Long-Term Solar Irradiance Forecasting

Braga D.[a], Chicco G.[b], Golovanov N.[c], Porumb R.[c]
Technical University of Moldova[a], University Politecnico di Torino[b], University Politehnica of Bucharest[c]
Chisinau, Republic of Moldova[a], Torino, Italy[b], Bucharest, Romania[c]

Abstract. The past decade has been characterized by considerable increase of the penetration level of solar photovoltaic systems in energy systems throughout the world. At the same time, solar irradiance has an intermittent nature. Thus, the efficient management of existing and new solar photovoltaic systems requires an accurate forecasting system of solar irradiance. The purpose of the paper is to develop and validate a long-term forecasting model for solar irradiance. This purpose is achieved by applying of clustering method and standard mathematical statistics. The modeling includes pre-processing of historical data used for forecasting and post-processing of the types of days resulted from the clustering analysis. Historical data include solar irradiance and sky coverage by clouds. Pre-processing supposes bi-normalization of the solar irradiance in time and amplitude, as well as clustering, and post-processing supposes denormalization to get the actual values. Error metrics and confusion matrix indices have been used to assess the accuracy of the proposed forecasting method. Four different model variants have been considered, which differ by pre-processing approach of initial data. The comparison of these model variants shows that for better accuracy it is required to use seasonality aspects of solar irradiance. The main result of paper is the created model, which can be used for the solar irradiance forecast with acceptable accuracy for this type of forecasting and for generating of the types of days for different annual scenarios. The importance of paper results consists in the possibility of using of these scenarios for feasibility assessment of the solar photovoltaic system and identifying of the best solutions for their integration in the energy system.
***Keywords*:** prediction, solar irradiance, forecasting, clustering, statistical error parameters, predictive model, confusion matrix, scenario analysis.
DOI: 10.5281/zenodo.3713424
UDC: 551.521.1

Prognoza pe termen lung al iradianței solare

Braga D.[a], Chicco G.[b], Golovanov N.[c], Porumb R.[c]
Universitatea Tehnică a Moldovei[a], Universitatea Politecnico di Torino[b], Universitatea Politehnica București[c]
Chișinău, Republica Moldova[a], Torino, Italia[b], București, Romania[c]

Abstract. Pe parcursul ultimului deceniu, în întreaga lume, s-a înregistrat o creștere considerabilă a numărului de sisteme solare fotoelectrice conectate la sistemul electroenergetic. În același timp, iradianța solară este caracterizată de un grad înalt de intermitență. Astfel, pentru un management eficient al sistemelor solare fotoelectrice existente și viitoare este necesară existența unui model de prognozare precisă a iradianței solare. Scopul lucrării constituie elaborarea și verificarea unui model pentru efectuarea prognozei iradianței solare pe termen lung. Scopul înaintat este realizat prin utilizarea metodei clusterelor și statistica matematică. Elaborarea modelului include etape de pregătire a datelor istorice necesare pentru realizarea prognozei și prelucrarea datelor, obținute în procesul de clusterizare și analiză. Datele istorice utilizate includ iradianța solară și nebulozitatea. Procesul de pregătire a datelor inițiale include procedeul de bi-normalizare și clusterizare a iradianței solare și al perioadei de strălucire a soarelui, iar prelucrarea rezultatului – procedeul de denormalizare pentru obținerea valorilor reale ale iradianței solare. Acuratețea modelului propus este verificată cu ajutorul indicatorilor de eroare și a matricei de confuzie. Sunt analizate patru variante ale modelului, care diferă prin abordarea pregătirii datelor inițiale pentru modelare. Compararea acestor variante de modele pentru prognoza iradianței solare au arătat că, pentru obținerea unei precizii mai mari este necesar să se țină cont particularitățile sezoniere ale iradianței solare. Cel mai semnificativ rezultat al lucrării constă în crearea modelului care poate fi utilizat pentru prognoza iradianței solare cu o precizie acceptabilă pentru acest tip de prognoză și generarea succesiunii tipului zilelor pentru diferite scenarii anuale. Valoarea rezultatului obținute în lucrare constă în posibilitatea utilizării scenariilor anuale generate pentru evaluarea fezabilității funcționării sistemelor solare fotoelectrice și identificarea celor mai bune soluții de integrare a acestora în sistemul energetic.
***Keywords*:** predicție, iradianță solară, prognoză, clusterizare, parametri de eroare statistică, model de predicție, matrice de confuzie, analiza scenariilor.

Долгосрочный прогноз солнечной иррадиации

Брага Д.[a], Кикко Ж.[b], Голованов Н.[c], Порумб Р.[c]
Технический Университет Молдовы[a], Туринский Политехнический Университет[b], Бухарестский Политехнический Университет[c]
Кишинев, Республика Молдова[a], Турин, Италия[b], Бухарест, Румыния[c]

Abstract. В последнее десятилетие значительно увеличилось число солнечных фотоэлектрических установок, подключённых к электроэнергетическим системам по всему миру. В то же время, солнечная иррадиация представляет собой непостоянную величину. Поэтому, для эффективного управления существующими и нововведёнными в эксплуатацию мощностями фотоэлектрических солнечных систем и эксплуатации электроэнергетических систем в нормальном режиме требуется точная модель для прогноза солнечной иррадиации. Главной целью данной работы является разработка и проверка модели для выполнения долгосрочного прогноза солнечной иррадиации. Предложенная цель в данной работе достигается с помощью метода кластеров и математической статистики. Разработка модели предусматривает этап подготовки исторических данных для прогнозирования и этап обработки и анализа результатов, полученных в процессе кластеризации. Использующиеся исторические данные включают солнечную иррадиацию и степень покрытия небесного свода облаками. Процесс подготовки исходных данных включает процесс би-нормализации и кластеризации солнечной иррадиации и периода солнечного свечения, а обработка результата – процесс денормализации. Точность предложенной модели проверяется с помощью стандартных индикаторов ошибок и матрицы путаницы. Для сравнения были проанализированы четыре варианта модели, которые отличаются подходом и подготовки исходных данных для моделирования. Сравнение этих вариантов моделей показало, что для получения более высокой точности необходимо учитывать сезонные особенности солнечной иррадиации.

Главным результатом работы является полученная модель, которая может быть использована для прогнозирования солнечной иррадиации с приемлемой точностью для этого типа прогнозирования и получения последовательности типа дней для различных годовых сценариев. Ценность модели состоит в том, что эти сценарии могут быть использованы для оценки эффективности работы фотоэлектрических солнечных систем и нахождения наилучших решений для их интеграции в электроэнергетическую систему.

***Keywords*:** предсказание, солнечная иррадиация, кластеризация, ошибки прогнозирования, матрица путаницы, анализ сценариев.

I. INTRODUCTION

During the last years, all around the world electricity generation from renewable energy sources (RES) has been increasing constantly. These increases led to transformation of energy systems from highly centralized systems with classical large power plants to systems with a growing number of territorial distributed small plants based on RES [1,2]. The most significant increase of installed capacity concerns photovoltaic (PV) sources: from 22.8 GW at the end of 2009 to 480.6 GW at the end of 2018 [3].

The PV source is highly intermittent and depends on meteorological and climate conditions, such as solar irradiance, cloudiness, air temperature, air humidity, etc. The intermittency poses difficulties in grid management with raising rate of electricity system penetration by solar PV systems, and represents great challenges for PV power generation forecasting [4]. In particular, solar irradiance is the main feature to be considered in short-term PV power forecasting [5] carried out by using different numerical techniques [6] and various methods for unsupervised and supervised learning [7]. Some review papers indicate the current trends in the different time horizons established to perform forecasting, partitioned in [8] into very short-term (from seconds to minutes, up to one hour), short-term (from one hour to one week), medium-term (from one week to one month), and long-term (from one month to one year) [9].

Thus, the task of solar irradiance forecasting is a crucial aspect for ensuring grid stability, reliability and efficient management of the existing or new RES power capacity. Without accurate prediction, it is difficult to promote adequate practices in energy production, transportation and transactions, and this fact conducts to the reduction of energy system efficiency and reliability [4,6].

However, meteorological conditions depend on period of day, season and year, and represent highly varying series of data [11], again making long-term forecasting not easy to be carried out [1,9,10].

During the last decade, different forecasting techniques have been developed. The major commonly used forecasting techniques are

persistence methods, statistical methods, physical methods, and hybrid methods [4,12].

Persistence method is the simplest method, but with too low accuracy. This method assumes that weather conditions at the certain moment in the future will be the same as it is when the forecasting is executed.

Statistical methods (including neural networks) are mathematical models that use the historical data to perform forecasting for next periods. These methods are good for short-term predictions due the fact that with increasing of the forecasting period the errors are increasing. The classical statistical techniques are defined by considering the data as a time series [12,13].

Physical methods take into account the physical aspects like topography, altitude, obstacles sheltering, atmospheric conditions etc. Often these methods are more accurate than other methods. They offer very good accuracy for long time horizons, but appliance of these methods require initial data of good quality [13,14]. The most used physical method is the Numerical Weather Prediction (NWP) model. This complex mathematical model usually requires to be run on super computers, which limits the usefulness of these models to very short time operation of power system.

The most common and effective method is the hybrid method, which represents a combination between individual techniques and permits to improve forecasting accuracy comparing with applications of standalone methods benefiting from the advantages of each model [4,7]. Conceptually, hybrid methods represent a multi-stage approach to forecasting, which applies different techniques at each stage [13,14]. For instance:

- Satellite-imaging and Artificial Neural Networks (ANN) for predictions of global solar irradiance on the horizontal surface for temporal horizons between 30 and 120 minutes [15];
- Satellite-imaging and Support Vector Machine (SVM) for intra-day predictions (in the range of 15 to 300 minutes) [16];
- Satellite-imaging, Exponential Smoothing State (ESS) and back propagated Multi-Layer Perceptron (MLP) model for hourly predictions [17];
- Autoregressive integrated moving average (ARIMA) and ANN [18];
- ANN and Clear sky model [19]
- NWP and ANN [20]; etc.

The next sections present a hybrid model based on time-series, Clear-sky and k-Nearest Neighbors (k-NN) methods for long-term solar irradiance forecasting on the horizontal surface.

II. CLUSTERING OF THE DAY TYPES AND FORECASTING MODEL

The main goal of this paper is the creation of a model for long-term solar irradiance forecasting. As initial data for the solar forecasting model, historical hourly data are used for the period of 1951 – 1990 for Chisinau (the capital of Republic of Moldova, emplaced geographically in the central part of country), and the data regarding weather features (sky nebulosity and temperature) obtaining from meteorological station from Chisinau for period of one year period (2018 – 2019).

For long-term forecasting, it is essential to predict the daily or weekly amount of generated energy. Theoretically, the daily solar irradiance in clear sky conditions is distributed in accordance with the Moon-Spencer model [21].

This model used the Sun position on the sky with respect to the daytime and year time. Solar Irradiance in this model includes Global Horizontal Irradiance (GHI), Direct Normal Irradiance (DNI) and Horizontal Diffuse Irradiance (HDI) [10,22].

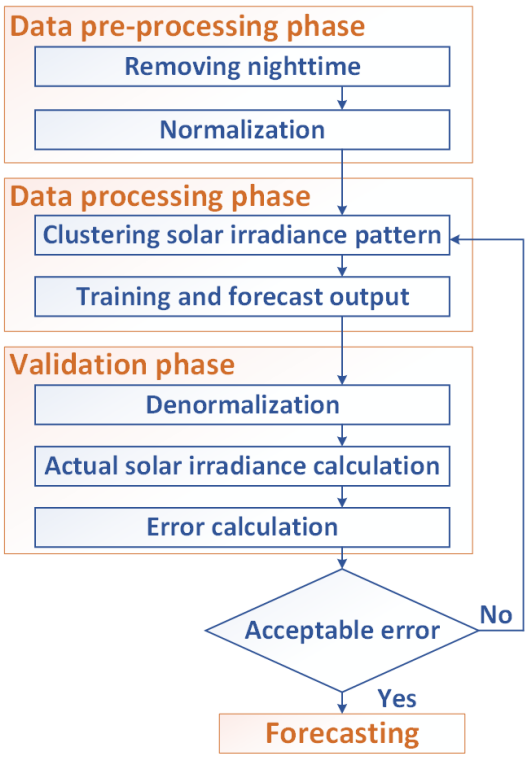


Fig. 1. Proposed solar irradiance forecasting model flowchart.

GHI data is the main component considered during forecasting in this paper. Besides historical GHI data, GHI in Clear-sky conditions represents the maximum GHI that can be received by PV systems during a clear sky day. GHI in Clear-sky conditions is constant for the same period of the year. Additional data required for the proposed model are cloud coverage of the sky, air temperature and humidity.

The development of the proposed solar irradiance forecasting model includes three phases: data pre-processing phase, data processing phase, and validation phase (Fig. 1).

A. Data pre-processing phase

The first phase of the proposed forecasting model includes removing nighttime, and data normalization. Removing nighttime supposes excluding the period between sunset time of previous day and sunrise time of the day considered.

Considering that the solar irradiance is a function of the sunshine period, which depends on the year period, it is difficult to compare solar irradiance characteristics for days with considerable different sunshine period (Table 1). For clear-sky conditions the differences among solar irradiance values in sunshine periods can be observed in Fig. 2.

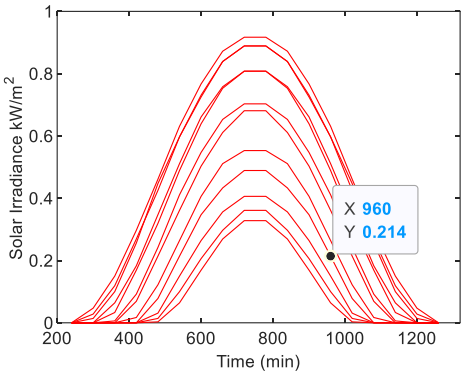


Fig.2. Solar irradiance in clear-sky condition.

To allow the comparison between solar irradiance data from different period of the year, it is necessary to normalize the solar irradiance data and the daily sunshine period (this process is called bi-normalization).

The bi-normalization consists of the representation of the solar irradiance and sunshine time in relative units, both with values between 0 and 1 [10]. For this purpose, the daily sunshine period for each day was limited from the sunrise to the sunset time periods for all days. Then, to represent the solar irradiance time series with the same number of points, the available data points are used for data alignment within an interpolation process [23] to obtain the same number of points (20 points in this paper) at the same locations onto the normalized horizontal axis. On the vertical axis, the daily solar irradiance was divided by GHI in Clear-sky conditions (the maximum solar irradiance) for this period [10,13].

For simplifying comparison between types of days in dependence of solar irradiance, for example sunny days in summer and winter, the normalization of solar irradiance was done separately for each month.

Table 1

Sunrise and sunset time and daytime hours

Day	Sunrise	Sunset	Sunshine period
20-Mar	06:07	18:15	12 h and 08 min
21-Jun	04:08	20:02	15 h and 54 min
23-Sep	05:51	18:00	12 h and 09 min
21-Dec	07:47	16:17	8 h and 30 min

In order to prepare data for clustering of daily solar irradiance and obtaining better accuracy of clustering, three types of patterns have been proposed:

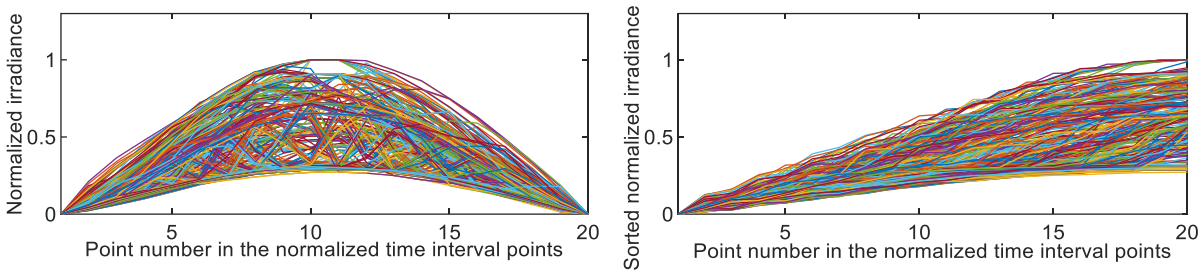
1. Normalized solar irradiance patterns NP;
2. Sorted normalized solar irradiance patterns SNP;
3. Differences between normalized solar irradiance patterns DNP.

For creating NP, the data regarding solar irradiance and sunshine time were normalized according to the procedure described above [10].

For creating SNP, the normalized solar irradiance data have been sorted in the ascending order. The DNP have been determined by considering two representative days: one for clear sky conditions, and one for cloudy sky conditions. Beginning from the irradiance features of these two days, for each day the normalized solar irradiance differences were calculated and sorted in ascending order.

The results of normalization of the solar irradiance and sunshine period are shown in Fig. 3. In particular, the available data are represented in bi-normalized form in Fig. 3a, and are sorted in the ascending order in Fig. 3b.

For clustering, it has been used the *k*-means method with the help of Classification Learner tool in MATLAB®, which carries out the daily solar irradiance pattern grouping into *K* exclusive clusters (groups).



a) Normalized solar irradiance patterns **b) Sorted normalized solar irradiance patterns**

Fig.3. Solar irradiance patterns

For the choice of the number of clusters, it is possible to consider the results of a parametric analysis by changing K, or to set up the number K according with a practical criterion. In the example shown in [10] for $K = 12$, the clustering results are fine, but it is not immediate to give a practical meaning to all the clusters on the basis of the results; for example, two clusters contain solutions close to the clear sky conditions, and the differences among the clusters are progressively lower. Conversely, with a smaller number of clusters it is easier to identify the type of days from practical considerations. In this paper, the chosen number of clusters is $K = 4$, with a practical meaning of having a simple categorization of the days into clear, mostly clear, mostly cloudy, and cloudy.

The results of the *k*-means clustering with $K = 4$ are shown in Fig. 4 for NP, Fig. 5 for SNP, and Fig. 6 for DNP. These results confirm the partitioning of the days into clear (cluster 1), mostly clear (cluster 2), mostly cloudy (cluster 3), and cloudy (cluster 4). The effectiveness of the choice $K = 4$ has been checked by repeating the *k*-means clustering with different number of clusters and tracking two clustering performance indicators, namely, the sum of the Euclidean distances between centroids for each cluster (the lower, the better), and the silhouette values (the higher, the better). Fig. 7 shows the results. The performance indicators for the practical solution chosen are relatively good, and are acceptable with respect to the higher difficulty of interpretation that would occur with higher numbers of clusters.

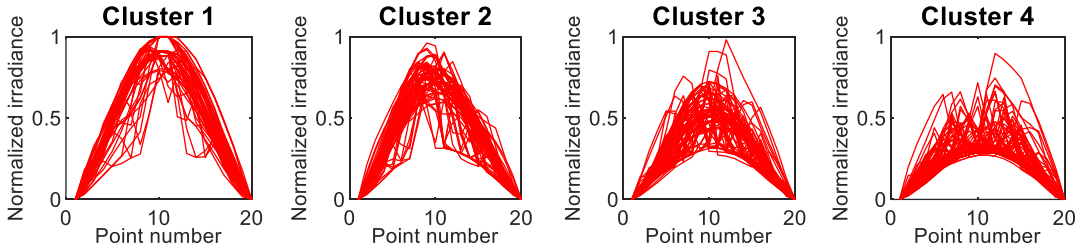


Fig.4. Clustering results based on normalized irradiance patterns

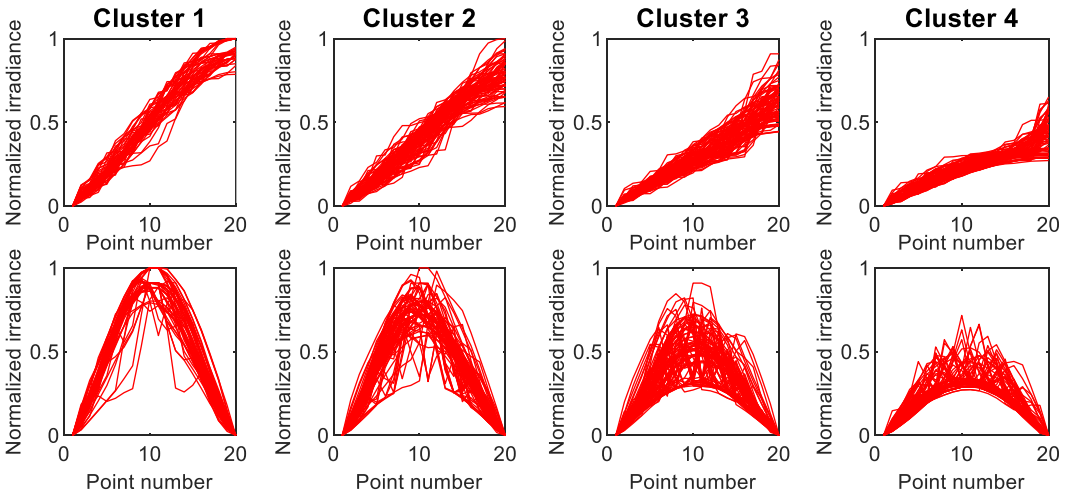


Fig.5. Clustering results based on the sorted normalized irradiance and corresponding normalized irradiance patterns

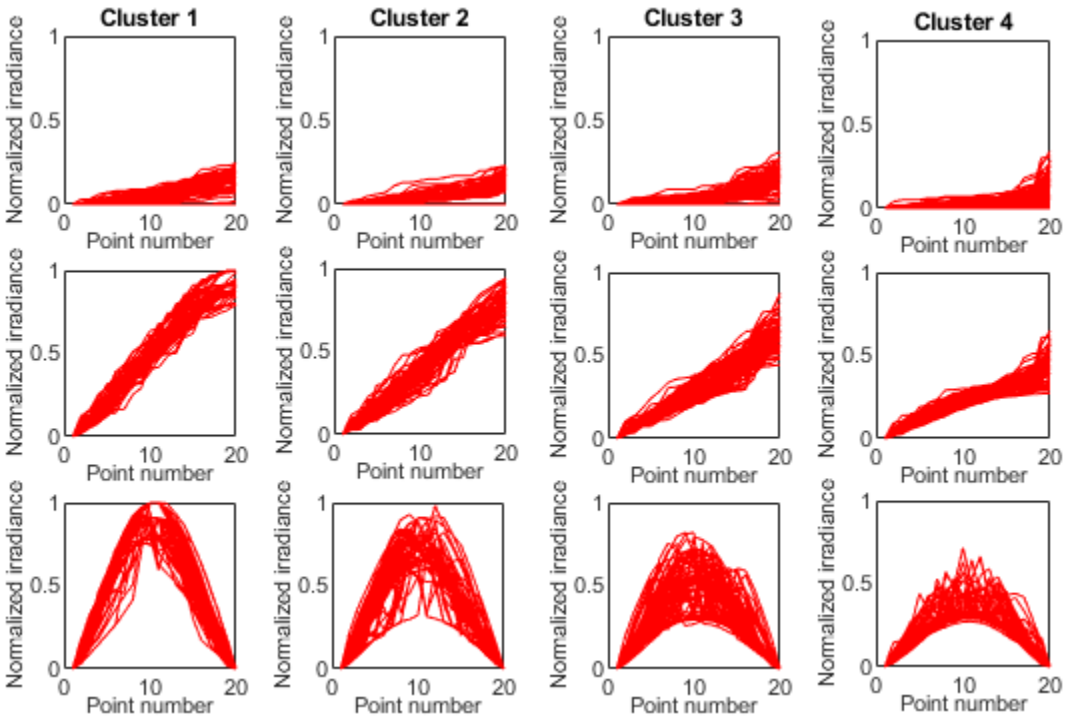


Fig.6. Clustering results based on the differences of the sorted normalized irradiance patterns and sorted and normalized irradiance patterns

Table 2

Succession of day type after knowing the type of preceding day

Type of data	Day type	Preceding day				
		Clear	Mostly clear	Mostly cloudy	Cloudy	Total
NP	Clear	22	16	7	10	55
	Mostly clear	12	46	16	15	89
	Mostly cloudy	11	11	16	37	75
	Cloudy	10	16	36	84	146
SNP	Clear	17	18	9	7	51
	Mostly clear	14	37	22	14	87
	Mostly cloudy	16	19	22	36	93
	Cloudy	4	13	40	77	134
DNP	Clear	20	21	11	7	59
	Mostly clear	14	33	19	15	81
	Mostly cloudy	17	18	20	34	89
	Cloudy	8	9	39	80	136

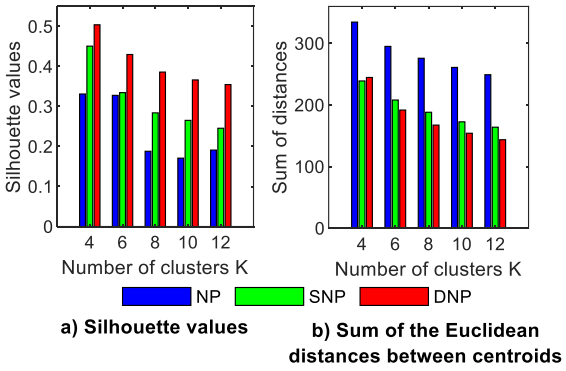


Fig. 7. Clustering performance for different numbers of clusters *K*. Better performance occurs for higher silhouette and lower sum of distances.

Table 3

The dissimilitude between three types of clustering approaches

Day type	Dissimilitude between types of clustering					
	NP vs. SNP		NP vs. DNP		SNP vs. DNP	
Clear	-4	1.1%	4	1.1%	8	2.2%
Mostly clear	-2	0.5%	-8	2.2%	-6	1.6%
Mostly cloudy	18	4.9%	14	3.8%	-4	1.1%
Cloudy	-12	3.3%	-10	2.7%	2	0.5%

Table 2 and Table 3 show the results and the dissimilitude between clustering with the three types of input data. It can be observed that the differences are not significant, but after visual assessment of solar irradiance pattern clusters, it can be concluded that the SNP and DNP approaches are more accurate.

In order to use the clustering results for solar irradiance forecasting, it was used the probability of the succession of days belonging to each cluster during the year in dependence of the type of preceding days (Table 4). In the columns there are the types of preceding days, and in the rows the following types of days and their probabilities of occurrence.

Table 5 shows an example of probability of succession type of day with known type of two preceding days.

B. Data processing phase

Data training consists of the simulation of a one-year day-by-day succession, with the scope of forecasting solar irradiance features for that year. The baseline for simulation is given by the probability of type of day successions, together with the average number of days of each cluster per year. Data training was carried out by using the available Matlab® application.

Forecasting is performed by using four models. Model 1 takes into account only one preceding day, without considering seasonal information. Model 2 takes into account the two preceding days, again without considering seasonal information.

Table 4

Probability of finding a given day type after knowing the type of the preceding day (for Model 1)

Type of data	Day type	Preceding day			
		Clear	Mostly clear	Mostly cloudy	Cloudy
NP	Clear	40%	18%	9%	7%
	Mostly clear	22%	52%	21%	10%
	Mostly cloudy	20%	12%	21%	25%
	Cloudy	18%	18%	48%	58%
SNP	Clear	33%	21%	10%	5%
	Mostly clear	27%	43%	24%	10%
	Mostly cloudy	31%	22%	24%	27%
	Cloudy	8%	15%	43%	57%
DNP	Clear	34%	26%	12%	5%
	Mostly clear	24%	41%	21%	11%
	Mostly cloudy	29%	22%	22%	25%
	Cloudy	14%	11%	44%	59%

Table 5

Probability of finding a given day type after knowing the types of two preceding days (for Model 2)

1st preceding day	**Clear**	**Mostly clear**	**Mostly-cloudy**	**Cloudy**	**Clear**	**Mostly clear**	**Mostly-cloudy**	**Cloudy**
2nd preceding day	**Clear**	**Clear**	**Clear**	**Clear**	**Mostly clear**	**Mostly clear**	**Mostly clear**	**Mostly clear**
Clear	35.3%	33.3%	55.6%	0.0%	35.7%	18.9%	22.7%	7.1%
Mostly clear	35.3%	27.8%	11.1%	28.6%	50.0%	43.2%	40.9%	35.7%
Mostly-cloudy	23.5%	22.2%	33.3%	71.4%	14.3%	27.0%	13.6%	28.6%
Cloudy	5.9%	16.7%	0.0%	0.0%	0.0%	10.8%	22.7%	28.6%
1st preceding day	**Clear**	**Mostly clear**	**Mostly-cloudy**	**Cloudy**	**Clear**	**Mostly clear**	**Mostly-cloudy**	**Cloudy**
2nd preceding day	**Mostly-cloudy**	**Mostly-cloudy**	**Mostly-cloudy**	**Mostly-cloudy**	**Cloudy**	**Cloudy**	**Cloudy**	**Cloudy**
Clear	12.5%	21.1%	4.5%	5.6%	0.0%	15.4%	5.0%	3.9%
Mostly clear	25.0%	31.6%	18.2%	22.2%	75.0%	7.7%	22.5%	1.3%
Mostly-cloudy	25.0%	21.1%	27.3%	22.2%	0.0%	53.8%	20.0%	27.3%
Cloudy	37.5%	26.3%	50.0%	50.0%	25.0%	23.1%	52.5%	67.5%

Table 6

Probability of finding a given day type after knowing the type of one preceding day (for Model 3)

Day type	Preceding day			
	Clear	Mostly clear	Mostly cloudy	Cloudy
Cold time period				
Clear	12.5%	20.0%	4.8%	4.2%
Mostly clear	0.0%	0.0%	9.5%	2.1%
Mostly cloudy	75.0%	20.0%	28.6%	22.9%
Cloudy	12.5%	60.0%	57.1%	70.8%
Warm time period				
Clear	38.6%	19.8%	13.7%	7.9%
Mostly clear	31.8%	45.7%	35.3%	31.6%
Mostly cloudy	22.7%	22.2%	19.6%	36.8%
Cloudy	6.8%	12.3%	31.4%	23.7%

Model 3 takes into account one preceding day and the probabilities are divided into two time periods (i.e., the cold period from November 1st to March 31st, and the warm period for the rest of the year). Finally, Model 4 takes into account one preceding day and the probabilities are divided into the four seasons. Another model could take into account two preceding days and the probabilities divided into the four seasons. However, the forecasting with this model would be quite problematic, due the fact that it would require two or four matrices with probabilities for each seasons with 64 cells each, and most of them with null values, leading to a rather impractical modeling. For this reason, this forecasting model was not applied.

In models 1, 3 and 4, at the initial stage of modeling one day preceding the ”forecast year” is extracted. Depending on the type of this day and of the probability of the following type of day, it is determined the type of the next day. Then, the types of the next days are determined in the same manner, with remark that the type of preceding day is taken as the type of the following day determined at the previous stage (it was not necessary to introduce manually the type of preceding day). This process continues until the simulation of all the days of the year.

Forecasting with two preceding days distinguishes from the previous variants of forecasting by taking into account the type of the two preceding days (not only one preceding day).

Table 7

Probability of finding a given day type after knowing the type of one preceding day (for Model 4)

Day type	Preceding day			
	Clear	Mostly clear	Mostly cloudy	Cloudy
Winter				
Clear	0.0%	20.0%	0.0%	3.8%
Mostly clear	0.0%	0.0%	10.0%	2.5%
Mostly cloudy	75.0%	20.0%	30.0%	22.5%
Cloudy	25.0%	60.0%	60.0%	71.3%
Spring				
Clear	33.3%	27.3%	5.0%	9.5%
Mostly clear	22.2%	18.2%	20.0%	14.3%
Mostly cloudy	33.3%	18.2%	25.0%	47.6%
Cloudy	11.1%	36.4%	50.0%	28.6%
Summer				
Clear	46.2%	17.5%	16.7%	0.0%
Mostly clear	30.8%	52.6%	43.3%	60.0%
Mostly cloudy	15.4%	24.6%	23.3%	40.0%
Cloudy	7.7%	5.3%	16.7%	0.0%
Autumn				
Clear	16.7%	28.6%	23.1%	9.1%
Mostly clear	33.3%	35.7%	15.4%	13.6%
Mostly cloudy	50.0%	14.3%	7.7%	18.2%
Cloudy	0.0%	21.4%	53.8%	59.1%

At the first stage, the types of these days are taken in accordance with the types of two days preceding the period of interest (Table 5). In the next stages, the types of preceding days are taken as the types of following days in the previous stage.

In Model 3 and Model 4 the probabilities of following a given type of day after the certain type of preceding day were extracted from Table 6 and Table 7, respectively, in accordance with the particularities of each season.

C. Validation phase

This phase includes de-normalization of the solar irradiance resulted from the forecasting process, calculation of the actual solar irradiance, and validation of forecasting.

De-normalization represents the opposite process of normalization, i.e., representation of

forecasting solar irradiance and time in natural units kW/m^2 and, respectively, hours. For obtaining the actual solar irradiance, the result of forecasting (in relative units) is multiplied by solar irradiance in Clear-sky conditions for the respective period of time [9,13].

In order to validate the forecasting results, the confusion matrix is constructed, and common errors used for accuracy assessment of forecasting (root mean square error, average of the errors, mean absolute error and mean absolute percentage error) are calculated.

Confusion Matrix

The Confusion matrix is a summary of prediction results and shows the ways in which the classification model performs in predictions.

The number of correct and incorrect predictions are summarized with counting values and broken down by each cluster. It shows not only the errors, but more importantly the types of errors made. The columns represent the predicted types of days, and the rows the actual types of days.

With the view of accuracy calculation for each predicted cluster, the data from the confusion matrix (Table 9) are classified as:

- True Positives (TP): placed in the top left cell, represents the data rows (type of day) belonging to the positive class (i.e., clear) and correctly classified as such;
- False Negatives (FN): placed in the first row at the right side of the TP cell, represents the data rows (type of day) belonging to the positive class (i.e., clear) and incorrectly classified as negative (i.e., mostly clear, mostly cloudy or cloudy);
- False Positives (FP): placed in the first column below the TP cell, represents the data rows (type of day) belonging to the negative class (i.e., mostly clear, mostly cloudy or cloudy) and incorrectly classified as positive (i.e., clear);
- True Negatives (TN): placed in rows 2 – 4 and columns 2 – 4, represents the data rows (type of day) belonging to the negative class (i.e., mostly clear, mostly cloudy or cloudy) and correctly classified as such.

Overall statistics summarize the accuracy of the forecasting model, represented by Overall Accuracy and the Overall Error. The Overall Accuracy of forecasting model is determined as the ratio of true predicted type of days to total number of days:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

The Overall error is determined as the ratio between the false predicted type of days and the total number of days:

$$Error = \frac{(FP+FN)}{(TP+FP+FN+TN)} \quad (2)$$

The class statistics summarizes the class performance for the positive class and the negative class, separately.

Sensitivity shows the capability of the model to detect positive classes. So if Cluster 1 is a positive class, the Sensitivity quantifies how many actual clear days are predicted correctly as clear. The Sensitivity is evaluated as:

$$Sensitivity = TP/(TP + FN) \quad (3)$$

Specificity shows the accuracy of assignment to the positive class:

$$Specificity = TN/(TN + FP) \quad (4)$$

Recall shows the ratio of the total number of days correctly classified as positive:

$$Recall = TP/(TP + FN) \quad (5)$$

Precision shows the capability of the model to assign positive events to the positive class:

$$Precision = TP/(TP + FP) \quad (6)$$

Recall and *Precision* are interconnected. If a stricter filter is used, it is increased the number of days reported correctly as Clear days, but at the same time is increased the number of days of other types reported incorrectly as Clear days. And vice versa, a less strict filter leads to increasing the number of Clear days reported incorrectly as days of other types. Often it is used the *F-measure*, which is the harmonic mean of *Recall* and *Precision*:

$$F - measure = 2\frac{Recall*Precision}{Recall+Precision} \quad (7)$$

Errors calculation

For quantitative estimation of forecasting, statistical methods are used. The estimation error ε is defined as the difference between the forecast irradiance I_{for} and actual irradiance I_{act}:

$$\varepsilon = I_{for} - I_{act} \quad (8)$$

The positive value of ε appears when the solar irradiance is overestimated, and vice versa, the negative value appears when the forecasting solar irradiance is underestimated.

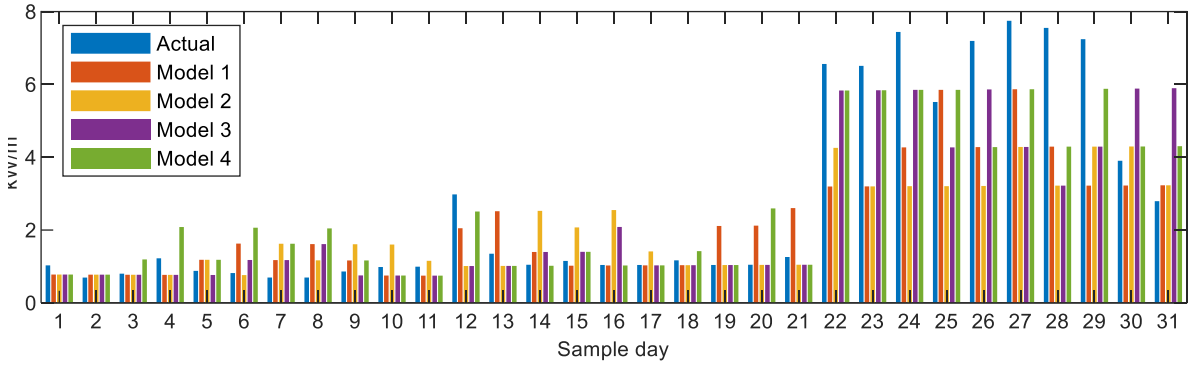


Fig. 8. Sample results of forecasting of solar irradiance

Table 8

Sample results of forecasting of the types of days

Day	Actual	Forecast			
		Model 1	Model 2	Model 3	Model 4
02/11/2017	Mostly cloudy	Cloudy	Cloudy	Cloudy	Cloudy
03/11/2017	Cloudy	Cloudy	Cloudy	Cloudy	Cloudy
04/11/2017	Cloudy	Cloudy	Cloudy	Cloudy	Mostly cloudy
05/11/2017	Mostly cloudy	Cloudy	Cloudy	Cloudy	Clear
06/11/2017	Cloudy	Mostly cloudy	Mostly cloudy	Cloudy	Mostly cloudy
07/11/2017	Cloudy	Mostly clear	Cloudy	Mostly cloudy	Clear
08/11/2017	Cloudy	Mostly cloudy	Mostly clear	Mostly cloudy	Mostly clear
09/11/2017	Cloudy	Mostly clear	Mostly cloudy	Mostly clear	Clear
10/11/2017	Cloudy	Mostly cloudy	Mostly clear	Cloudy	Mostly cloudy
11/11/2017	Mostly cloudy	Cloudy	Mostly clear	Cloudy	Cloudy
12/11/2017	Mostly cloudy	Cloudy	Mostly cloudy	Cloudy	Cloudy
01/02/2018	Clear	Mostly clear	Cloudy	Cloudy	Clear
02/02/2018	Mostly cloudy	Clear	Cloudy	Cloudy	Cloudy
03/02/2018	Cloudy	Mostly cloudy	Clear	Mostly cloudy	Cloudy
04/02/2018	Cloudy	Cloudy	Mostly clear	Mostly cloudy	Mostly cloudy
05/02/2018	Cloudy	Cloudy	Clear	Mostly clear	Cloudy
06/02/2018	Cloudy	Cloudy	Mostly cloudy	Cloudy	Cloudy
07/02/2018	Cloudy	Cloudy	Cloudy	Cloudy	Mostly cloudy
08/02/2018	Cloudy	Mostly clear	Cloudy	Cloudy	Cloudy
09/02/2018	Cloudy	Mostly clear	Cloudy	Cloudy	Clear
10/02/2018	Mostly cloudy	Clear	Cloudy	Cloudy	Cloudy
01/06/2018	Mostly clear	Cloudy	Mostly cloudy	Mostly clear	Mostly clear
02/06/2018	Mostly clear	Cloudy	Cloudy	Mostly clear	Mostly clear
03/06/2018	Clear	Mostly cloudy	Cloudy	Mostly clear	Mostly clear
04/06/2018	Mostly clear	Mostly clear	Cloudy	Mostly cloudy	Mostly clear
05/06/2018	Clear	Mostly cloudy	Cloudy	Mostly clear	Mostly cloudy
06/06/2018	Clear	Mostly clear	Mostly cloudy	Mostly cloudy	Mostly clear
07/06/2018	Clear	Mostly cloudy	Cloudy	Cloudy	Mostly cloudy
08/06/2018	Clear	Cloudy	Mostly cloudy	Mostly cloudy	Mostly clear
09/06/2018	Cloudy	Cloudy	Mostly cloudy	Mostly clear	Mostly cloudy
10/06/2018	Cloudy	Cloudy	Cloudy	Mostly clear	Mostly cloudy

The most common indices presented in the literature are [1,2,13,24-26]:

- The root mean square error (*RMSE*), the most popular error used for forecasting accuracy assessment, calculated as:

$$RMSE = \sqrt{1/N \sum_{i=1}^{N} \varepsilon_i^2} \quad (9)$$

- The average of the errors (*MBE*), defined as the mean difference between forecast and actual irradiance, represents the systematic part of the error:

$$MBE = \bar{\varepsilon} = 1/N \sum_{i=1}^{N} \varepsilon_i \quad (10)$$

- The mean absolute error (*MAE*), more sensitive to high-value errors, is useful in those applications insensitive to minor errors, is defined as the absolute mean difference between forecast and actual irradiance, and represents the systematic part of the error:

$$MAE = 1/N \sum_{i=1}^{N} |\varepsilon_i| \quad (11)$$

- The mean absolute percentage error (*MAPE*), which assesses uniform prediction errors:

$$MAPE = 1/N \sum_{i=1}^{N} \left| \left(I_{act} - I_{for}\right)/I_{act} \right| \quad (12)$$

III. RESULTS AND DISCUSSIONS

A. Results and performance assessment of the prediction of the day types

The proposed forecasting model has been used for forecasting GHI for the period 02 November 2017 – 01 November 2018 for the Chisinau Municipality, Republic of Moldova. The actual type of day and the solar irradiance value that presents real solar features has been compared with the forecast type of day and solar irradiance. An example of this comparison is presented in Table 8 and Fig. 8. The sample has been chosen randomly, and three different periods of the year (autumn, winter and spring) are presented.

For accuracy assessment of the day type prediction, the confusion matrix shown in Table 9 presents the results of comparing the forecast and actual succession of days.

The overall performance of forecasting model and class prediction statistics is presented in Table 10 and Table 11. The overall statistics of the forecasting models shows that Model 4, which takes into account seasonality aspects of solar irradiance, is the most exact.

At the same time, Model 2, which neglects these aspects, is the most inexact model. Thus, for better results it is necessary to take into account probabilities determined per seasons, but taking two preceding days for forecasting was practically ineffective.

Table 9

Confusion Matrix for Clear class (Model 4)

		Forecast			
		Clear	Mostly clear	Mostly cloudy	Cloudy
Actual	Clear	18 (TP)	20 (FP)	13 (FP)	13 (FP)
	Mostly clear	22 (FN)	51 (TN)	17 (TN)	7 (TN)
	Mostly cloudy	10 (FN)	9 (TN)	9 (TN)	31 (TN)
	Cloudy	23 (FN)	25 (TN)	13 (TN)	84 (TN)

Table 10

Overall statistics per forecasting models

Model	Accuracy	Error	Correctly Classified	Incorrectly Classified
1	0.312	0.688	114	251
2	0.238	0.762	87	278
3	0.362	0.638	132	233
4	0.444	0.556	162	203

Table 11

Class statistics (Model 4)

Type of day	*TP*	*FP*	*TN*	*FN*	*Recall*	*Precision*	*Sensitivity*	*Specificity*	*F-measure*
Clear	18	46	246	55	0.247	0.281	0.247	0.842	0.263
Mostly clear	51	46	214	54	0.486	0.526	0.486	0.823	0.505
Mostly cloudy	9	50	263	43	0.173	0.153	0.173	0.840	0.162
Cloudy	84	61	169	51	0.622	0.579	0.622	0.735	0.600

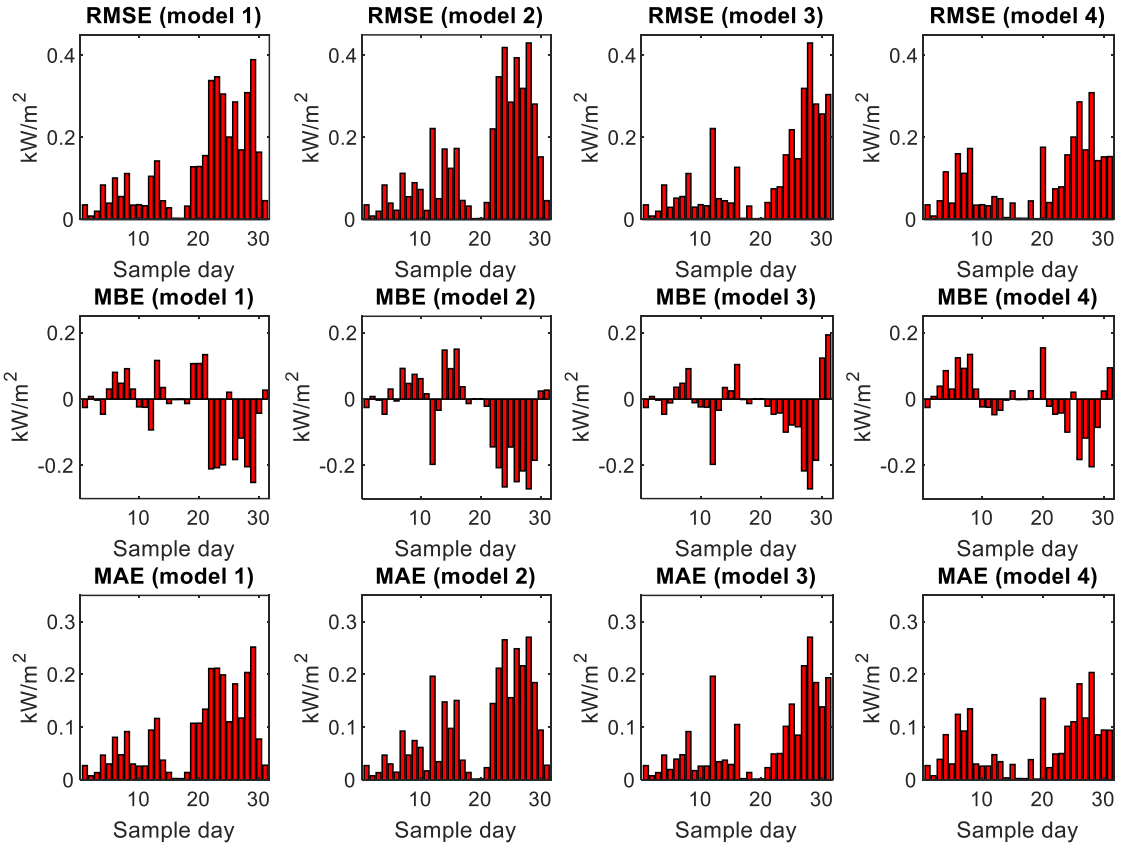


Fig. 10. Error indices per sample days.

The performance of the forecasting model is quantifying by calculation of error samples for the days and for the entire year. The results of the error calculations are presented in Table 12, Fig. 10 and Fig. 11. Error analysis demonstrates that taking into account of the seasonality aspects is crucial for forecasting accuracy. Models 3 and 4 show lower level of error indices. Overall Model 4 has lower MBE than Model 3.

Table 12

Error indices per year

Model	*RMSE* kW/m²	*MBE* kW/m²	*MAE* kW/m²	*MAPE* %
1	0.154	-0.034	0.098	39.8
2	0.157	-0.027	0.103	42.1
3	0.131	-0.017	0.083	33.6
4	0.131	-0.001	0.082	33.3

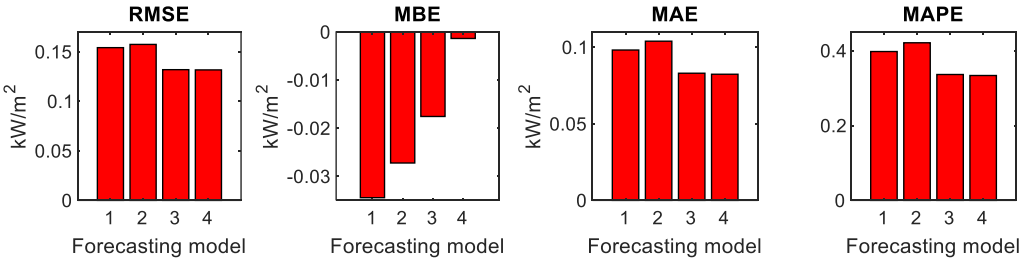


Fig. 11. Error indices per year.

B. Discussion and applications to scenario generation

The results of the comparisons carried out among the four variants indicate that the best performance is obtained by using Model 4, which takes into account seasonality aspects of the solar irradiance. The calculations of the classical errors such as *MAPE* show relatively high values of the errors. However, it has to be considered that these errors are obtained by trying to identify the type of day in a long-term forecasting context. Trying to guess the type of day that will occur in a long-term time horizon is not the appropriate way to proceed, as the uncertainty concerning the future is so high that it is virtually impossible to identify the day type for a single day.

The *MAPE* error itself is a limited metric to address this kind of problem. Better metrics should be found, based on the aggregate behavior of the days in a given period. Indeed, the comparison presented above had only the goal to compare the four models, and to identify Model 4 as the most suitable one. Model 4 is now used to create a mechanism of *scenario generation.* In each scenario, the day types for an entire year are generated by using the information available. Since the process of scenario generation depends on random number extractions, it is possible to construct a large set of scenarios that can be then used for different types of applications in which the definition of multiple scenarios is useful to make a probabilistic characterization of the operation of a PV system.

An example of scenario generation has been executed by constructing 100 scenarios starting from different days of the year, considering the partitioning into the four seasons. The type of day has been chosen randomly with the probability given by the relative occurrence of the types of day in the corresponding season. Fig. 12 shows the results, in the form of the Cumulative Distribution Function (CDF) of the number of days. The number of days found in the 100 scenarios for the day types in the four seasons are included in relatively wide ranges. This confirms

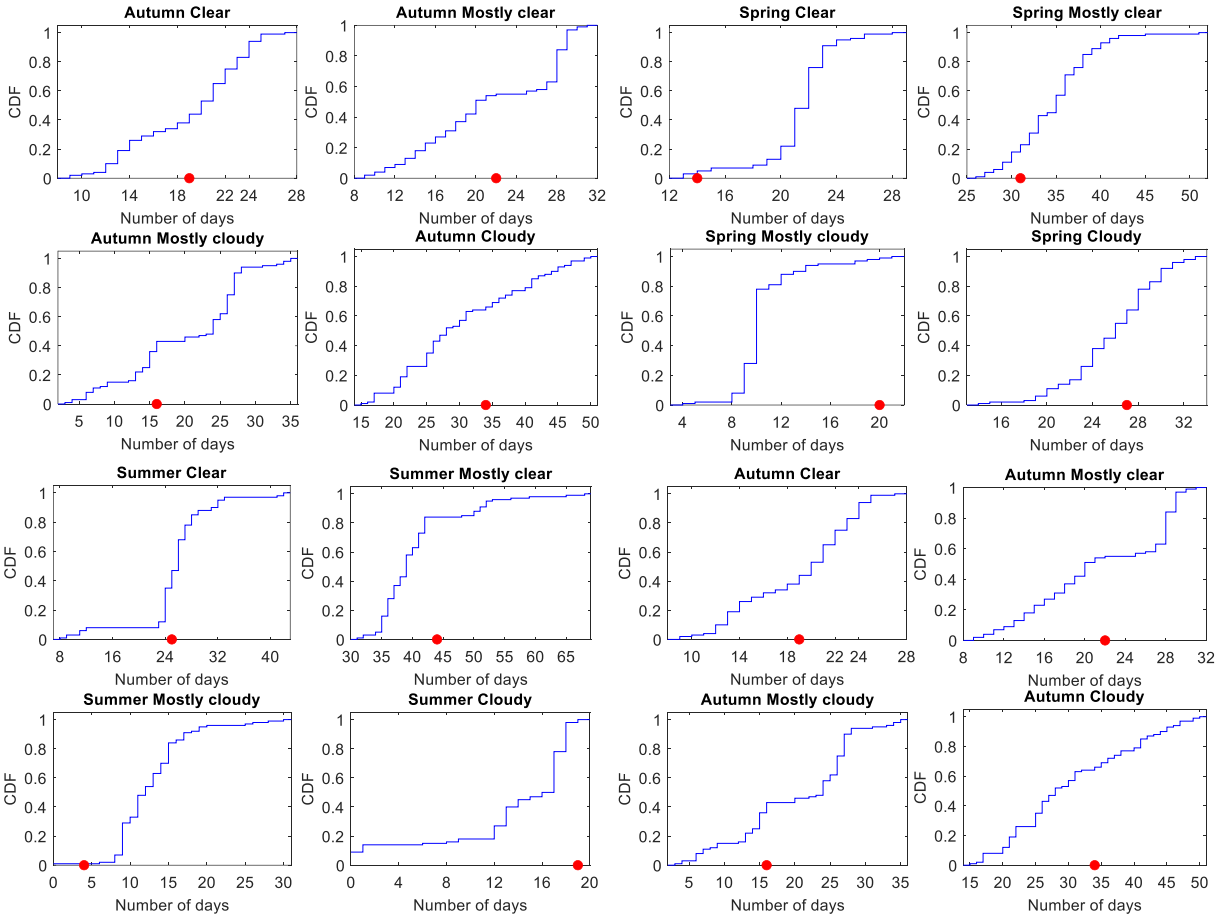
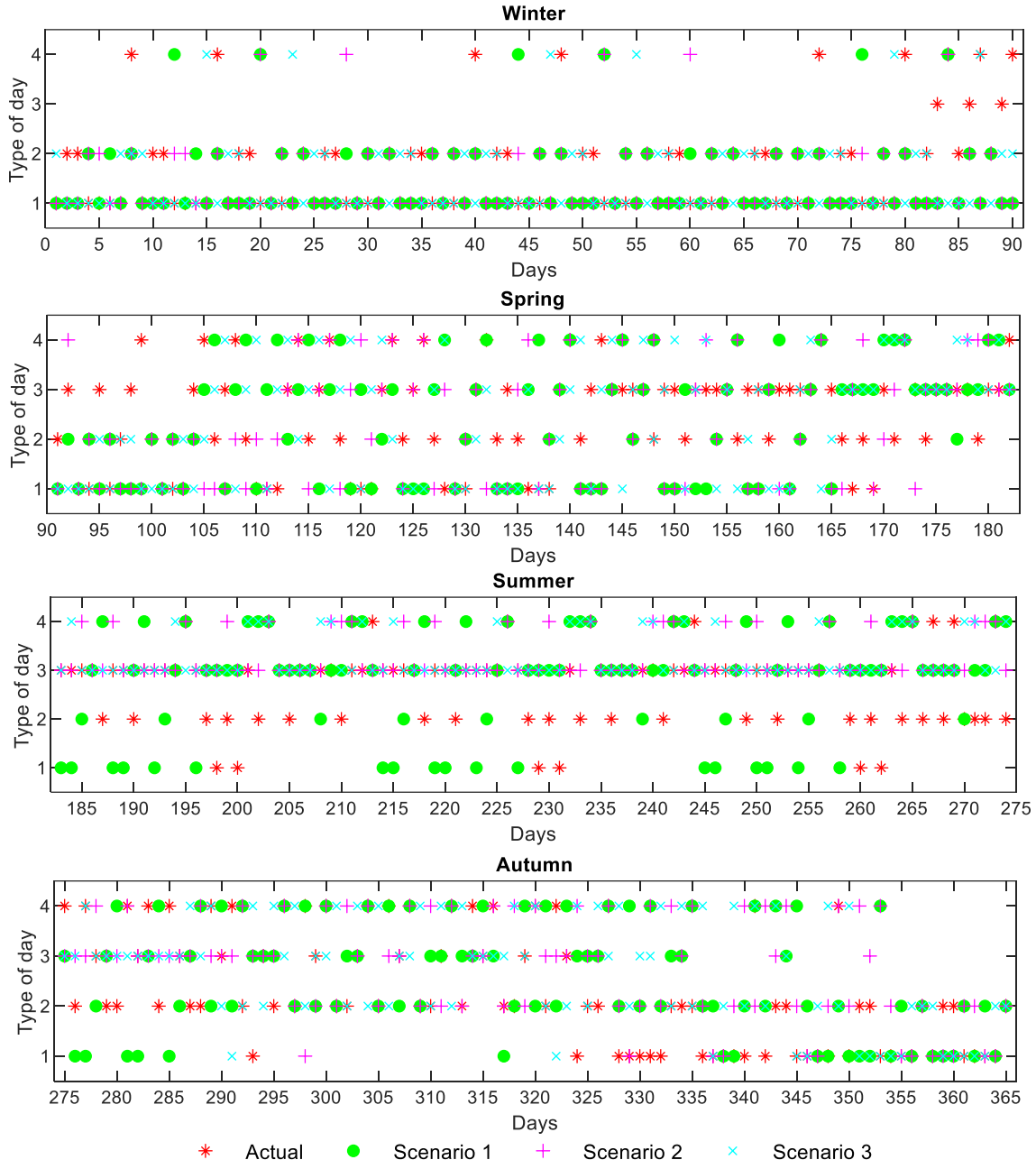


Fig. 12. CDF of the generation of 100 scenarios for one year. The red dots indicate a real situation.



1 – Cloudy; 2 – Mostly cloudy; 3 – Mostly Clear; 4 – Clear

Fig. 13. Comparison between Actual type of day and three different modeled scenarios.

the possibility to generate various scenarios with different numbers of days, following the variability that occurs in real conditions.

To confirm the validity of the ranges obtained, a red dot is positioned on each figure to show the number of days that occurred in a real case for one year.

It can be seen that the real number of days varies considerably, but it is included in the ranges. In any case, the numbers of days are linked together by the fact that the sum of the days in a given season is fixed. Thereby, if more days of a given type appear during one year, the number of days of all the other types will be lower.

A further result concerning the generation of the scenarios is presented in Fig. 13. The actual types of day found in a real situation are superposed to the types of days found in three scenarios arbitrarily taken from the 100 scenarios generated.

It can be seen that the distribution of the types of days is consistent in the various cases, even though the same day can be different in the various scenarios. For example, during the Winter

there are some clear days in all scenarios, at different locations, with a situation that resembles the real case in which the clear days appear occasionally.

Furthermore, there are groups of successive days with similar characteristics in each scenario, which represent what may happen in reality, with a sequence of corresponding days that does not appear regularly every year in the same period. These results confirm the practical usefulness of the proposed way to generate the day type scenarios.

CONCLUSIONS

In recent years the diffusion of solar PV systems grew up considerably. However, solar irradiance has intermittent nature, so that for efficient planning of existing capacities and new capacity to be installed it is extremely important to carry out long-term forecasting with acceptable accuracy. This paper has developed four variants of a forecasting model using clustering method and standard statistical instruments. These models have been compared, showing that for better accuracy it is recommended to use seasonality aspects of the solar irradiance. The most suitable model has then been used to generate a number of scenarios that represent the possible variability of the type of day during one year. These scenarios are useful to carry out any probabilistic analysis in which it is important to incorporate the variability of the type of day during the year.

Further research will aim at enhance the accuracy of the presented model by testing its capabilities in multiple sites with different meteorological conditions.

REFERENCES

[1] Chicco G., Cocina V., Di Leo P., Spertino F., Weather forecast-based power predictions and experimental results from photovoltaic systems, 2014 International Symposium on Power Electronics, Electrical Drives, Automation and Motion, Ischia, 2014, pp. 342-346. doi: 10.1109/SPEEDAM.2014.6872086

[2] Ghofrani M., Suherli A., Time Series and Renewable Energy Forecasting, chapter 12 in Time Series Analysis and Applications, Intech open, 2018. doi:10.5772/intechopen.71501.

[3] IRENA, Renewable Energy Statistics, The International Renewable Energy Agency, Abu Dhabi, 2019.

[4] Akhter M. N., Mekhilef S., Mokhlis H., Mohamed Shah N., Review on forecasting of photovoltaic power generation based on machine learning and metaheuristic techniques, IET Renewable Power Generation, vol. 13, no. 7, 2019, pp. 1009-1023, doi: 10.1049/iet-rpg.2018.5649.

[5] Wang F., Zhen Z., Mi Z., Sun H., Su S., Yang G., Solar irradiance feature extraction and support vector machines based weather status pattern recognition model for short-term photovoltaic power forecasting, Energy and Buildings, vol. 86, 2015, pp. 427-438, doi: 10.1016/j.enbuild.2014.10.002

[6] Wang F., Zhen Z., Wang B., Mi Z., Comparative Study on KNN and SVM Based Weather Classification Models for Day Ahead Short Term Solar PV Power Forecasting. Applied Sciences. 8. 28. 2017, doi: 10.3390/app8010028.

[7] Voyant C,. Notton G., Kalogirou S., Nivet M.L., Paoli C., Motte F., Fouilloy A., Machine learning methods for solar radiation forecasting: A review, Renewable Energy, vol. 105, 2017, pp. 569-582, doi: 10.1016/j.renene.2016.12.095.

[8] Das U.K., Tey K.S., Seyedmahmoudian M., Mekhilef S., Idris M.Y.I., Van Deventer W., Horan B., Stojcevski A., Forecasting of photovoltaic power generation and model optimization: A review, Renewable and Sustainable Energy Reviews, vol. 81, Part 1, 2018, pp. 912-928, doi; 10.1016/j.rser.2017.08.017.

[9] Song Y. et al., Medium and long term load forecasting considering the uncertainty of distributed installed capacity of photovoltaic generation, 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA), Wuhan, China, 2018, pp. 1691-1696. doi: 10.1109/ICIEA.2018.8397982.

[10] Chicco G., Cocina V., Spertino F., Characterization of solar irradiance profiles for photovoltaic system studies through data rescaling in time and amplitude, 2014 49th International Universities Power Engineering Conference (UPEC), Cluj-Napoca, 2014, pp. 1-6. doi: 10.1109/UPEC.2014.6934619.

[11] Sun, Y., Wang, F., Wang, B., Chen, Q., Engerer, N., Mi, Z. Correlation Feature Selection and Mutual Information Theory Based Quantitative Research on Meteorological Impact Factors of Module Temperature for Solar Photovoltaic Systems. Energies , 10, 7, 2017, doi: 10.3390/en10010007.

[12] Yousefi M., Hajizadeh A., Soltani M.N., A Comparison Study on Stochastic Modeling Methods for Home Energy Management Systems, in IEEE Transactions on Industrial Informatics, vol. 15, no. 8, pp. 4799-4808, Aug. 2019. doi: 10.1109/TII.2019.2908431

[13] Alanazi M., Alanazi A., Khodaei A., Long-term solar generation forecasting, 2016 IEEE/PES Transmission and Distribution Conference and Exposition (T&D), Dallas, TX, 2016, pp. 1-5. doi: 10.1109/TDC.2016.7519883

[14] Akhter M.N., Mekhilef S., Mokhlis H., Mohamed Shah N., Review on forecasting of photovoltaic power generation based on machine learning and metaheuristic techniques, IET Renewable Power Generation, vol. 13, no. 7, pp. 1009-1023, 2019. doi: 10.1049/iet-rpg.2018.5649

[15] Perez R., Kivalov S., Schlemmer J., Hemker K., Renné D., Hoff T.E., Validation of short and medium term operational solar radiation forecasts in the US, Solar Energy, vol. 84, Issue 12, 2010, pp. 2161-2172, doi: 10.1016/j.solener.2010.08.014.

[16] Jang H.S., Bae K.Y., Park H., Sung D. K., Solar Power Prediction Based on Satellite Images and Support Vector Machine, in IEEE Transactions on Sustainable Energy, vol. 7, no. 3, pp. 1255-1263, July 2016, doi: 10.1109/TSTE.2016.2535466.

[17] Dong Z, Yang D., Reindl T., Walsh W.M., Satellite image analysis and a hybrid ESSS/ANN model to forecast solar irradiance in the tropics, Energy Conversion and Management, vol. 79, 2014, pp. 66-73, ISSN 0196-8904, doi: 10.1016/j.enconman.2013.11.043.

[18] Kardakos E.G., Alexiadis M.C., Vagropoulos S.I., Simoglou C.K., Biskas P.N., Bakirtzis A.G., Application of time series and artificial neural network models in short-term forecasting of PV power generation, 2013 48th International Universities' Power Engineering Conference (UPEC), Dublin, 2013, pp. 1-6, doi: 10.1109/UPEC.2013.6714975.

[19] Mosaico G., Saviozzi M., A hybrid methodology for the day-ahead PV forecasting exploiting a Clear Sky Model or Artificial Neural Networks, IEEE EUROCON 2019 -18th International Conference on Smart Technologies, Novi Sad, Serbia, 2019, pp. 1-6. doi: 10.1109/EUROCON.2019.8861551.

[20] Lima F.J.L., Martins F.R., Pereira E.B., Lorenz E., Heinemann D., Forecast for surface solar irradiance at the Brazilian Northeastern region using NWP model and artificial neural networks, Renewable Energy, Volume 87, Part 1, 2016, pp. 807-818, doi: 10.1016/j.renene.2015.11.005.

[21] Moon P., Spencer D.E., Illumination from a non-uniform sky, Trans. of the Illumination Engineering Society, vol. 37 (12), pp. 707-7261, 1942.

[22] Batrinu F., Carpaneto E., Chicco G., Gagliano S., Spertino F., Tina G.M., Assessing the performance of photovoltaic sites and grid-connected plants: A study case. Proc. VI World Energy System Conference, Torino, Italy, pp. 386-393, 10-12 July 2006.

[23] Chicco G., Cocina V., Mazza A., Spertino F., Data Pre-Processing and Representation for Energy Calculations in Net Metering Conditions, Proc. IEEE Energycon 2014, Dubrovnik, Croatia, 13-16, May 2014, paper 262.

[24] Alimohammadi S., He D., Multi-stage algorithm for uncertainty analysis of solar power forecasting, 2016 IEEE Power and Energy Society General Meeting (PESGM), Boston, MA, 2016, pp. 1-5. doi: 10.1109/PESGM.2016.7741199

[25] Sobri S., Koohi-Kamali S., Abd Rahim N.. Solar photovoltaic generation forecasting methods: A review. Energy Conversion and Management, vol. 156, pp. 459-497, 2018. doi:10.1016/j.enconman.2017.11.019.

[26] Antonanzas J., Osorio N., Escobar R., Urraca R., Ascacibar F.J., Antonanzas F. Review of photovoltaic power forecasting. Solar Energy, vol. 136, pp. 78-111, 2016. doi:10.1016/j.solener.2016.06.069.

Information about authors.

Dumitru Braga,
PhD student, University Politehnica Bucharest; Lecturer, Technical University of Moldova, Energy and Electrical Engineering Faculty.
E-mail:dumitru.braga@tme.utm.md

Gianfranco Chicco
PhD, Full Professor at Politecnico di Torino, Dipartimento Energia "Galileo Ferraris".
E-mail: gianfranco.chicco@polito.it

Nicolae Golovanov,
PhD, Professor, Power Enginering Faculty, Electrical Power Systems Department,
E-mail:
nicolae_golovanov@yahoo.com

Radu Porumb,
PhD, Associate Professor, Power Enginering Faculty, Electrical Power Systems Department,
E-mail: radu.porumb@upb.ro