series models [35, 36] or post-processing methods [37]. Thus, to improve  
263 the performance of the post-processing step, we tested the addition of easy  
264 to compute variables as input to the three tested probabilistic models. The  
265 tested additional variables are: solar zenith angle, current and past global  
266 horizontal irradiances, beam normal irradiances and clear sky indices, mean  
267 and variability over past observed clear sky indices. The best combination  
268 of inputs, based on the BS, is:

- 269 • the deterministic forecast of cloud presence  $\hat{y}_{t+h}$ ,
- 270 • the current clear sky index  $CSK_t$ ,
- 271 • the mean over the 5 past clear sky indices  $\overline{CSK}$ .

272 Finally, we created one post-processing model by forecast horizon. Consider-  
273 ing a time resolution of 1 minute and horizons up to 30 minutes, we trained 30  
274 different models for each of the three probabilistic methods presented above.

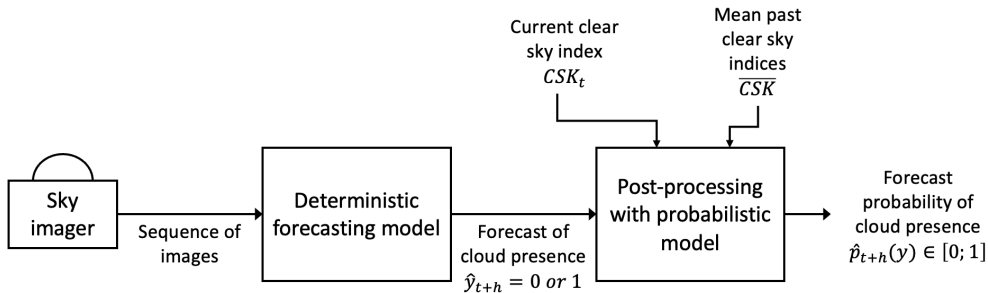


Figure 3: Diagram of the implementation of the forecasting models at time  $t$  and an horizon of forecast  $h$

## 275 5. Case study and data

276 In this study, images from a sky camera with rotational shadow band  
 277 (*TSI-880* model) have been used to provide a hemispheric vision of sky (fish-  
 278 eye vision). Additionally, the measurements of diffuse and global irradiance  
 279 from two *CMP11 Kipp & Zonen* pyranometers and direct irradiance from a  
 280 *CH1 Kipp & Zonen* pyrliometer were used, and all the instruments were  
 281 installed on a two-axis solar tracker. The testing facility is located at the  
 282 Center of Research of Solar Energy (CIESOL) at the University of Almería  
 283 in a region in southern Spain. The facility has a Mediterranean climate with  
 284 a large presence of maritime aerosols and is located at  $36.8^\circ$  N latitude and  
 285  $2.4^\circ$  W longitude at sea level. Data are collected every minute, as this was  
 286 proposed to be an suitable frequency [3]. An appropriate maintenance was  
 287 performed on the sensors and sky camera. The sensors are cleaned with ethyl  
 288 alcohol every day. The sky camera mirror is cleaned using a soft rag with  
 289 distilled water three times a week.

290 Images were taken with  $352 \times 288$  color pixel resolution, which corre-  
 291 sponds to 24 bits in JPEG format. They have three different channels that  
 292 represent the red, green and blue levels. Each pixel of the image is repre-  
 293 sented by 8 bits, with values between 0 and 255.

294 For the cloud nowcasting, data from 2010 and 2011 were used, for mo-  
 295 ments where solar altitude degree was higher than  $5^\circ$ . For 2010, a total of  
 296 137794 moments were analyzed for each interval of prediction (1 to 30 min-  
 297 utes) independently, whereas for 2011, 134993 predictions where processed,  
 298 also for each forecast interval. Year 2010 has been used to train the post-

299 processing models and year 2011 to test them.

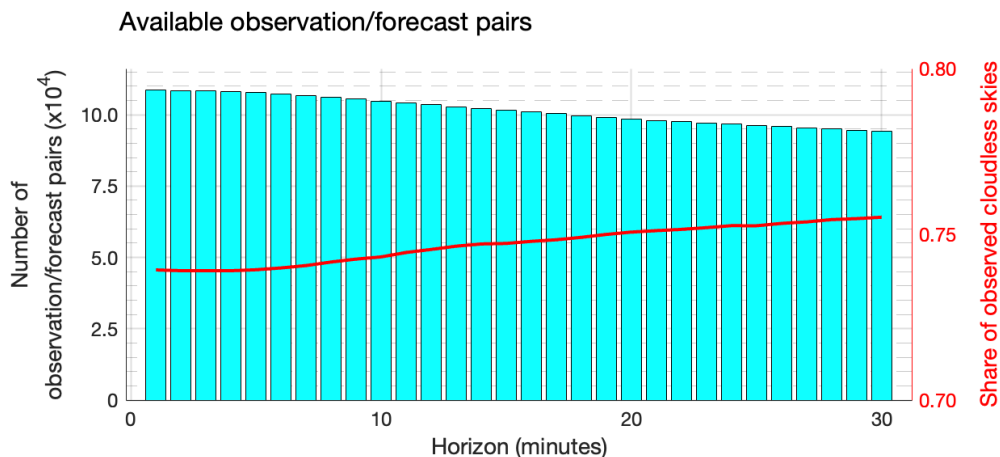


Figure 4: Number of observation/forecast pairs (blue bars) and ratio of observed cloudless skies (red line) in the test set (2011) for forecast horizons ranging from 1 to 30 minutes.

300 It should be noted that the number of observation/forecast pairs and the  
301 ratio of observed cloudless skies in the test set (2011) are not identical for the  
302 different forecast horizons. Indeed, the ASI fails to predict the presence of  
303 clouds for long horizons when the cloud speed is high and/or of the cloud base  
304 height is low. Specifically, under these conditions, predicted cloud locations  
305 have a high probability of leaving the ASI's field of view before an horizon of  
306 30 minutes. As a consequence, the total number of observation/forecast pairs  
307 decreases while the ratio of observed cloudless skies in the test set slightly  
308 increases with forecast horizon as presented in figure 4. As the clear skies  
309 are easier to forecast, this pattern will impact the assessment of the models  
310 accuracy.

## 311 6. Results and discussions

312 The post-processing of the deterministic forecasts gives a probability level  
313 of the possible future cloud presence. But, it can also be seen as a calibration  
314 of the deterministic forecasts based on the training set statistics. Further-  
315 more, it is common to transform the probability level resulting from the  
316 discrete choice models in a new binary and deterministic forecasts. To do  
317 so, we assume that a probability above 0.5 ( $> 50\%$ ) corresponds to a "yes"

318 event, i.e. in our case a clear sky. While a probability below 0.5 ( $< 50\%$ ) is  
319 a "no" event corresponding to the presence of clouds. To assess the ability of  
320 the selected discrete choice models to improve the deterministic forecast, this  
321 transformation was applied to the probabilistic forecasts. Thus, the evaluation  
322 of the generated forecasts will be performed in two steps. First we will  
323 evaluate the improvement of the quality of the deterministic forecasts before  
324 and after the post-processing step. Second, we will assess the quality of the  
325 probabilistic forecasts and the improvement compared to the corresponding  
326 deterministic forecasts. Table 1, at the end of this paper, gives the detailed  
327 numeric results used to plot the graphs evaluating the quality of the forecasts.

### 328 *6.1. Deterministic forecasts quality*

329 Figure 5 shows the evaluation of the quality of the deterministic forecasts  
330 before and after the post-processing with the three discrete choice models  
331 tested in this work. Surprisingly, with longer horizons, the accuracy of the  
332 initial forecasts done with the sky imager increases (solid black line). This  
333 observation results from the share of clear and cloudy skies available in the  
334 test sets presented previously in section 5 and figure 4. Indeed, for longer  
335 horizons, the share of clear skies, which are easier to forecast when there is  
336 no cloud in the field of view of the ASI, is more important. We could have  
337 homogenized the test sets of the different horizons to cancel this effect. How-  
338 ever, removing conditions with fast moving clouds from the shorter horizons,  
339 would have biased the analysis of the improvement brought by the prob-  
340 abilistic approach that is more interesting when forecasting becomes more  
341 uncertain. Even if this effect does not influence significantly the results of  
342 this work, the reader must keep in mind that for the longest horizons, the  
343 test sets leads to a higher share of situations that are easier to forecast.

344 As expected, the post-processing with the discrete choice models im-  
345 proves significantly the accuracy of the forecasting system. For horizons  
346 from 1 to 15 minutes, the accuracy resulting from the 3 models decreases.  
347 Above a 15 minute horizon, as for the original ASI forecasts, the accuracy  
348 increases slightly. Among the 3 tested models, the RF model, which is a non-  
349 parametric method, shows the best improvement with an accuracy of 93.4%  
350 and 90.3% for horizons of 1 minute and 30 minutes respectively. Compared  
351 to the initial ASI forecasts, this improvement correspond to a gain of 11.6  
352 percentage points for the shortest horizon (i.e. 1 minute) and 7.5 percentage  
353 point for the longest one (i.e. 30 minutes). Regarding the two parametric  
354 techniques, the logit model, which has an accuracy close to the RF, clearly

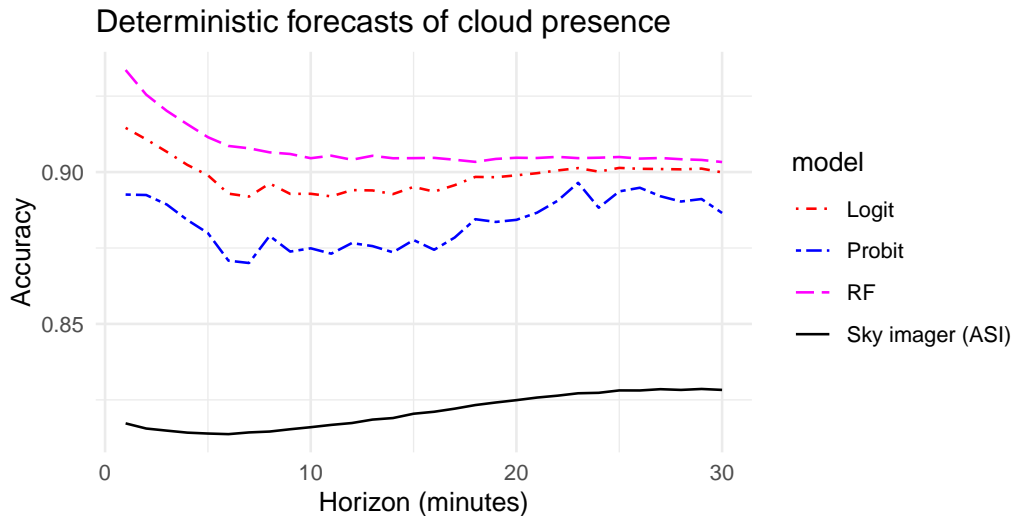


Figure 5: Accuracy (or fraction correct) of the deterministic ASI forecasts before and after the post-processing

355 outperforms the probit model. However, both of them show a significant  
 356 improvement over the ASI original forecasts. Given the simplicity and the  
 357 low computational efforts of the logit model, the former offers a very good  
 358 trade off for the study case selected in this work.

### 359 6.2. Probabilistic forecasts quality

360 As previously discussed in section 2, the reliability of the probabilistic  
 361 forecast is the the first attribute to verify. Figure 6 gives the reliability  
 362 diagrams of the 3 discrete choice models and of the climatology model for all  
 363 the horizons of forecast (i.e. overall reliability). The climatology is a very  
 364 simple model used as a reference, which forecasts the average probability of  
 365 the event whatever the weather conditions and the horizon. Here, the average  
 366 probability to have a clear sky computed from the test set is 74.8%. The  
 367 reliability diagram is a visual tool that gives a qualitative assessment of the  
 368 reliability. A perfectly reliable model should result in a reliability curve that  
 369 sticks to the diagonal. Here, none of the 3 tested models presents a perfect  
 370 reliability. Conversely to the RF models, the probit and logit models never  
 371 generate forecast probabilities of 0 and 1. As a consequence, their reliability  
 372 curves do not reach the lower and upper limit of the diagram. The important  
 373 deviations from the diagonal of the probit and logit models indicate a peak

374 of under-confidence for a forecast probability of 0.5 and an over-confidence  
 375 for forecast probabilities ranging between 0.75 and 0.9. In other word, when  
 376 these two models issue a forecast probability of 0.5 (i.e. 50% probability of a  
 377 clear sky), the actual observed frequency is higher than 0.75. The RF model  
 378 shows a better overall reliability than the 2 parametric models with a high  
 379 reliability when it forecasts a clear sky with forecast probabilities above 0.5.

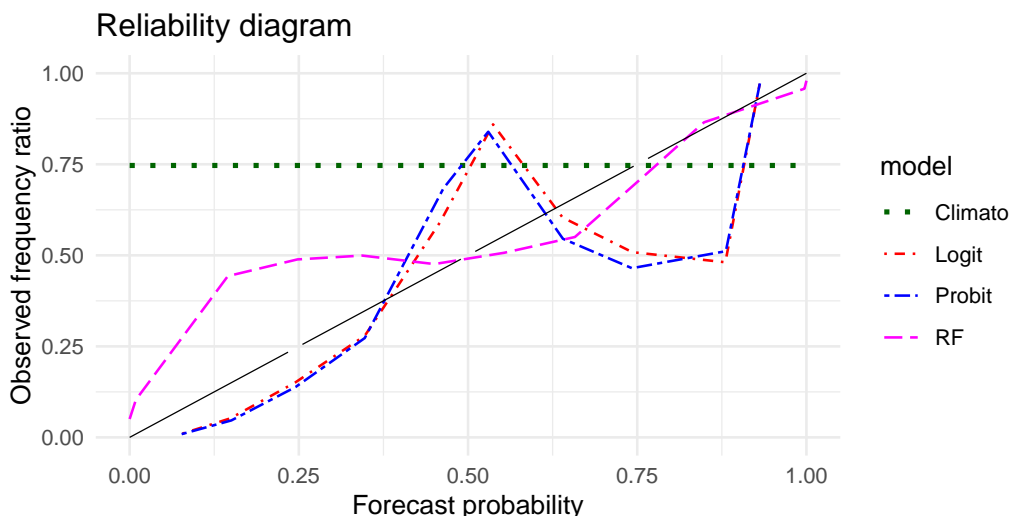


Figure 6: Reliability diagrams of the 3 discrete choice models and of the climatology

380 In addition to the reliability assessment, the BS provides quantitative  
 381 information on the quality of the forecast. The BS is negatively oriented and  
 382 a lower value indicates better quality. Figures 7 shows the BS of the original  
 383 ASI forecasts, of the climatology model, based on the training set, and of  
 384 the 3 discrete choice models. For the original ASI forecasts (solid black line),  
 385 which are deterministic, the BS is derived from the accuracy as detailed in  
 386 section 2. First, we can observe that the quality of the climatology increases  
 387 slightly with the horizon. Again, this trend results from the increased share  
 388 of clear skies for the longer horizons in the test set. Second, the BS of the  
 389 probit and logit models is almost the same regardless of the horizon. This  
 390 result, which differs from that obtained with their deterministic counterparts,  
 391 highlights that the information included in a probabilistic forecast cannot  
 392 be translated into a deterministic forecast. Finally, the RF model clearly  
 393 outperforms the 2 parametric models. The good performance of this non-  
 394 parametric model comes from several advantages. Indeed, RF is able to map



395 non-linear relationships between inputs and output. It is designed to handle  
 396 different types of variables, which can be binary, categorical or continuous.  
 397 Finally, and unlike probit and logit models, RF issues probability forecasts  
 398 of 0 and 1 with high reliability.

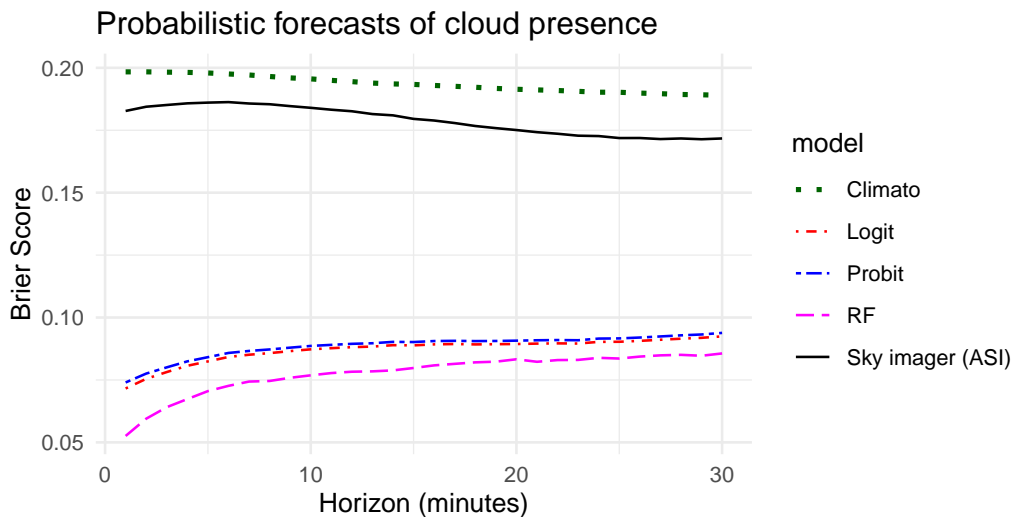


Figure 7: Brier scores of the original ASI forecasts, of the climatology model used as a reference and of the probabilistic forecasts resulting from post-processing with the 3 discrete-choice models

## 399 7. Conclusions

400 This work is a first attempt, in the field of solar energy, to propose a  
 401 methodology to generate very short-term probabilistic forecasts of the pres-  
 402 ence of clouds as a binary event. The objective is to anticipate the moment  
 403 when the direct normal irradiance is higher than a defined threshold, suitable  
 404 to the operation of concentrated solar power plants. The proposed approach  
 405 combines binary forecast based on a sky imager with discrete choice models  
 406 commonly used in various decision-making problems to generate probability  
 407 forecast of cloud presence. Two parametric (probit and logit) and one non-  
 408 parametric (RF) discrete choice models have been tested in this work. The  
 409 RF clearly outperforms the widely used probit and logit models. Beyond  
 410 a better quality assessed with the reliability diagram and the BS, the RF  
 411 provides better features, like the ability to forecast probability levels of 0 or  
 412 1 with high reliability.

413 As this work is a the first one on the topic, no comparison with other  
414 models or approaches is possible and it is difficult to evaluate the actual  
415 performance of the proposed method. However, the generated forecasts show  
416 a good quality. Indeed, the accuracy of the deterministic forecasts derived  
417 from the probability level is above 90% with an improvement ranging from 7.5  
418 to 11.6 percentage points compared to the original ASI forecasts. Regarding  
419 the probability forecasts obtained with the three tested models, their BS are  
420 below 0.1, regardless the horizon of forecast.

#### 421 **Acknowledgment**

422 The authors want to acknowledge the project MAPVSpain, with reference  
423 PID2020-118239RJ-I00, financed by the Ministerio de Ciencia e Innovación,  
424 and co-financed by the European Regional Development Fund. This research  
425 also received funding from the European Union's Horizon Europe research  
426 and innovation programme under Grant No.101076447 (TwInSolar project).

Table 1: Numerical results of the evaluation of deterministic and probabilistic forecasts for the different horizons

Horizon (minutes)	Accuracy				Brier Score			
	ASI	Probit	Logit	RF	Climato	Probit	Logit	RF
1	0.817	0.893	0.915	0.934	0.198	0.074	0.072	0.053
2	0.815	0.892	0.911	0.925	0.198	0.078	0.075	0.059
3	0.814	0.889	0.907	0.920	0.198	0.080	0.078	0.064
4	0.814	0.884	0.902	0.916	0.198	0.082	0.081	0.067
5	0.813	0.880	0.899	0.911	0.198	0.084	0.082	0.071
6	0.813	0.871	0.893	0.909	0.198	0.086	0.084	0.073
7	0.814	0.870	0.892	0.908	0.197	0.087	0.085	0.074
8	0.814	0.879	0.896	0.906	0.197	0.087	0.086	0.075
9	0.815	0.874	0.893	0.906	0.196	0.088	0.087	0.076
10	0.816	0.875	0.893	0.905	0.196	0.089	0.087	0.077
11	0.816	0.873	0.892	0.905	0.195	0.089	0.088	0.078
12	0.817	0.877	0.894	0.904	0.194	0.089	0.088	0.078
13	0.818	0.876	0.894	0.905	0.194	0.090	0.088	0.078
14	0.819	0.874	0.893	0.905	0.194	0.090	0.089	0.079
15	0.820	0.878	0.895	0.905	0.193	0.090	0.089	0.080
16	0.821	0.874	0.894	0.905	0.193	0.091	0.089	0.081
17	0.822	0.878	0.896	0.904	0.193	0.091	0.089	0.081
18	0.823	0.884	0.898	0.903	0.192	0.091	0.089	0.082
19	0.824	0.884	0.898	0.904	0.192	0.091	0.089	0.082
20	0.824	0.884	0.899	0.905	0.191	0.091	0.089	0.083
21	0.825	0.887	0.900	0.905	0.191	0.091	0.090	0.082
22	0.826	0.891	0.901	0.905	0.191	0.091	0.090	0.083
23	0.827	0.896	0.901	0.905	0.191	0.091	0.090	0.083
24	0.827	0.888	0.900	0.905	0.190	0.092	0.090	0.084
25	0.828	0.894	0.901	0.905	0.190	0.092	0.090	0.084
26	0.828	0.895	0.901	0.904	0.190	0.092	0.091	0.084
27	0.828	0.892	0.901	0.905	0.190	0.092	0.091	0.085
28	0.828	0.890	0.901	0.904	0.189	0.093	0.092	0.085
29	0.828	0.891	0.901	0.904	0.189	0.093	0.092	0.085
30	0.828	0.887	0.900	0.903	0.189	0.094	0.092	0.086
Overall	0.821	0.883	0.899	0.908	0.198	0.088	0.087	0.077

427 **References**

- 428 [1] M. Diagne, M. David, P. Lauret, J. Boland, N. Schmutz, Review of  
429 solar irradiance forecasting methods and a proposition for small-scale  
430 insular grids, *Renewable and Sustainable Energy Reviews* 27 (2013) 65–  
431 76. doi:10.1016/j.rser.2013.06.042.
- 432 [2] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. M. de Pison,  
433 F. Antonanzas-Torres, Review of photovoltaic power forecasting, *Solar*  
434 *Energy* 136 (2016) 78–111. doi:10.1016/j.solener.2016.06.069.
- 435 [3] E. Lorenz, J. A. Ruiz-Arias, L. Martin, S. Wilbert, C. Köhler, R. Fritz,  
436 A. Betti, P. Lauret, M. David, J. Huang, R. Perez, A. Kazantzidis,  
437 P. Wang, Y.-M. Saint-Drenan, Forecasting Solar Radiation and Pho-  
438 tovoltaic Power, in: *Best Practices Handbook for the Collection and*  
439 *Use of Solar Resource Data for Solar Energy Applications: Third Edi-*  
440 *tion*, no. NREL/TP-5D00-77635, National Renewable Energy Labora-  
441 tory, Golden, CO, 2021.  
442 URL <https://www.nrel.gov/docs/fy21osti/77635.pdf>
- 443 [4] M. Ajith, M. Martínez-Ramón, Deep learning algorithms for very short  
444 term solar irradiance forecasting: A survey, *Renewable and Sustainable*  
445 *Energy Reviews* 182 (2023). doi:10.1016/j.rser.2023.113362.
- 446 [5] J. Alonso-Montesinos, F. Batlles, C. Portillo, Solar irradiance forecasting  
447 at one-minute intervals for different sky conditions using sky camera  
448 images, *Energy Conversion and Management* 105 (2015) 1166 – 1177.  
449 doi:10.1016/j.enconman.2015.09.001.
- 450 [6] S.-A. Logothetis, V. Salamalikis, S. Wilbert, J. Remund, L. F. Zarza-  
451 lejo, Y. Xie, B. Nouri, E. Ntavelis, J. Nou, N. Hendrikx, L. Visser,  
452 M. Sengupta, M. Pó, R. Chauvin, S. Grieu, N. Blum, W. van Sark,  
453 A. Kazantzidis, Benchmarking of solar irradiance nowcast performance  
454 derived from all-sky imagers, *Renewable Energy* 199 (2022) 246 – 261.  
455 doi:10.1016/j.renene.2022.08.127.
- 456 [7] J. Alonso-Montesinos, R. Monterreal, J. Fernandez-Reche, J. Ballestrín,  
457 G. López, J. Polo, F. J. Barbero, A. Marzo, C. Portillo, F. J. Batlles,  
458 Nowcasting system based on sky camera images to predict the solar flux  
459 on the receiver of a concentrated solar plant, *Remote Sensing* 14 (7)  
460 (2022). doi:10.3390/rs14071602.

- 461 [8] Q. Paletta, G. Arbod, J. Lasenby, Omnivision forecasting: Combining  
462 satellite and sky images for improved deterministic and probabilistic  
463 intra-hour solar energy predictions, *Applied Energy* 336 (2023). doi:  
464 10.1016/j.apenergy.2023.120818.
- 465 [9] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, R. J. Hyndman,  
466 Probabilistic energy forecasting: Global Energy Forecasting Competi-  
467 tion 2014 and beyond, *International Journal of Forecasting* 32 (3) (2016)  
468 896–913. doi:10.1016/j.ijforecast.2016.02.001.
- 469 [10] D. van der Meer, J. Widén, J. Munkhammar, Review on probabilistic  
470 forecasting of photovoltaic power production and electricity consump-  
471 tion, *Renewable and Sustainable Energy Reviews* 81 (2018) 1484–1512.  
472 doi:10.1016/j.rser.2017.05.212.
- 473 [11] J. Alonso-Montesinos, J. Polo, J. Ballestrín, F. Batlles, C. Portillo, Im-  
474 pact of dni forecasting on csp tower plant power production, *Renewable*  
475 *Energy* 138 (2019) 368 – 377. doi:10.1016/j.renene.2019.01.095.
- 476 [12] D. L. McFadden, Econometric models of probabilistic choice, in: C. F.  
477 Manski, D. L. McFadden (Eds.), *Structural Analysis of Discrete Data*  
478 *with Econometric Applications*, MIT Press, Cambridge, MA, USA,  
479 1981, pp. 198–272.
- 480 [13] Daniel S. Wilks, *Statistical methods in the atmospheric sciences*, 2nd  
481 Edition, no. 91 in *International geophysics series*, Elsevier [u.a.], Ams-  
482 terdam, 2009.
- 483 [14] J. Alonso, F. Batlles, Short and medium-term cloudiness forecasting  
484 using remote sensing techniques and sky camera imagery, *Energy* 73  
485 (2014) 890–897. doi:10.1016/j.energy.2014.06.101.
- 486 [15] H. Escrig, F. J. Batlles, J. Alonso, F. M. Baena, J. L. Bosch, I. B.  
487 Salbidegoitia, J. I. Burgaleta, Cloud detection, classification and motion  
488 estimation using geostationary satellite imagery for cloud cover forecast,  
489 *Energy* 55 (2013) 853–859.
- 490 [16] D. Yang, S. Alessandrini, J. Antonanzas, F. Antonanzas-Torres,  
491 V. Badescu, H. G. Beyer, R. Blaga, J. Boland, J. M. Bright, C. F.  
492 Coimbra, M. David, A. Frimane, C. A. Gueymard, T. Hong, M. J.

- 493 Kay, S. Killinger, J. Kleissl, P. Lauret, E. Lorenz, D. van der  
494 Meer, M. Paulescu, R. Perez, O. Perpiñán-Lamigueiro, I. M. Peters,  
495 G. Reikard, D. Renné, Y.-M. Saint-Drenan, Y. Shuai, R. Urraca,  
496 H. Verbois, F. Vignola, C. Voyant, J. Zhang, Verification of deter-  
497 ministic solar forecasts, *Solar Energy* (2020) S0038092X20303947doi:  
498 10.1016/j.solener.2020.04.019.
- 499 [17] P. Lauret, M. David, P. Pinson, Verification of solar irradiance prob-  
500 abilistic forecasts, *Solar Energy* 194 (2019) 254–271. doi:10.1016/j.  
501 solener.2019.10.041.
- 502 [18] T. Gneiting, S. Lerch, B. Schulz, Probabilistic solar forecasting: Bench-  
503 marks, post-processing, verification, *Solar Energy* 252 (2023) 72–80.  
504 doi:10.1016/j.solener.2022.12.054.
- 505 [19] A. H. Murphy, What Is a Good Forecast? An Essay on the Nature  
506 of Goodness in Weather Forecasting, *Weather and Forecasting* 8 (2)  
507 (1993) 281–293. doi:10.1175/1520-0434(1993)008<0281:WIAGFA>2.  
508 0.CO;2.
- 509 [20] H. Brooks, B. Brown, B. Brown, C. Ferro, J. Jolliffe, T.-Y. Koh, P. Roeb-  
510 ber, D. Stephenson, Joint working group on forecast verification research  
511 (Jan 2015).  
512 URL <https://cawcr.gov.au/projects/verification/>
- 513 [21] R. J. Hogan, I. B. Mason, *Deterministic Forecasts of Binary Events*,  
514 John Wiley & Sons, Ltd, 2011, Ch. 3, pp. 31–59. doi:10.1002/  
515 9781119960003.ch3.
- 516 [22] P. Pinson, H. A. Nielsen, J. K. Møller, H. Madsen, G. N. Karinio-  
517 takis, Non-parametric probabilistic forecasts of wind power: required  
518 properties and evaluation, *Wind Energy* 10 (6) (2007) 497–516. doi:  
519 10.1002/we.230.
- 520 [23] T. M. Hamill, Reliability diagrams for multicategory probabilistic fore-  
521 casts, *Weather and Forecasting* 12 (4) (1997) 736 – 741. doi:10.1175/  
522 1520-0434(1997)012<0736:RDFMPF>2.0.CO;2.
- 523 [24] K. Lahiri, L. Yang, Chapter 19 - forecasting binary outcomes, in: G. El-  
524 liott, A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Vol. 2

- 525 of Handbook of Economic Forecasting, Elsevier, 2013, pp. 1025–1106.  
526 doi:10.1016/B978-0-444-62731-5.00019-1.
- 527 [25] J. Sánchez, J. Marcos, M. de la Fuente, A. Castro, A logistic regression  
528 model applied to short term forecast of hail risk, *Physics and Chemistry  
529 of the Earth* 23 (5) (1998) 645–648. doi:10.1016/S0079-1946(98)  
530 00102-5.
- 531 [26] P. McCullagh, J. A. Nelder, *Generalized Linear Models*, Springer US,  
532 Boston, MA, 1989. doi:10.1007/978-1-4899-3242-6.
- 533 [27] R Core Team, *R: A Language and Environment for Statistical Com-  
534 puting*, R Foundation for Statistical Computing, Vienna, Austria, ISBN  
535 3-900051-07-0 (2022).  
536 URL <http://www.R-project.org/>
- 537 [28] M. Frölich, Non-parametric regression for binary dependent variables,  
538 *The Econometrics Journal* 9 (3) (2006) 511–540. doi:10.1111/j.1368-  
539 423X.2006.00196.x.
- 540 [29] Q. Li, J. S. Racine, *Nonparametric econometrics: theory and practice*,  
541 Princeton University Press, Princeton, N.J, 2007, oCLC: ocm67346329.
- 542 [30] W. K. Newey, Nonparametric continuous/discrete choice models\*, *In-  
543 ternational Economic Review* 48 (4) (2007) 1429–1439. doi:10.1111/  
544 j.1468-2354.2007.00469.x.
- 545 [31] Q. Li, J. S. Racine, Nonparametric estimation of conditional cdf and  
546 quantile functions with mixed categorical and continuous data, *Journal  
547 of Business & Economic Statistics* 26 (4) (2008) 423–434. doi:10.1198/  
548 073500107000000250.
- 549 [32] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, *Classification  
550 and regression trees*, Routledge, 2017.
- 551 [33] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learn-  
552 ing*, Springer Series in Statistics, Springer New York.
- 553 [34] L. Breiman, Random Forests, *Machine Learning* 45 (1) (2001) 5–32.  
554 doi:10.1023/A:1010933404324.

- 555 [35] R. Alonso-Suárez, M. David, V. Branco, P. Lauret, Intra-day solar prob-  
556 abilistic forecasts including local short-term variability and satellite in-  
557 formation, *Renewable Energy* 158 (2020) 554–573. doi:10.1016/j.  
558 *renene*.2020.05.046.
- 559 [36] P. Lauret, M. David, H. Pedro, Probabilistic Solar Forecasting Using  
560 Quantile Regression Models, *Energies* 10 (10) (2017) 1591. doi:10.  
561 3390/en10101591.
- 562 [37] E. Lorenz, J. Hurka, D. Heinemann, H. G. Beyer, Irradiance Forecasting  
563 for the Power Prediction of Grid-Connected Photovoltaic Systems, *IEEE*  
564 *Journal of Selected Topics in Applied Earth Observations and Remote*  
565 *Sensing* 2 (1) (2009) 2–10. doi:10.1109/JSTARS.2009.2020300.