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

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

Mathieu David[a,*], Joaquín Alonso-Montesinos[b,c], Josselin Le Gal La Salle[a], Philippe Lauret[a]

[a]*PIMENT, University of La Réunion, 97715 Saint-Denis, Reunion*
[b]*Department of Chemistry and Physics, University of Almería, 04120 Almería, Spain*
[c]*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. Introduction

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

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

It is well-known that weather forecasts are uncertain because the evolution of the weather and consequently solar irradiance are chaotic processes. Thus, in decision-making operations that use solar forecasts, such as the management of power plants, probabilistic forecasts are crucial. Indeed, probabilistic forecasts assign probability levels to future events and allow their users to assess the associated risks. Research works dealing with probabilistic solar forecast are relatively recent but numerous have been released on the topic in the last 10 years [9, 10, 3]. However, as mentioned previously, very few works concerning ASI proposed a method to generate probabilistic solar forecast. This work will contribute to fill this gap in the literature.

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

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

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

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

2. Overall methodology and forecasts evaluation

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

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

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

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

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

Figure 1: Contingency table for a binary forecast

$$Accuracy = \frac{hits + correct\ negatives}{N} \tag{1}$$

Regarding the verification of probabilistic forecasts, two main attributes define the quality: the reliability and the resolution. Reliability refers to the statistical consistency between the forecasts and the observations. In other words, the forecast probability should be equal to the observed probability of the event (e.g. 20% of the events should happens for a forecast probability of 20%). The reliability is a crucial prerequisite as non reliable forecasts would lead to a systematic bias in subsequent decision-making processes [22]. The most used visual tools to assess the reliability is the reliability diagram [23]. It plots the correspondence between the forecast probability (x axis) and the observed frequency of the event (y axis). Perfectly reliable forecasts should be as close to the diagonal as possible. Figure 6 shows the reliability diagram for the different models tested in this work. Resolution refers to the ability of a forecasting system to generate case-dependent forecasts. For example, the climatology model, which predicts the average probability of the event (i.e. always the same probability regardless the horizon or the weather conditions) has no resolution. Unfortunately, no graphical tool exists to evaluate the resolution.

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

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

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

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

3. Deterministic forecasts of cloud presence with a sky camera

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

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

- The cloud motion vector (CMV) is calculated for each sector by applying the maximum cross-correlation method.
- Different quality tests are applied to ensure the correct determination of the cloud motion.

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

4. Post-processing with binary probabilistic models

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

4.1. Parametric approach

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

For the logit and probit models, the index function Z is a linear combination of independent explanatory variables (x_1, ..., x_k) and corresponding regression coefficients (β_0, ..., β_k), written as follow:

$$Z = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k. \tag{3}$$

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

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

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

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

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

4.2. Non-parametric approach

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

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

Decision Trees (DT) and by extension Random Forests (RF), which belong to the supervised machine learning methods are appealing non-parametric models to predict discrete choice. Indeed, they can predict either a numerical value (regression tree), a class or a discrete choice (classification tree). They can use either continuous, categorical or binary variable as input. They require less computational effort than classical non-parametric regression methods. Indeed, the computation time of regression methods increases exponentially with the number of variables which is not the case for DT and RF.

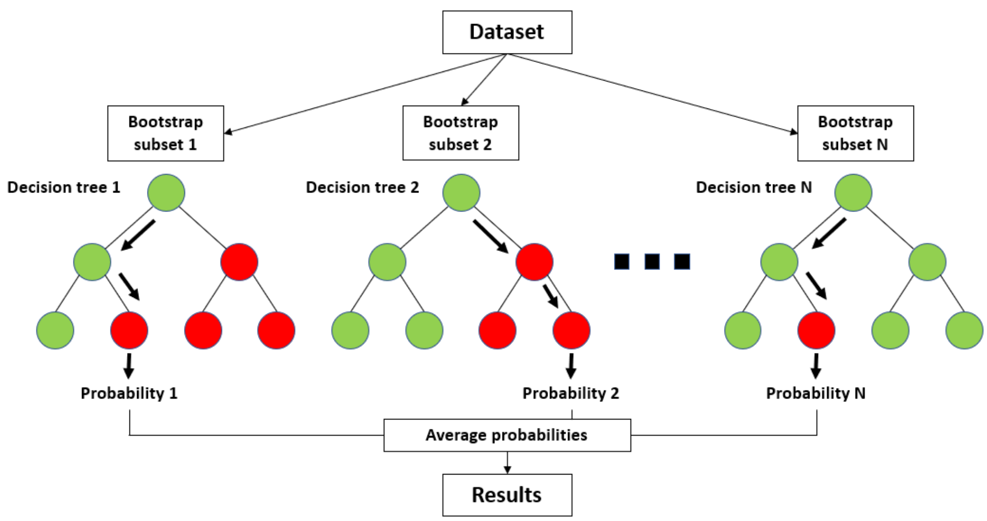

Figure 2: Simplified illustration of a Random Forest classifier used to predict class probability

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

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

4.3. Implementation

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

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

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

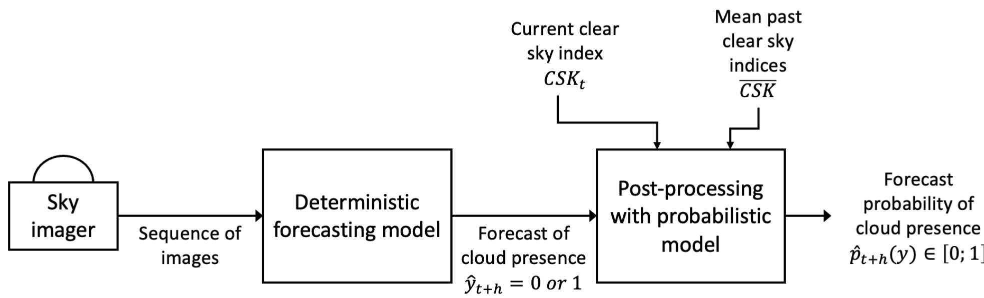

Figure 3: Diagram of the implementation of the forecasting models at time t and an horizon of forecast h

5. Case study and data

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

Images were taken with 352 x 288 color pixel resolution, which corresponds to 24 bits in JPEG format. They have three different channels that represent the red, green and blue levels. Each pixel of the image is represented by 8 bits, with values between 0 and 255.

For the cloud nowcasting, data from 2010 and 2011 were used, for moments where solar altitude degree was higher than 5^o. For 2010, a total of 137794 moments were analyzed for each interval of prediction (1 to 30 minutes) independently, whereas for 2011, 134993 predictions where processed, also for each forecast interval. Year 2010 has been used to train the post-processing models and year 2011 to test them.

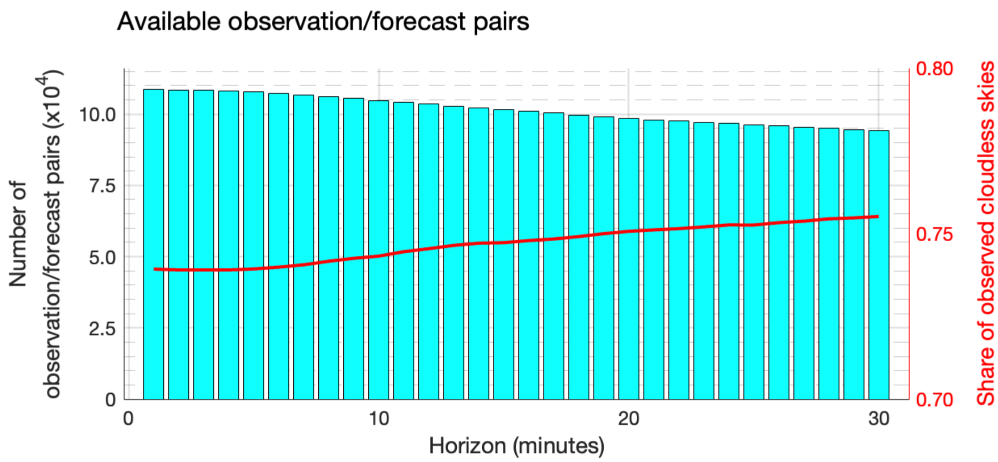

Figure 4: Number of observation/forecast pairs (blue bars) and ratio of observed cloudless skies (red line) in the test set (2011) for forecast horizons ranging from 1 to 30 minutes.

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

6. Results and discussions

The post-processing of the deterministic forecasts gives a probability level of the possible future cloud presence. But, it can also be seen as a calibration of the deterministic forecasts based on the training set statistics. Furthermore, it is common to transform the probability level resulting from the discrete choice models in a new binary and deterministic forecasts. To do so, we assume that a probability above 0.5 ($> 50\%$) corresponds to a "yes" event, i.e. in our case a clear sky. While a probability below 0.5 ($< 50\%$) is a "no" event corresponding to the presence of clouds. To asses the ability of the selected discrete choice models to improve the deterministic forecast, this transformation was applied to the probabilistic forecasts. Thus, the evaluation of the generated forecasts will be performed in two steps. First we will evaluate the improvement of the quality of the deterministic forecasts before and after the post-processing step. Second, we will assess the quality of the probabilistic forecasts and the improvement compared to the corresponding deterministic forecasts. Table 1, at the end of this paper, gives the detailed numeric results used to plot the graphs evaluating the quality of the forecasts.

6.1. Deterministic forecasts quality

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

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

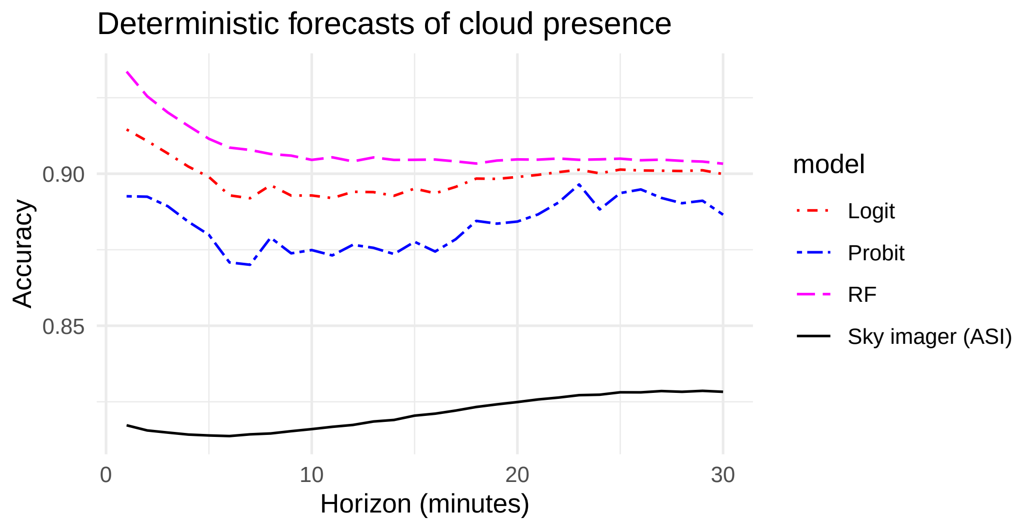

Figure 5: Accuracy (or fraction correct) of the deterministic ASI forecasts before and after the post-processing

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

6.2. Probabilistic forecasts quality

As previously discussed in section 2, the reliability of the probabilistic forecast is the the first attribute to verify. Figure 6 gives the reliability diagrams of the 3 discrete choice models and of the climatology model for all the horizons of forecast (i.e. overall reliability). The climatology is a very simple model used as a reference, which forecasts the average probability of the event whatever the weather conditions and the horizon. Here, the average probability to have a clear sky computed from the test set is 74.8%. The reliability diagram is a visual tool that gives a qualitative assessment of the reliability. A perfectly reliable model should result in a reliability curve that sticks to the diagonal. Here, none of the 3 tested models presents a perfect reliability. Conversely to the RF models, the probit and logit models never generate forecast probabilities of 0 and 1. As a consequence, their reliability curves do not reach the lower and upper limit of the diagram. The important deviations from the diagonal of the probit and logit models indicate a peak of under-confidence for a forecast probability of 0.5 and an over-confidence for forecast probabilities ranging between 0.75 and 0.9. In other word, when these two models issue a forecast probability of 0.5 (i.e. 50% probability of a clear sky), the actual observed frequency is higher than 0.75. The RF model shows a better overall reliability than the 2 parametric models with a high reliability when it forecasts a clear sky with forecast probabilities above 0.5.

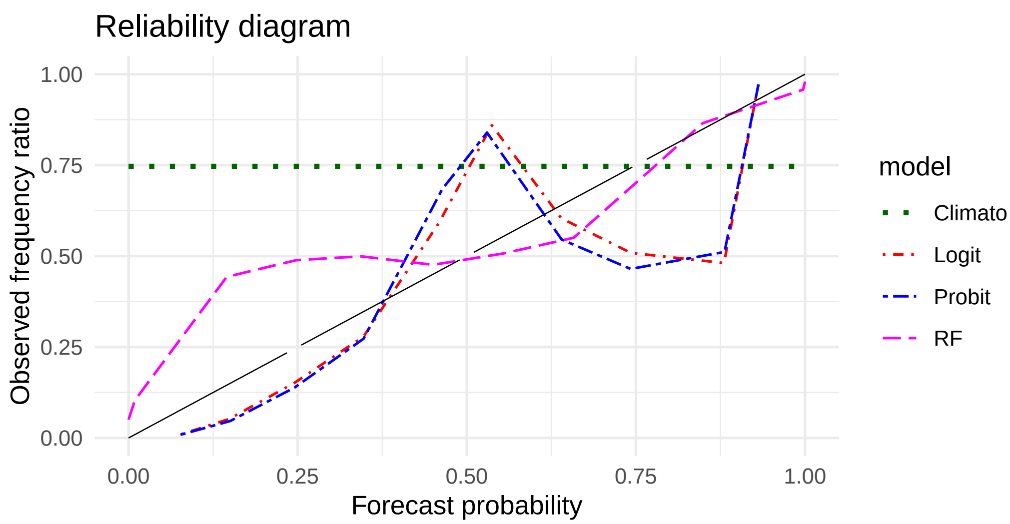

Figure 6: Reliability diagrams of the 3 discrete choice models and of the climatology

In addition to the reliability assessment, the BS provides quantitative information on the quality of the forecast. The BS is negatively oriented and a lower value indicates better quality. Figures 7 shows the BS of the original ASI forecasts, of the climatology model, based on the training set, and of the 3 discrete choice models. For the original ASI forecasts (solid black line), which are deterministic, the BS is derived from the accuracy as detailed in section 2. First, we can observe that the quality of the climatology increases slightly with the horizon. Again, this trend results from the increased share of clear skies for the longer horizons in the test set. Second, the BS of the probit and logit models is almost the same regardless of the horizon. This result, which differs from that obtained with their deterministic counterparts, highlights that the information included in a probabilistic forecast cannot be translated into a deterministic forecast. Finally, the RF model clearly outperforms the 2 parametric models. The good performance of this non-parametric model comes from several advantages. Indeed, RF is able to map non-linear relationships between inputs and output. It is designed to handle different types of variables, which can be binary, categorical or continuous. Finally, and unlike probit and logit models, RF issues probability forecasts of 0 and 1 with high reliability.

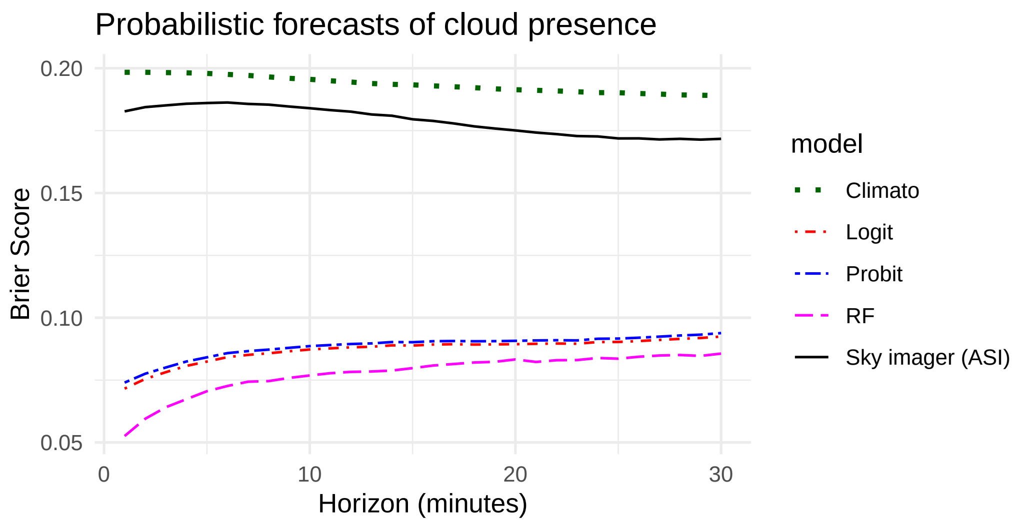

Figure 7: Brier scores of the original ASI forecasts, of the climatology model used as a reference and of the probabilistic forecasts resulting from post-processing with the 3 discrete-choice models

7. Conclusions

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

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

Acknowledgment

The authors want to acknowledge the project MAPVSpain, with reference PID2020-118239RJ-I00, financed by the Ministerio de Ciencia e Innovación, and co-financed by the European Regional Development Fund. This research also received funding from the European Union's Horizon Europe research 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	Accuracy				Brier Score			
(minutes)	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

References

[1] M. Diagne, M. David, P. Lauret, J. Boland, N. Schmutz, Review of solar irradiance forecasting methods and a proposition for small-scale insular grids, Renewable and Sustainable Energy Reviews 27 (2013) 65–76. doi:10.1016/j.rser.2013.06.042.

[2] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. M. de Pison, F. Antonanzas-Torres, Review of photovoltaic power forecasting, Solar Energy 136 (2016) 78–111. doi:10.1016/j.solener.2016.06.069.

[3] E. Lorenz, J. A. Ruiz-Arias, L. Martin, S. Wilbert, C. Köhler, R. Fritz, A. Betti, P. Lauret, M. David, J. Huang, R. Perez, A. Kazantzidis, P. Wang, Y.-M. Saint-Drenan, Forecasting Solar Radiation and Photovoltaic Power, in: Best Practices Handbook for the Collection and Use of Solar Resource Data for Solar Energy Applications: Third Edition, no. NREL/TP-5D00-77635, National Renewable Energy Laboratory, Golden, CO, 2021.
URL https://www.nrel.gov/docs/fy21osti/77635.pdf

[4] M. Ajith, M. Martínez-Ramón, Deep learning algorithms for very short term solar irradiance forecasting: A survey, Renewable and Sustainable Energy Reviews 182 (2023). doi:10.1016/j.rser.2023.113362.

[5] J. Alonso-Montesinos, F. Batlles, C. Portillo, Solar irradiance forecasting at one-minute intervals for different sky conditions using sky camera images, Energy Conversion and Management 105 (2015) 1166 – 1177. doi:10.1016/j.enconman.2015.09.001.

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

[7] J. Alonso-Montesinos, R. Monterreal, J. Fernandez-Reche, J. Ballestrín, G. López, J. Polo, F. J. Barbero, A. Marzo, C. Portillo, F. J. Batlles, Nowcasting system based on sky camera images to predict the solar flux on the receiver of a concentrated solar plant, Remote Sensing 14 (7) (2022). doi:10.3390/rs14071602.

[8] Q. Paletta, G. Arbod, J. Lasenby, Omnivision forecasting: Combining satellite and sky images for improved deterministic and probabilistic intra-hour solar energy predictions, Applied Energy 336 (2023). doi:10.1016/j.apenergy.2023.120818.

[9] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, R. J. Hyndman, Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond, International Journal of Forecasting 32 (3) (2016) 896–913. doi:10.1016/j.ijforecast.2016.02.001.

[10] D. van der Meer, J. Widén, J. Munkhammar, Review on probabilistic forecasting of photovoltaic power production and electricity consumption, Renewable and Sustainable Energy Reviews 81 (2018) 1484–1512. doi:10.1016/j.rser.2017.05.212.

[11] J. Alonso-Montesinos, J. Polo, J. Ballestrín, F. Batlles, C. Portillo, Impact of dni forecasting on csp tower plant power production, Renewable Energy 138 (2019) 368 – 377. doi:10.1016/j.renene.2019.01.095.

[12] D. L. McFadden, Econometric models of probabilistic choice, in: C. F. Manski, D. L. McFadden (Eds.), Structural Analysis of Discrete Data with Econometric Applications, MIT Press, Cambridge, MA, USA, 1981, pp. 198–272.

[13] Daniel S. Wilks, Statistical methods in the atmospheric sciences, 2nd Edition, no. 91 in International geophysics series, Elsevier [u.a.], Amsterdam, 2009.

[14] J. Alonso, F. Batlles, Short and medium-term cloudiness forecasting using remote sensing techniques and sky camera imagery, Energy 73 (2014) 890–897. doi:10.1016/j.energy.2014.06.101.

[15] H. Escrig, F. J. Batlles, J. Alonso, F. M. Baena, J. L. Bosch, I. B. Salbidegoitia, J. I. Burgaleta, Cloud detection, classification and motion estimation using geostationary satellite imagery for cloud cover forecast, Energy 55 (2013) 853–859.

[16] D. Yang, S. Alessandrini, J. Antonanzas, F. Antonanzas-Torres, V. Badescu, H. G. Beyer, R. Blaga, J. Boland, J. M. Bright, C. F. Coimbra, M. David, A. Frimane, C. A. Gueymard, T. Hong, M. J.

Kay, S. Killinger, J. Kleissl, P. Lauret, E. Lorenz, D. van der Meer, M. Paulescu, R. Perez, O. Perpiñán-Lamigueiro, I. M. Peters, G. Reikard, D. Renné, Y.-M. Saint-Drenan, Y. Shuai, R. Urraca, H. Verbois, F. Vignola, C. Voyant, J. Zhang, Verification of deterministic solar forecasts, Solar Energy (2020) S0038092X20303947doi:10.1016/j.solener.2020.04.019.

[17] P. Lauret, M. David, P. Pinson, Verification of solar irradiance probabilistic forecasts, Solar Energy 194 (2019) 254–271. doi:10.1016/j.solener.2019.10.041.

[18] T. Gneiting, S. Lerch, B. Schulz, Probabilistic solar forecasting: Benchmarks, post-processing, verification, Solar Energy 252 (2023) 72–80. doi:10.1016/j.solener.2022.12.054.

[19] A. H. Murphy, What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting, Weather and Forecasting 8 (2) (1993) 281–293. doi:10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

[20] H. Brooks, B. Brown, B. Brown, C. Ferro, J. Jolliffe, T.-Y. Koh, P. Roebber, D. Stephenson, Joint working group on forecast verification research (Jan 2015).
URL https://cawcr.gov.au/projects/verification/

[21] R. J. Hogan, I. B. Mason, Deterministic Forecasts of Binary Events, John Wiley & Sons, Ltd, 2011, Ch. 3, pp. 31–59. doi:10.1002/9781119960003.ch3.

[22] P. Pinson, H. A. Nielsen, J. K. Møller, H. Madsen, G. N. Kariniotakis, Non-parametric probabilistic forecasts of wind power: required properties and evaluation, Wind Energy 10 (6) (2007) 497–516. doi:10.1002/we.230.

[23] T. M. Hamill, Reliability diagrams for multicategory probabilistic forecasts, Weather and Forecasting 12 (4) (1997) 736 – 741. doi:10.1175/1520-0434(1997)012<0736:RDFMPF>2.0.CO;2.

[24] K. Lahiri, L. Yang, Chapter 19 - forecasting binary outcomes, in: G. Elliott, A. Timmermann (Eds.), Handbook of Economic Forecasting, Vol. 2

of Handbook of Economic Forecasting, Elsevier, 2013, pp. 1025–1106. doi:10.1016/B978-0-444-62731-5.00019-1.

[25] J. Sánchez, J. Marcos, M. de la Fuente, A. Castro, A logistic regression model applied to short term forecast of hail risk, Physics and Chemistry of the Earth 23 (5) (1998) 645–648. doi:10.1016/S0079-1946(98)00102-5.

[26] P. McCullagh, J. A. Nelder, Generalized Linear Models, Springer US, Boston, MA, 1989. doi:10.1007/978-1-4899-3242-6.

[27] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0 (2022).
URL http://www.R-project.org/

[28] M. Frölich, Non-parametric regression for binary dependent variables, The Econometrics Journal 9 (3) (2006) 511–540. doi:10.1111/j.1368-423X.2006.00196.x.

[29] Q. Li, J. S. Racine, Nonparametric econometrics: theory and practice, Princeton University Press, Princeton, N.J, 2007, oCLC: ocm67346329.

[30] W. K. Newey, Nonparametric continuous/discrete choice models*, International Economic Review 48 (4) (2007) 1429–1439. doi:10.1111/j.1468-2354.2007.00469.x.

[31] Q. Li, J. S. Racine, Nonparametric estimation of conditional cdf and quantile functions with mixed categorical and continuous data, Journal of Business & Economic Statistics 26 (4) (2008) 423–434. doi:10.1198/073500107000000250.

[32] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, Classification and regression trees, Routledge, 2017.

[33] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer Series in Statistics, Springer New York.

[34] L. Breiman, Random Forests, Machine Learning 45 (1) (2001) 5–32. doi:10.1023/A:1010933404324.

[35] R. Alonso-Suárez, M. David, V. Branco, P. Lauret, Intra-day solar probabilistic forecasts including local short-term variability and satellite information, Renewable Energy 158 (2020) 554–573. doi:10.1016/j.renene.2020.05.046.

[36] P. Lauret, M. David, H. Pedro, Probabilistic Solar Forecasting Using Quantile Regression Models, Energies 10 (10) (2017) 1591. doi:10.3390/en10101591.

[37] E. Lorenz, J. Hurka, D. Heinemann, H. G. Beyer, Irradiance Forecasting for the Power Prediction of Grid-Connected Photovoltaic Systems, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 2 (1) (2009) 2–10. doi:10.1109/JSTARS.2009.2020300.