Solar power forecasting using machine learning and ECMWF weather variables

Viktória Soós
Department of Energy Engineering, Faculty of Mechanical Engineering, Budapest University of Technology and Economics
Budapest, Hungary
soos.viktoria@eszk.org

Dávid Markovics
Department of Energy Engineering, Faculty of Mechanical Engineering, Budapest University of Technology and Economics
Budapest, Hungary
markovics.david@edu.bme.hu

Martin János Mayer
Department of Energy Engineering, Faculty of Mechanical Engineering, Budapest University of Technology and Economics
Budapest, Hungary
mayer@energia.bme.hu

Abstract—Day-ahead power forecasts play a key role in managing the electricity distribution system, but the weather dependency of renewables is a major challenge. State-of-the-art solar power forecasts are based on numerical weather prediction model outputs. Forecasts of a multitude of weather variables issued by the European Centre for Medium-Range Weather Forecasts are utilized to provide a solid basis for this research. The 31 weather variables, zenith and azimuth angle were first investigated based on correlation coefficients, which resulted in a reduction to 27 variables to avoid multicollinearity. Stepwise regression is used to identify the order of importance of the predictors and to determine the combination of predictors that results in the most accurate forecasts. The post-processing reduced the root mean square error of the global horizontal irradiance (GHI) predictions from 68.1 W/m^2 to 64.4 W/m^2. The corrected GHI forecasts are converted to photovoltaic (PV) power by a physical model chain. The PV power forecasts created from the post-processed GHI input data are more accurate than those created from the raw weather forecasts.

Keywords—machine learning, ECMWF, global horizontal irradiance, feature importance, solar power, PV power forecasting

I. Introduction

To achieve sustainable development goals, the global installed capacity of solar photovoltaic (PV) systems has grown rapidly worldwide over the past decade and is expected to continue to grow. In Hungary, installed capacity has tripled in the last 5 years, currently more than 7.8 GW of PV power is installed [1]. This means that the share of PV in the total installed capacity of the electricity system is more than 35%, which is a significant share in terms of weather dependence [1]. Global horizontal irradiance (GHI) forecasts are essential in predicting the performance of solar power plants, which contributes to the integration of solar power plants into the grid. The grid fluctuations caused by solar power plants pose significant challenges for transmission system operators. The post-processing of GHI forecasts is becoming increasingly important, as these forecasts are the main input to PV power forecasting models. When more accurate solar power forecasts are available, balancing power demand and costs are reduced. The balancing power is usually provided by fast-start gas turbines, which have much higher CO_2 emissions. In order to achieve sustainability objectives and for security of supply reasons, improvements in the accuracy of weather forecasting should be ensured. Power forecasts are relevant for grid operators to optimize dispatching and allocate the necessary reserves to maintain grid stability, and for PV power plant operators participating in the electricity market, because they have to submit generation bids and keep to the planned schedule. If they do not keep to the schedule, they will have to pay a penalty. For this reason, day-ahead power forecasts play a key role in the management of the electricity distribution system.

The weather system is an extremely complex dynamic system. Accurate forecasting of solar radiation or solar energy is challenging because of the complex meteorological processes involved. There are different techniques for obtaining exogenous data that have different temporal and spatial coverage: local sensing, sky camera, remote sensing, satellite, and numerical weather prediction (NWP) model [2][3]. Hourly power forecasts use the latest production data and sky camera data, intraday forecasts use satellite data to detect cloud movement [4]. Day-ahead forecasts most commonly rely on data from an NWP model [4]. NWP forecasts can be obtained, for example, from national meteorological centers. However, these forecasting methods are physics-based, the forecasts produced are subject to bias because the atmosphere is a chaotic system where small errors in the initial state estimation can lead to large forecast errors as forecast lead times increase, and it is therefore challenging to produce accurate forecasts because of our limited ability to describe the initial state of the atmosphere [3][5]. For this reason, even state-of-the-art NWP models remain inaccurate, but their errors are, to some extent, systematic and often exhibit patterns [6]. Therefore, post-processing of forecasts plays a key role in reducing systematic error and increasing forecast accuracy by learning from past NWP forecast errors [6]. Machine learning post-processing methods can be applied to any raw forecast [3]. NWP forecasts are relevant for both intraday and day-ahead markets because of their time scale [7].

The GHI depends on several meteorological variables, such as cloud cover and air temperature near the surface. However, it is also affected by liquid water or aerosols because of their ability to reflect or absorb solar radiation [5]. The error of the NWP model also depends on the predicted weather conditions [6]. Using a larger input set is preferable for models because it can help to better track weather conditions [6]. The choice of predictors has a greater impact on the results than the predictive algorithm because the ranking of models varies from site to site because the algorithm is influenced by local weather and climate [6]. Many researchers use zenith angle, day of year, or time-related variables as inputs, and some also consider, for example, the cloud cover and ambient temperature [6]. However, there are also studies where multiple NWP outputs are used, they do not conduct any manual variable selection, suggesting that using multiple predictors significantly improves the post-processing

979-8-3315-2600-9/25/$31.00 ©2025 IEEE

of NWP forecasts [6]. In the past, solar radiation is considered a less important variable than temperature, humidity, wind, or pressure because they have a much stronger impact on our daily lives [3]. However, with the proliferation and rapid expansion of solar power plants, there is now a strong emphasis on solar radiation forecasting in the energy sector.

In deterministic (or point) forecasting, the forecaster gives a single value for each location and time horizon in the future, providing no information about the forecast uncertainty [5]. However, since most of the practical applications of the forecasts still rely on point forecasts, only deterministic forecasts are considered in this paper.

This paper presents a comprehensive study on post-processing GHI forecasts using a wide range of further NWP forecast variables, and converting GHI values into PV power. Most practitioners use only the few most commonly used predictors from the literature, which limits their forecasting accuracy. This article moves a step forward by testing the added value of 27 predictors to the post-processing of GHI forecasts. The main objective of this research is to improve the accuracy of GHI and PV power forecasts by identifying the weather variables that have a major impact on the post-processing of GHI forecasts.

II. DATA DESCRIPTION

The research in this paper is based on a city in Hungary. Paks is located in the middle of Hungary, on the bank of the Danube, with coordinates of 46.6220° N and 18.8559° E. This area is part of the Great Plain of Hungary, which is the country's largest flat area, but it is also close to the Transdanubian hills. The climate is temperate continental, with the most frequent wind direction being north-west [8]. December is the month with the least sunshine in the Paks area, while May-September are the sunniest months [8]. The summer half-year has almost two and a half times as much sunshine as the winter half-year [8].

The data of a PV plant and a meteorological station are used in this study, located only around 2 km apart, so the comparison of the results gives a representative picture.

A. *PV power plant and observation GHI data*

The PV power data are available for 4 years between 2020 and 2023 with quarterly resolution. Solar power plant data are summarized in Table I.

TABLE I. PARAMETERS OF THE SOLAR POWER PLANT

Angle of inclination	35°
Orientation	south
Nominal DC power	20680.2 kW
Nominal AC power	17244 kW

The GHI measurement dataset of the Paks meteorological station is used. The 10-min observation data are summed into hourly intervals for the same period.

B. *ECMWF data*

ECMWF (European Centre for Medium-Range Weather Forecasts) is a research institute and a 24/7 operational service that produces weather forecasts with a global coverage. They operate a world-class weather forecasting supercomputing facility and have one of the largest meteorological data archives [9]. The Integrated Forecasting System (IFS) is a global-scale NWP forecasting system developed and run operationally by ECMWF at several resolutions. The HRES model is the highest resolution configuration of IFS. It runs four times a day, at 00Z, 06Z, 12Z and 18Z, and each time provides forecasts to 90 hours ahead with an hourly temporal resolution, 9 km horizontal resolution and 137 vertical levels. In addition to hourly resolution forecasts, 3 and 6 hourly forecasts are also produced, covering a forecast horizon of 93-144 and 150-240 hours, respectively. The 3 and 6 hourly forecasts are only available for 00Z and 12Z runs [10][11]. The HRES dataset contains deterministic forecasts with hourly resolution. ECMWF forecasts are considered in literature as the best global forecasts [3][11].

The NWP model provides forecasts for hundreds of weather variables, of which 31 are investigated, plus zenith and azimuth angle. These variables and their units are listed in Table II. The HRES model forecasts are available in hourly resolution for the same interval as PV power and GHI measurement datasets. The *ssrd* variable corresponds to the raw GHI forecast.

TABLE II. WEATHER VARIABLES

Short name	Long name	Unit
p3020	Visibility	m
fdir	Total sky direct solar radiation at surface	W/m^2
dsrp	Surface direct normal solar radiation	W/m^2
par	Photosynthetically active radiation at the surface	W/m^2
cape	Convective available potential energy	W/kg
tclw	Total column cloud liquid water	kg/m^2
sp	Surface pressure	Pa
tcw	Total column water	kg/m^2
tcwv	Total column vertically-integrated water vapor	kg/m^2
sd	Snow depth	m of water equivalent
sshf	Surface sensible heat flux	W/m^2
tcc	Total cloud cover	(0 - 1)
u10	10-meter U wind component	m/s
v10	10-meter V wind component	m/s
t2m	2-meter temperature	K
d2m	2-meter dewpoint temperature	K
ssrd	Surface solar radiation downwards	W/m^2
strd	Surface thermal radiation downwards	W/m^2
ssr	Surface net solar radiation	W/m^2
str	Surface net thermal radiation	W/m^2
e	Evaporation	m of water equivalent
lcc	Low cloud cover	(0 - 1)
mcc	Medium cloud cover	(0 - 1)
hcc	High cloud cover	(0 - 1)
tco3	Total column ozone	kg/m^2
ssrc	Surface net solar radiation, clear sky	W/m^2
tisr	TOA incident solar radiation	W/m^2
tp	Total precipitation	m
si200	200-meter wind speed	m/s
fal	Forecast albedo	(0 - 1)
totalx	Total totals index	K
ghi	Global Horizontal Irradiance	W/m^2

III. METHODOLOGY

In this chapter describes the methodological steps of the research and the logical relationship between the methods. The learning period is the period 2020-2022, the testing period is 2023. To process the data, we use the *scikit-learn* library in Python programming language. This open source machine learning library is used to implement the data-driven post-processing of weather forecast inputs. The machine learning model is used the *Extremely Randomized Trees Regressor* (*Extra Trees*), which uses a collection of decision trees to perform regression tasks [12]. Using this model can be an advantage when analyzing a dataset with many variables.

Machine learning-based models can also solve problems that cannot be represented by explicit algorithms and are thus able to identify characteristic trends even in a very chaotic and changing climate. The use of these models is excellent for solving forecasting and post-processing problems.

A. Post-processing of GHI data

The first step in the post-processing is to examine the correlation coefficient between weather variables. If equal to zero, there is no linear relationship between the two quantities, if equal to 1, they vary together, if equal to (-1), they vary in opposite directions. The correlation coefficient (ρ) can be used to measure the relationship between forecasts and observations and can be calculated using equation (1) [11]. In equation (1), $cov(f,\ o)$ denotes the covariance between forecast and observation, which is the linear relationship between the two quantities. σ_f is the forecast and σ_o is the variance of the observation.

$$\rho = \frac{cov(f,\ o)}{\sigma_f \sigma_o} \tag{1}$$

In the case of a large dataset, not all parameters may carry useful content, the dataset is redundant. When the co-movement of weather variables is statistically significant, multicollinearity can have a number of negative consequences for the results, so the solution may be to omit parameters that are significantly correlated.

Stepwise regression is an iterative method that implements a predictor selection process. It determines which parameters are included in the set of predictors that will be inputs to the model during the iterations. The purpose of this is to find the most important predictors that can be used to estimate the GHI most accurately. In our work, we use forward selection, whereby a new predictor is added to the model in each iteration. The parameter with the lowest RMSE (root mean square error) in the current iteration is added in the iteration. The RMSE values are recalculated in each iteration. In the very first iteration, the model makes 27 forecasts for each weather variable separately. It then performs a verification and ranks the parameters in ascending order based on the RMSE values. The parameter with the lowest RMSE becomes the predictor in the iteration. In the last iteration, the model makes predictions once, using all 27 weather variables as input. The predictor of the first iteration is the most important weather variable. In all iterations, the *Extra Trees* machine learning model is used. A total of 378 cases are considered in the stepwise regression.

B. Physical model chain for irradiance-to-power conversion

The Hungarian electricity market has a balancing clearing resolution of 15 minutes, therefore the AC power forecast is also made with 15-minute resolution [11]. For this purpose, the hourly resolution ECMWF data, GHI measurements, and best GHI forecast have to be converted to 15-minute resolution, for which clear sky interpolation is applied. This method takes into account the daily cycle of the GHI [11].

GHI predictions are converted into PV power by a physical model chain [11]. The *pvlib* Python library is used to model PV performance, which offers reliable, physics-based models for solar position, irradiance transposition, and system output calculation. The modelling process is shown on Fig. 1. As a first step, the zenith angle and azimuth angle are calculated for each quarter hour by solar position algorithm [13]. This algorithm calculated values describing the position of the sun by the location and time parameters.

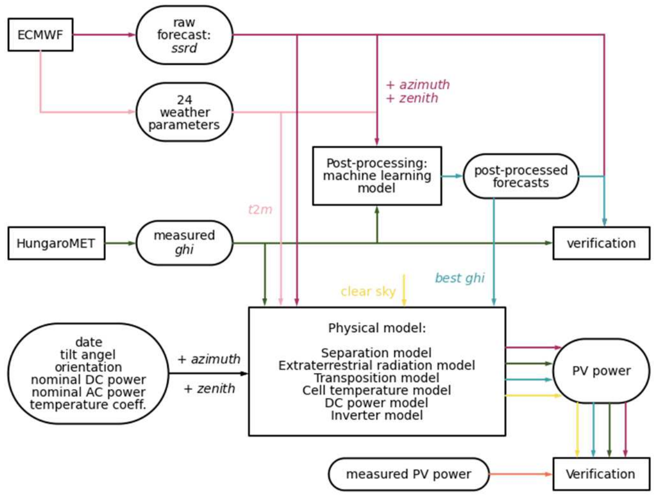

Fig. 1. The process of the research

Then the *Erbs* model is used as the separation model. The *Erbs* model estimates the diffuse fraction from global horizontal irradiance through an empirical relationship [14]. The input parameters are *ghi*, zenith angle and day of the year. The model decomposes *ghi* into *dni* (direct normal irradiance) and *dhi* (diffuse horizontal irradiance) components. The next step is determining extraterrestrial radiation from day of year [15]. To do this, we use the extra radiation function, whose input is just the zenith angle, and whose output is the extraterrestrial radiation presents on a surface which is normal to the sun (*dni extra*). A transposition model is used to determine the total in-plane irradiance and its radiative, sky diffuse and ground reflected components, in effect converting the radiative components to the in-plane PV module [16]. There are several types of sky diffuse models, in this paper the *Perez* model is used because it is widely recognized as the best transposition model [11]. The inputs include panel tilt angle, panel azimuth angle, solar zenith and azimuth angles, *dni*, *dhi*, *ghi*, *dni extra*. As output, we obtain the plane of array irradiance (*poa global*) reaching the PV modules. The temperature model calculates the cell temperature using the empirical heat loss factor model used in PVsyst [17]. The temperature of a PV cell depends not only on the air temperature but also on the irradiance falling on it, since the solar radiation heat directly heats the cells, thus degrading the efficiency of the PV system. The inputs to the model are the global plane of array irradiance and the air temperature, and the output is the cell temperature. As a next step, the *Evans* DC power model is used, which calculates the DC power output of the PV modules [18]. Its inputs include the global plane of array irradiance, the cell temperature, the DC input limit of the inverter and the temperature coefficient of PV power, which latter is typically -0.002 to -0.005 per °C [18], and the value of -0.4%/°C is considered herein. DC power is obtained as the output of the model. A linear loss factor of 10% is applied to take account of all other losses on the DC side. The inverter model calculates the inverter efficiency as a function of the input DC power [19]. It has two additional inputs, the inverter DC input limit and the nominal inverter efficiency, which is 0.97. As output, we obtain the AC power.

The calculations are repeated with four input irradiance data series. The input GHI is replaced by the following data series:

- *ghi*: GHI measurements,
- *ssrd*: surface short-wave (solar) radiation downwards, raw GHI forecast of the ECMWF,
- *best ghi*: the best post-processed GHI forecasts.

The examination of several data sets allows the validation of the physical model based on the data.

The modelling can be further refined, for example the reflection losses due to the module encapsulation can also be modelled [11]. The wind speed is not taken into account when calculating the cell temperature, but it cools the PV modules and thus improves the efficiency of the system. We also do not account for degradation loss, which means that the modules' performance degrades over time. There are two types of degradation: one type is the light-induced degradation (LID), which occurs during the first hours of the exposure to the solar radiation, the other type is the continuous degradation over the whole lifetime [20]. We also do not model shading loss, which is due to the covering of the PV module rows. We also do not estimate losses of cables, transformers and other devices, which can be calculated based on relevant information in the design documentation of PV power plants [11]. It is advantageous if all design parameters in the design documentation are available, because then detailed physical modelling can be implemented. However, physical model chains can be used, and sufficient accuracy can be achieved even if some design data are not known [11]. It is sufficient if only the nominal DC and AC power, tilt and orientation of the PV plant are known [11]. Overall, it can be concluded that the availability of design data does not limit the prediction of PV performance [11].

C. *Verification*

The most comprehensive analyses of deterministic forecasts use the mean bias error (MBE), mean absolute error (MAE), root mean square error (RMSE), the variance ratio (F) and the correlation coefficient (ρ) [11][21]. Equations (1), (2), (3) and (4) show how these error values can be calculated. In these equations, f_i is the i-th GHI prediction, o_i is the i-th observed radiation, and n is the number of data points.

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(f_i - o_i) \tag{2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - o_i| \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - o_i)^2} \tag{4}$$

If the MBE, MAE and RMSE values are normalized, divided by the average of the observed GHI, we obtain relative measures (nMBE, nMAE, nRMSE) [11]. This makes it easier to compare results even when testing other models that have different ranges of values.

The variance ratio (F) between predictions and observations can be calculated using equation (5), which shows how much of the variability in observations is due to the predictions [11]. $\mathbb{V}()$ denotes the variance operator, f the prediction, o the observation. If $F < 1$, $\mathbb{V}(f) < \mathbb{V}(o)$, then we speak of underpredictions, and if $F > 1$, $\mathbb{V}(f) > \mathbb{V}(o)$, then we speak of overpredictions.

$$F = \frac{\mathbb{V}(f)}{\mathbb{V}(o)} \tag{5}$$

TABLE III. THE REFERENCE VERIFICATION VALUES

	Reference verification
MAE [W/m^2]	30.33
RMSE [W/m^2]	68.13
MBE [W/m^2]	5.36
nMAE	0.19
nRMSE	0.44
nMBE	0.03
F	0.98
ρ	0.96

Table III shows the verification of *ssrd*, the raw forecast. This is taken as the reference.

Equations (1), (2), (3), (4), and (5) are used to verify the PV power predictions. In the literature, the accuracy of the forecasting models is characterized by the error between the model predictions and the actual measured power [4].

IV. RESULTS AND DISCUSSION

Plotting the correlation matrix of parameters shows which parameters are significantly correlated with each other. Fig. 2 shows the high correlation coefficient values and parameters. The green color indicates both high unidirectional and bidirectional correlation, while the red color indicates lower correlation.

	fdir	*dsrp*	*par*	*ssrd*	*ssr*	*ssrc*	*tisr*	*zenith*	*tcw*	*tcwv*	*d2m*	*strd*	*t2m*
fdir	1	0.93	0.96	0.97	0.97	0.85	0.84	-0.73	0.13	0.13	0.19	0.20	0.51
dsrp	0.93	1	0.89	0.90	0.90	0.78	0.78	-0.73	0.05	0.05	0.14	0.09	0.45
par	0.96	0.89	1	1.00	1.00	0.95	0.94	-0.83	0.17	0.18	0.22	0.27	0.54
ssrd	0.97	0.90	1.00	1	1.00	0.95	0.94	-0.82	0.16	0.16	0.21	0.26	0.53
ssr	0.97	0.90	1.00	1.00	1	0.95	0.94	-0.82	0.15	0.16	0.20	0.25	0.53
ssrc	0.85	0.78	0.95	0.95	0.95	1	1.00	-0.88	0.18	0.18	0.21	0.32	0.50
tisr	0.84	0.78	0.94	0.94	0.94	1.00	1	-0.90	0.20	0.20	0.23	0.33	0.52
zenith	-0.73	-0.73	-0.83	-0.82	-0.82	-0.88	-0.90	1	-0.24	-0.25	-0.29	-0.37	-0.55
tcw	0.13	0.05	0.17	0.16	0.15	0.18	0.20	-0.24	1	1.00	0.90	0.89	0.79
tcwv	0.13	0.05	0.18	0.16	0.16	0.18	0.20	-0.25	1.00	1	0.91	0.89	0.79
d2m	0.19	0.14	0.22	0.21	0.20	0.21	0.23	-0.29	0.90	0.91	1	0.86	0.86
strd	0.20	0.09	0.27	0.26	0.25	0.32	0.33	-0.37	0.89	0.89	0.86	1	0.84
t2m	0.51	0.45	0.54	0.53	0.53	0.50	0.52	-0.55	0.79	0.79	0.86	0.84	1

Fig. 2. Correlation coefficients of highly correlated parameters

It should be stressed that the different radiation components are significantly correlated with each other, hence their impact on the accuracy of the forecasts is similar, and individually they do not add any additional value to the forecasting process. Due to the multicollinearity and in order to obtain more accurate forecast results, the parameters *par*, *ssr*, *fdir*, *tcwv*, *tisr* and *dsrp* are not considered in the following. These are different solar irradiance (*par*, *ssr*, *fdir*, *tisr* and *dsrp*) or airborne water vapor (*tcwv*) values. This also reduces the dimension of the model input. If the data set are not filtered, false conclusions can be reached.

The RMSE matrix of the stepwise regression shows the order in which the parameters are placed in the set of predictors and the RMSE of each case. The results can be seen in Fig. 3, where the brown color scale is narrow, but there is a steady improvement in the first iterations. In order to better show the small differences between the RMSE values, the range of the color scale is reduced.

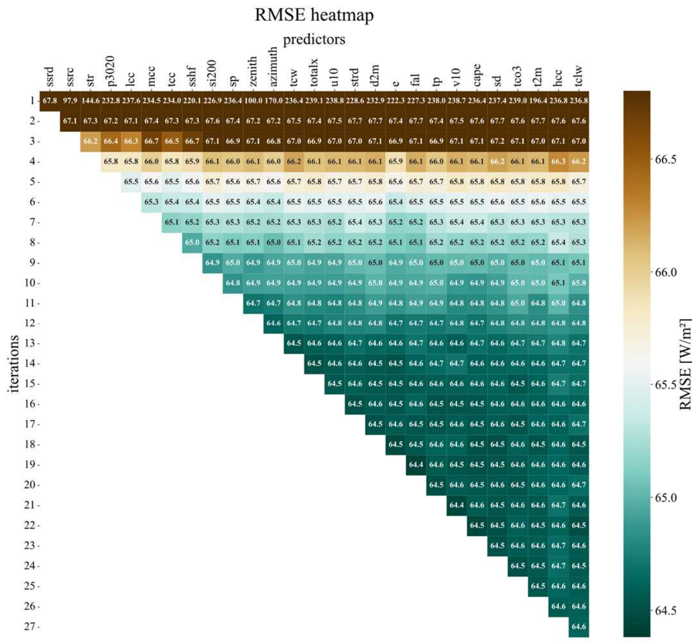

Fig. 3. The RMSE heatmap of the stepwise regression

First, the parameter *ssrd* is used as a predictor, being the most important weather variable. This is not surprising, because *ssrd* represents the ECMWF global horizontal irradiance forecast. The 19th iteration is the most accurate as it has the lowest RMSE of all the iterations. The predictors of iteration 19 are *ssrd*, *ssrc*, *str*, *p3020*, *lcc*, *mcc*, *tcc*, *sshf*, *si200*, *sp*, *zenith*, *azimuth*, *tcw*, *totalx*, *u10*, *strd*, *d2m*, *e* and *fal*, it is the best predictor combination. In addition to the solar radiation parameters (*ssrd*, *ssrc*), we find cloud cover (*lcc*, *tcc*, *mcc*), thermal radiation (*str*, *strd*), visibility (*p3020*), heat flux (*sshf*), surface pressure (*sp),* evaporation (*e*), solar position variables (*azimuth*, *zenith*), wind speeds (*si200*, *u10*), total column water (*tcw*), precipitation (*totalx*), and dewpoint temperature (*d2m*) and forecast albedo (*fal*). It can be observed that after the 13th iteration there is no significant difference between the minimum RMSE values, but until the 19th iteration the new predictors can improve the accuracy of the prediction slightly. However, the order of the predictors differs from location to location, it is not possible to generalize the most important predictors for a large outlier location, but they have to be examined on a location by location basis.

Compared to the reference RMSE value, an accuracy improvement of more than 5.43% is achieved thanks to post-processing. The minimum RMSE value is 64.43 W/m^2 in rounded form, while the reference RMSE value is 68.13 W/m^2. The GHI predictions are underdispersed, which is inherited from the NWP model itself, with the variance being only slightly affected by the scaling to 15-min resolution, as described in [11].

Table IV shows the error metrics calculated for the test period for PV power. The *best ghi* data the predictions from the most accurate 19th iteration. For *best gh*i the forecasts are underdispersed, while for the other cases they are overdispersed. There is a small difference between the RMSE values of the *best ghi* and *ssrd* power forecasts, which confirms the importance of post-processing. The lowest RMSE value is obtained for the measured GHI.

TABLE IV. THE VERIFICATION VALUES OF THE PV POWER FORECASTS

	MAE [kW]	nMAE [%]	RMSE [kW]	nRMSE [%]	MBE [kW]	nMBE [%]	*ρ* [%]	*F* [%]
Forecast *ssrd*	1576	26.0	2636	43.6	511	8.4	87.7	92.8
Measured *ghi*	961	15.9	1859	30.7	288	4.8	94.0	101.9
Forecast *best ghi*	1549	25.6	2516	41.6	208	3.4	88.2	86.8

The normalized errors of PV power forecasts are larger than those of GHI forecasts. This is due to the error amplification of the model chains [22]. More accurate radiation predictions lead to more accurate PV performance, so it is important to focus on the post-processing of radiation predictions. The bias in PV power forecasting is highly dependent on GHI input data [11]. The choice of NWP model affects the accuracy.

Fig. 4 shows time series PV power data. There is no significant difference between the post-processed and measured PV power data, the difference between them being mainly in the statistics. Predictions are mainly influenced by the underlying numerical weather prediction model and the physical model, so it is very important to develop and improve them.

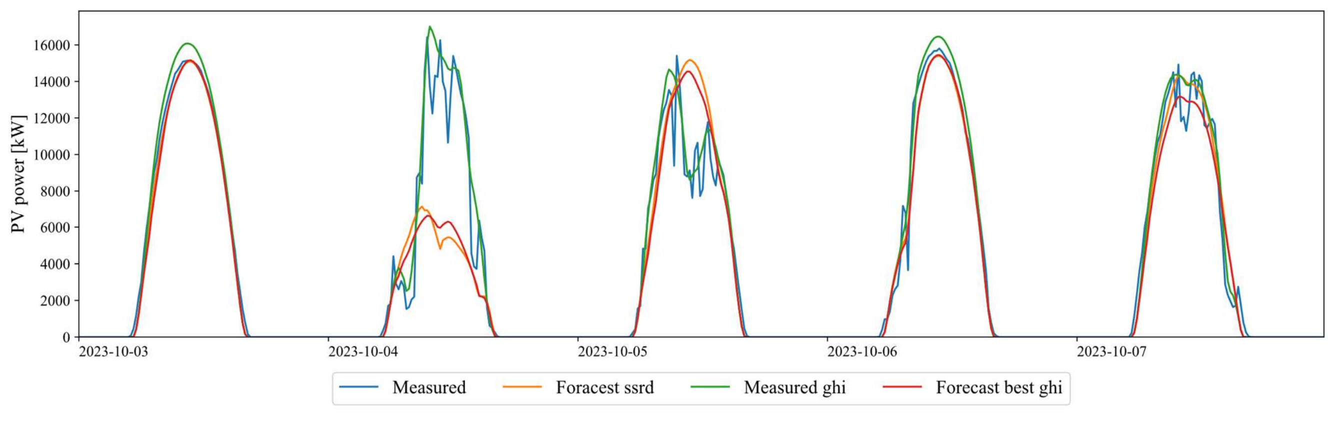

Fig. 4. PV power measurement and forecasts for 3-7 October 2023

V. CONCLUSION

The numerical weather prediction (NWP) is the most widely used method for estimating global horizontal irradiance (GHI). Forecasts are trendwise subject to error due to unique weather properties. The measured data allows for the identification of systematic error patterns that can be corrected by post-processing. In this study, 31 weather variables forecast by the NWP of the ECMWF are tested as potential predictors for the post-processing of GHI forecasts. The examination of correlation coefficients is a particularly important initial step, allowing for the filtering out of such data that can negatively affect the results and avoid multicollinearity. Stepwise regression is used to reveal the order of importance of the predictors. The initial root mean square error (RMSE) of 68.1 W/m^2 is reduced to 64.4 W/m^2 by a machine-learning-based post-processing using the optimal set of predictors. After interpolating the data for a quarter of an hour, the GHI is converted into power by a physical model chain. While the improvement in RMSE may appear moderate, it is important to emphasize the operational relevance of even small gains in forecasting accuracy because better predictability can lead to more efficient grid balancing and reduced reserve requirements. The RMSE of the raw PV power forecast is 2635.6 kW, which is reduced to 2516.4 kW due to post-processing. The methodology can support day-ahead or intraday planning, particularly for transmission system operators and energy traders managing large PV portfolios. This workflow can be used without actual PV production, which is relevant for proactive curtailment. The results of this research provide an excellent basis for development ideas, such as hyperparameter tuning, which can help to further avoid overfitting. The hyperparameters define the structure of the model, control the learning period, so they should be set before the model is run. The best hyperparameters depend largely on the application task, and the values should always be set for the problem. If the hyperparameters of the machine learning model are optimized during each iteration run, the prediction accuracy would probably increase, but the run time can also increase.

ACKNOWLEDGMENT

The authors would like to thank the Hungarian Meteorological Service for providing weather forecast data. This work was supported by the National Research, Development and Innovation Fund, project no. OTKA-FK 142702 and project no. KDP-IKT-2023-900-I1-00000957/0000003, and by the Hungarian Academy of Sciences through the Sustainable Development and Technologies National Programme (FFT NP FTA) and the János Bolyai Research Scholarship.

REFERENCES

[1] MAVIR Zrt.: "Evolution of installed capacity and other main characteristics of photovoltaic power plants", [online], available: https://mavir.hu/web/mavir/energia-mix-eromuvi-beepitett-teljesitokepesseg-adatok, 31.05.2025.

[2] Y. Chu, Y. Wang, D. Yang, S. Chen, M. Li: "A review of distributed solar forecasting with remote sensing and deep learning", Renewable and Sustainable Energy Reviews Volume 198, July 2024, 114391, DOI: 10.1016/j.rser.2024.114391

[3] W. Wang, D. Yang, T. Hong, J. Kleissl: "An archived dataset from the ECMWF Ensemble Prediction System for probabilistic solar power forecasting", Solar Energy Volume 248, December 2022, Pages 64-75, DOI: 10.1016/j.solener.2022.10.062

[4] M. Marrocu, L. Massidda: "Estimating the value of ECMWF EPS for photovoltaic power forecasting", Solar Energy Volume 279, 1 September 2024, 112801, DOI: 10.1016/j.solener.2024.112801

[5] S. Sperati, S. Alessandrini, L. D. Monache: "An application of the ECMWF Ensemble Prediction System for short-term solar power forecasting", Solar Energy Volume 133, August 2016, Pages 437-450, DOI: 10.1016/j.solener.2016.04.016

[6] H. Verbois, Y. Saint -Drenan, A. Thiery, P. Blanc: "Statistical learning for NWP post-processing: A benchmark for solar irradiance forecasting", Solar Energy Volume 238, 15 May 2022, Pages 132-149, DOI: 10.1016/j.solener.2022.03.017

[7] D. Yang, Y. Kong, B. Liu, J. Wang, D. Sun, G. Yang, W. Wang: "Comparing calibrated analog and dynamical ensemble solar forecasts", Solar Energy Advances Volume 4, 2024, 100048, DOI: 10.1016/j.seja.2023.100048

[8] MVM Paks II. Zrt.: "Climatic characterisation of a radius of Paks and 30 km", [online], available: https://paks2.hu/documents/20124/60046/10.%20fejezet%20-%20Paks%20%C3%A9s%2030%20km%20sugar%C3%BA%20k%C3%B6rnyezet%C3%A9nek%20%C3%A9ghajlati%20jellemz%C3%A9se.pdf/125f9e14-2270-dcab-5722-fe36bc16b094, 31.05.2025.

[9] ECMWF: "About us", [online], available: https://www.ecmwf.int/en/about, 31.05.2025.

[10] D. Yang, W. Wang, T. Hong: "A historical weather forecast dataset from the European Centre for Medium-Range Weather Forecasts (ECMWF) for energy forecasting", Solar Energy Volume 232, 15 January 2022, Pages 263-274, DOI: 10.1016/j.solener.2021.12.011

[11] M. J. Mayer, D. Yang, B. Szintai: "Comparing global and regional downscaled NWP models for irradiance and photovoltaic power forecasting: ECMWF versus AROME", Applied Energy Volume 352, 15 December 2023, 121958, DOI: 10.1016/j.apenergy.2023.121958

[12] P. Geurts, D. Ernst, L. Wehenkel: "Extremely randomized trees", Volume 63, pages 3–42, (2006), DOI: 10.1007/s10994-006-6226-1

[13] pvlib: "pvlib.location.Location.get_solarposition", [online], available: https://pvlib-python.readthedocs.io/en/stable/reference/generated/pvlib.location.Location.get_solarposition.html, 31.05.2025.

[14] pvlib: "pvlib.irradiance.erbs", [online], available: https://pvlib-python.readthedocs.io/en/stable/reference/generated/pvlib.irradiance.erbs.html, 31.05.2025.

[15] pvlib: „pvlib.irradiance.get_extra_radiation", [online], available: https://pvlib-python.readthedocs.io/en/stable/reference/generated/pvlib.irradiance.get_extra_radiation.html, 31.05.2025.

[16] pvlib: „pvlib.irradiance.get_total_irradiance", [online], available: https://pvlib-python.readthedocs.io/en/stable/reference/generated/pvlib.irradiance.get_total_irradiance.html, 31.05.2025.

[17] pvlib: „pvlib.temperature.pvsyst_cell", [online], available: https://pvlib-python.readthedocs.io/en/stable/reference/generated/pvlib.temperature.pvsyst_cell.html, 31.05.2025.

[18] pvlib: „pvlib.pvsystem.pvwatts_dc", [online], available: https://pvlib-python.readthedocs.io/en/stable/reference/generated/pvlib.pvsystem.pvwatts_dc.html, 31.05.2025.

[19] pvlib: „pvlib.inverter.pvwatts", [online], available: https://pvlib-python.readthedocs.io/en/stable/reference/generated/pvlib.inverter.pvwatts.html, 31.05.2025.

[20] M. J. Mayer, Gy. Gróf: "Techno-economic optimization of grid-connected, ground-mounted photovoltaic power plants by genetic algorithm based on a comprehensive mathematical model", Solar Energy Volume 202, 15 May 2020, Pages 210-226, DOI: 10.1016/j.solener.2020.03.109

[21] D. Markovics, M. J. Mayer: "Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction", Renewable and Sustainable Energy Reviews Volume 161, 2022/112364, DOI: 10.1016/j.rser.2022.112364

[22] M. J. Mayer: "Impact of the tilt angle, inverter sizing factor and row spacing on the photovoltaic power forecast accuracy", Applied Energy Volume 323, 1 October 2022, 119598, DOI: 10.1016/j.apenergy.2022.119598