

Cross-Regional Solar Irradiance Forecasting in Iraq Using Explainable Machine Learning: Evidence from Najaf, Basra, and Erbil

Hiba Qasim Mohammed

Kufa Institute, Al-Furat Al-Awsat Technical University, Al-Najaf, Iraq

ABSTRACT: Day-ahead forecasting of global horizontal irradiance (GHI) is required for photovoltaic planning, operational scheduling, and grid integration, especially in countries where strong solar resources coexist with marked climatic heterogeneity. In Iraq, the southern, central, and northern regions differ substantially in temperature, humidity, rainfall, wind, and seasonality, and these contrasts are expected to affect how forecasting models behave and transfer between regions. In this study, an explainable forecasting framework was developed for Basra, Najaf, and Erbil using daily NASA POWER solar and meteorological data for 2020–2025. Next-day GHI was defined as the target variable, and a leakage-controlled predictor set was constructed from solar, meteorological, temporal, lagged, and rolling features. Six machine-learning models were evaluated against a persistence benchmark: linear regression, random forest, support vector regression, XG Boost, Light GBM, and a feed-forward artificial neural network. The lowest test error in 2025 was obtained by linear regression ($MAE = 0.486 \text{ kWh/m}^2/\text{day}$, $RMSE = 0.723 \text{ kWh/m}^2/\text{day}$, $MAPE = 14.974\%$, $R^2 = 0.860$), although its advantage over XG Boost was not statistically significant. Similar behavior was confirmed by walk-forward testing, which indicated that forecast quality was governed more by feature design than by model complexity. The lowest errors were obtained in Najaf and Basra, whereas the highest errors were observed in Erbil, where stronger seasonality and higher precipitation were present. The explainability analysis further indicated that same-day GHI, lagged and rolling irradiance terms, seasonal indicators, surface pressure, temperature, and humidity contributed most strongly to the forecasts. It was therefore demonstrated that, once temporally consistent features were constructed without leakage, a simple transparent model could perform competitively with more complex ensemble approaches across the main climatic regions of Iraq.

KEYWORDS: day-ahead forecasting; explainable machine learning; global horizontal irradiance; Iraq; NASA POWER; photovoltaic planning

1. INTRODUCTION

Solar energy has become central to low-carbon power planning, energy-security strategies, and large-scale photovoltaic deployment. Among solar-resource variables, global horizontal irradiance (GHI) is routinely used in resource assessment, system sizing, yield estimation, operational scheduling, and grid-integration studies. Reliable short-term GHI forecasts are therefore required not only for site evaluation but also for day-to-day operational decisions in photovoltaic systems [1–6].

Iraq is characterized by high solar potential, yet its climatic conditions are not spatially uniform. Central and southern Iraq are predominantly hot and arid, whereas the north is cooler, wetter, and more seasonally variable. Differences in temperature, humidity, rainfall, wind regime, pressure, cloud cover, and dust loading are expected to influence both GHI variability and forecast behavior across regions [7–11].

Short-term solar forecasting is affected by forecast horizon, temporal resolution, predictor selection, data quality, validation design, and local climate. Daily irradiance is governed by both seasonal solar geometry and rapidly changing atmospheric processes, including clouds, humidity, precipitation, aerosols, and transient weather systems. For these reasons, empirical and statistical methods have increasingly been complemented by data-driven and machine-learning approaches [12–21].

Solar-irradiance and photovoltaic (PV) forecasts have been used in a variety of machine-learning models, such as linear regression, support vector regression, random forests, artificial neural networks, XGBoost, and LightGBM. There is no single model that has yet been proven to be better than the others in all circumstances. For many time-series applications, the gains achieved have been due to feature engineering and leakage-free validation, rather than to the complexity of the model [22–38].

Satellite and reanalysis products are extensively exploiting for solar-resource assessment and preliminary forecasting as they have widely been used in many areas where continuous ground-based irradiance measurements are still limited. NASA POWER offers solar radiation and meteorological data globally and is widely utilized in various renewable-energy projects. It has also been tested in the case of Iraq and it is a realistic alternative source of data for regional forecasting in situations where dense networks of pyranometers are not available [7–8,39–44].

Cross-Regional Solar Irradiance Forecasting in Iraq Using Explainable Machine Learning: Evidence from Najaf, Basra, and Erbil

Besides prediction accuracy, interpretability is needed if the behavior of forecasts is to be related to physical conditions and operational decisions. Using model explanations, the relative importance of irradiance persistence, seasonal structure and meteorological co-variables (humidity, precipitation, pressure, wind and temperature) can be determined for the association between GHI and irradiance on the next day. This is especially important if transparent models are considered against more complex ensemble models [34–38,45–48].

Solar-resource assessment and photovoltaic (PV) potential have been studied in Iraq and neighboring countries, but limited research has focused on day-ahead global horizontal irradiances (GHI) forecasting in climatically diverse regions in Iraq, with the use of multi-year daily data provided by the National Aeronautics and Space Administration (NASA) POWER program. There is very little evidence so far which can be used to make direct comparisons between central, southern and northern Iraq [7-10,22-33,44]. Thus, a leakage controlled and explanatory day-ahead forecasting system has been created for the locations of Najaf, Basra, and Erbil based on the daily NASA POWER data for 2020–2025. Solar, meteorological, temporal, lagged and rolling predictors were created, multiple machine-learning models were compared against a persistence model, and the behavior of the models across city-level was analyzed. The study aimed to give an idea about forecasting short term GHI for future use in Iraq and not to claim the methodological novelty.

2. MATERIALS AND METHODS

2.1 Study area

The cities studied were Najaf, Basra and Erbil. The cities were chosen to represent the Middle, South, and North of Iraq, respectively, and to give a comparative picture of the behavior of solar radiation under varying climatic conditions in Iraq. Najaf was an example of the hot and dry weather of central Iraq, Basra of the warm and humid weather of southern Iraq, and Erbil of the cooler and wetter weather of northern Iraq. The aforementioned sites were therefore chosen for testing the reliability of the use of a single climate-forecasting methodology in the major climatic regions of Iraq.

Table 1. Cities selected for study

City	Region represented	Latitude	Longitude
Basra	Southern Iraq	30.5016° N	47.6607° E
Najaf	Central Iraq	31.9992° N	44.2959° E
Erbil	Northern Iraq	36.1075° N	44.0063° E

2.2 Dataset description

Single-point extraction for each city was used to retrieve the daily solar-radiation and meteorological data from the NASA POWER program for the period 01 January 2020 to 31 December 2025. A daily resolution was chosen due to the balance of forecast relevance and data stability, while also being easily interpretable to assess resources at the regional scale.

There were 6,576 records from the three locations, with each city providing 2,192 daily records. The variables that were retained were solar-radiation components and meteorological indices that are typically used in the calculation of irradiance variability. The variables used to predict the next day global horizontal irradiance (GHI) were created.

The target parameter for forecast was the global horizontal solar radiation for the coming day $GHI(t+1)$. This was considered to be the GHI reported by NASA POWER the next day. It is a measurement that has been derived from satellite data and reanalysis, not from a ground-based measurement made by a radiometer. This definition guaranteed that the modeling task was still a short-term forecasting problem and not a same day reverse calculation problem.

Table 2. NASA Power program variables used in the study.

Variable	Description	Unit
ALLSKY_SFC_SW_DWN	Global horizontal irradiance under all-sky conditions	kWh/m ² /day
ALLSKY_SFC_SW_DNI	Direct normal irradiance under all-sky conditions	kWh/m ² /day
ALLSKY_SFC_SW_DIFF	Diffuse shortwave irradiance under all-sky conditions	kWh/m ² /day
T2M	Air temperature at 2 m	°C
T2M_MAX	Maximum air temperature at 2 m	°C
T2M_MIN	Minimum air temperature at 2 m	°C
RH2M	Relative humidity at 2 m	%
IMERG_PRECTOT	Daily precipitation	mm/day
PS	Surface pressure	kPa
WS10M	Wind speeds at 10 m	m/s

Cross-Regional Solar Irradiance Forecasting in Iraq Using Explainable Machine Learning: Evidence from Najaf, Basra, and Erbil

2.3 Data preprocessing

Each downloaded data series was checked for completeness, consistency, and chronological order. Dates were converted to standard date and time format, and city and region identifiers were assigned to each record before the three city-level datasets were merged for comparative modeling.

There were no missing values in the downloaded files; therefore, missing values were not imputed. The data set was additionally inspected for the irrational values and duplicate records. The order was maintained in the data throughout the preprocessing stage, so that no new information was added to the training set after the stage (45-49).

The target was chosen as to move the time series one day forward within each city when creating the time series. The data obtained from the previous day (t) was then fed into a model to predict the global horizontal irradiance for the next day ($t+1$). The final data point in each city series were deleted due to the shift since no target was available for the next day.

2.4 Feature engineering

The temporal and climatic structure of irradiance were represented using feature engineering. To augment the original solar and meteorological variables, so-called temporal indicators, lagged terms and rolling statistics were created.

The date was used to create a series of calendar variables such as month, day of the year, and season. Irradiance is also dependent on the seasonal solar geometry as well as the atmospheric conditions, so the annual cycle was also coded using sine and cosine transformations of the day of the year.

The continuity of solar irradiance and weather over short time span were captured using lagged features. The main delays that were taken into consideration for irradiance were $GHI(t-1)$, $GHI(t-2)$, $GHI(t-3)$, and $GHI(t-7)$, along with the relevant lagged terms for the meteorological variables: temperature, relative humidity, precipitation, wind speed, and surface pressure. Three-day and seven-day moving averages were also presented, based on only previous data to avoid unintended "leakage."

The final modeling table was made with the original solar and meteorological variables, calendar indicators, city and region identifiers, lagged terms and rolling statistics. Thus, the local data as well as a pattern disseminated to the three cities were preserved within the model structure.

2.5 Machine Learning Models and Their Application

Model selection was initiated with a simple linear baseline and then extended to more flexible nonlinear models. The purpose was to test whether increased model complexity produced a meaningful gain in prediction accuracy. All models were trained with the same predictor set to ensure a consistent comparison framework.

Multiple linear regression was adopted as a transparent baseline. Random Forest was included to represent nonlinear tree aggregation, SVR was implemented with a radial basis function kernel, and XG Boost and Light GBM were selected because of their strong performance on tabular datasets. A feed-forward artificial neural network was also included to assess whether additional flexibility improved next-day GHI prediction(50-52).

Table 3. Machine learning models and key configuration settings.

Model	Main configuration used
Linear Regression	Ordinary least squares baseline with standardized numerical predictors
Random Forest	120 trees, maximum depth = 12, minimum samples per leaf = 3
SVR	Radial basis function kernel, $C = 10$, $\epsilon = 0.05$, $\gamma = \text{scale}$
XG Boost	250 estimators, maximum depth = 3, learning rate = 0.04, subsample = 0.85, col sample _by tree = 0.85
Light GBM	250 estimators, learning rate = 0.04, num_ leaves = 31, subsample = 0.85, col sample _by tree = 0.85
ANN	Two hidden layers (48 and 24 neurons), Re LU activation, Adam optimizer, early stopping enabled

2.6 Training, Validation, and Testing Strategy

The data was a time series and so were divided chronologically, not randomly. Data from 2020-23 were used for initial training, data from 2024 for validation/adjustment, and unseen data from 2025 for testing. Once the model settings had been determined, the final models were then refitted to all the data until 2024 and evaluated on the data from 2025.

This design is representative of the reality of forecasting: models learn from previous observations and are tested using future observations. Random segmentation was not used as this might have confounded the future and past records, and led to excessive optimism.

One model was trained for all three cities; city and region identifiers were added to the list of predictors. The performance in tests was next reported by city, to investigate regional behavior.

Cross-Regional Solar Irradiance Forecasting in Iraq Using Explainable Machine Learning: Evidence from Najaf, Basra, and Erbil

Table 4. Data segmentation strategy based on chronological order.

Subset	Period	Purpose
Training	2020-2023	Initial model fitting
Validation	2024	Model tuning and selection
Testing	2025	Final unseen performance evaluation

2.7 Model Evaluation Measures

Prediction performance was estimated using mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), and coefficient of determination (R^2) [49]. These measures collectively describe the mean error, sensitivity to large errors, percentage relative error, and explained variance.

The metrics were calculated as follows for n observations:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2)$$

$$MAPE = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4)$$

Where y_i represents the observed value, $\hat{y}_i$ the predicted value, $\bar{y}$ the average observed value, and n the number of observations.

Lower MAE, RMSE, and MAPE values indicate better performance, while a higher R^2 value indicates stronger agreement between observed and predicted GHI values. MAPE has been interpreted with caution because the error ratio becomes exaggerated when irradiance is low, particularly in winter or under cloudy conditions. This is especially important in Erbil, where seasonal variability is greater.

2.8 Explainable Machine Learning Analysis

Explanatory capacity was analyzed post training of models. The model that resulted in the lowest overall error was linear regression, so standardized coefficients were used from this model for final interpretation. The importance scores from the tree model were also examined to gain some insight into the relative impact of the solar, meteorological, temporal and location-based predictors. The results were interpreted as model-level associations, since the irradiance, the temperature, the seasonal variables and the lagged terms were correlated.

It was added to prevent the work from being seen as a "predictive exercise. Therefore, the feature importance was also analyzed in conjunction with descriptive, seasonal and correlation analyses, since a significance score does not imply causation.

2.9 General Methodological Workflow

The methodological workflow comprised six main steps: extraction of daily solar and meteorological data from NASA POWER, verification and merging of the city-level datasets, feature generation, chronological model training, testing and model evaluation, and feature-importance analysis.

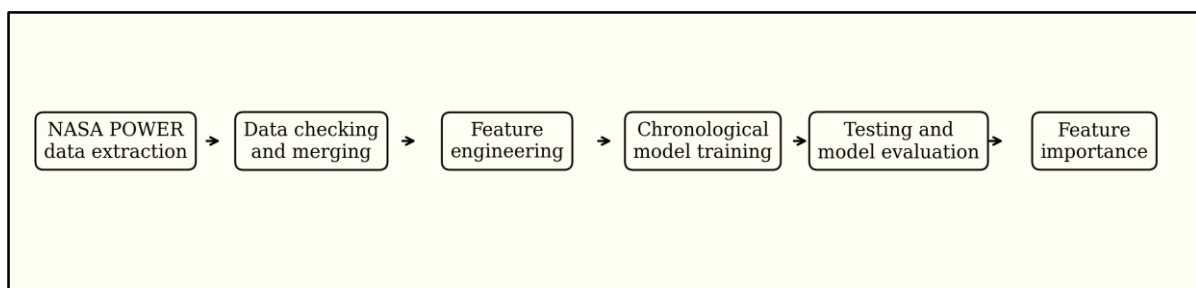


Figure 1. Overall methodological workflow used in the study.

3. RESULTS AND DISCUSSION

3.1 Descriptive Analysis of Solar and Climatic Conditions

Table 5 presents the descriptive statistics for the period 2020–2025. The spatial differences are clearly evident: Basra has the highest average daily GHI, followed closely by Najaf, while Erbil has the lowest annual average.

Cross-Regional Solar Irradiance Forecasting in Iraq Using Explainable Machine Learning: Evidence from Najaf, Basra, and Erbil

Table 5. Descriptive statistics for selected cities during the period 2020-2025.

City	Region	Avg GHI	SD GHI	Min GHI	Max GHI	Avg Temp	Max Temp	Avg RH	Avg Wind	Total Rain
Basra	Southern Iraq	5.598	1.824	0.733	8.797	27.29	52.73	31.46	4.66	1054.25
Najaf	Central Iraq	5.533	1.815	0.576	8.575	25.43	52.36	32.81	3.82	821.13
Erbil	Northern Iraq	5.052	2.307	0.323	9.235	22.02	49.16	39.77	3.03	2388.08

The average temperature of the warmest city was Basra and that of the coldest was Erbil, thus showing greater convective level in southern Iraq. Even though it had the lowest average irradiance, Erbil showed the maximum solar radiation standard deviation for the three cities.

The differences in precipitation were also significant. Erbil experienced the highest amount of rainfall, while Basra experienced a medium amount and Najaf experienced the least amount of rainfall during the study period. This variability allows data from the three cities to be used for the evaluation of the forecast performance in truly different climate conditions.

3.2 Seasonal Variations in Solar Radiation

Table 6 summarizes the seasonal averages of GHI and the major climatic variables. In all cities, radiation reached its maximum value in summer and its minimum value in winter, reflecting the combined effects of solar geometry, day length, atmospheric humidity, and cloud cover.

Table 6. Seasonal averages of GHI and climate variables.

City	Season	Avg GHI	Avg Temp	Avg RH	Total Rain
Basra	Winter	3.655	14.10	52.82	465.65
Basra	Spring	6.304	26.69	30.62	313.67
Basra	Summer	7.460	39.38	14.74	13.02
Basra	Autumn	4.931	28.75	28.01	261.91
Najaf	Winter	3.514	12.00	54.26	366.68
Najaf	Spring	6.329	24.46	33.13	257.21
Najaf	Summer	7.445	37.91	15.67	14.33
Najaf	Autumn	4.801	27.13	28.54	182.91
Erbil	Winter	2.592	8.86	61.04	953.12
Erbil	Spring	5.532	20.13	46.41	1024.00
Erbil	Summer	7.640	35.13	18.49	14.48
Erbil	Autumn	4.391	23.73	33.44	396.48

Solar-radiation intensity in summer was similar in Basra and Najaf, whereas Erbil recorded the highest summer average but a much sharper winter decline than the other two cities, indicating a stronger seasonal cycle in the north. Seasonal rainfall helps explain this pattern. Erbil is wetter in winter and spring than Najaf or Basra, and the resulting cloud cover and scattering reduce solar-radiation intensity. These seasonal findings therefore show why Iraqi forecasting models must account for both regional and temporal variation rather than relying solely on annual averages.

Cross-Regional Solar Irradiance Forecasting in Iraq Using Explainable Machine Learning: Evidence from Najaf, Basra, and Erbil

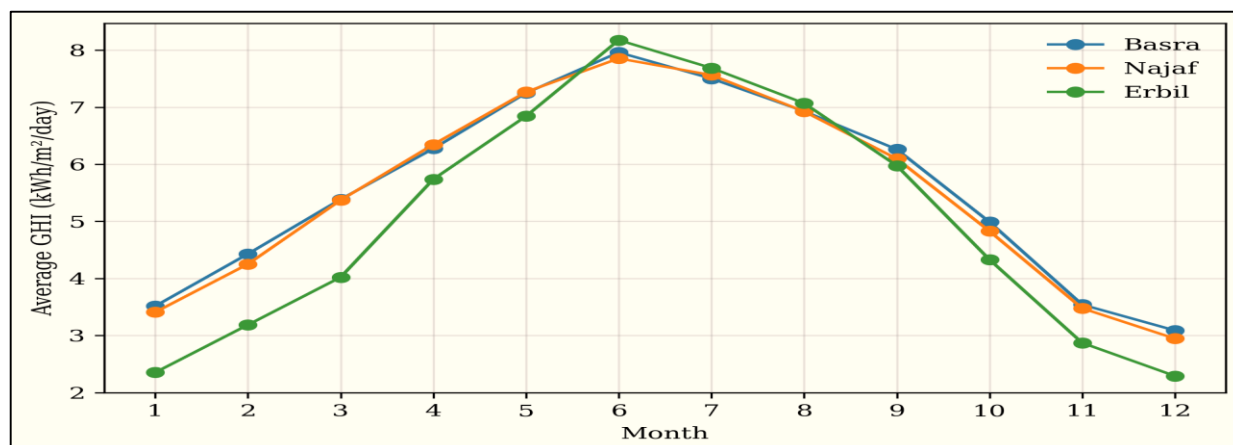


Figure 2. Average global monthly solar radiation in Najaf, Basra and Erbil during the period 2020-2025.

3.3 Correlation Analysis and the Importance of Predictive Variables

The linear relationship of the selected predictors with next-day GHI was assessed with Pearson correlation analysis. Model development was preceded by the preliminary screening of solar, climatic and temporal variables in the analysis. Table 7 shows the strength of the correlation with the target and ranks the predictors.

Table 7. Pearson correlation coefficients between the selected predictive variables and the GHI the following day.

Predictor	Correlation with next-day GHI
Same-day GHI	0.893
Day-of-year seasonal component	-0.881
GHI lag 1 day	0.847
GHI lag 2 days	0.831
GHI lag 3 days	0.827
GHI lag 7 days	0.807
Maximum temperature	0.770
Mean temperature	0.762
Previous-day temperature	0.761
Relative humidity	-0.731
Minimum temperature	0.724
Previous-day relative humidity	-0.697
Diffuse irradiance	0.579
Direct normal irradiance	0.525
Wind speed lag 1 day	0.361
Wind speed	0.355
Precipitation	-0.233
Surface pressure	-0.203
Precipitation lag 1 day	-0.174

Same-day GHI is closely related to next-day GHI, which is a short time continuity and justifies the use of radiation forecast for tomorrow's forecast. The lagged GHI variables also did very well, and particularly the one-day lag was significant, indicating that there is significant information in the history of irradiance.

The seasonal term also had good correlation in size, which further substantiates the seasonal effect on daily irradiance. The negative sign is there because the seasonal variable was encoded in a specific way, and it's not necessarily a physical effect.

The positive correlation between the temperature variables and GHI on the following day was observed, as would be expected, warmer and clearer weather is correlated with greater irradiance. Relative humidity was found to be negatively correlated with precipitation, which is expected as more rain reduces the radiance.

While correlation is helpful as a starting point, it fails to capture nonlinear interactions between predictors. To answer this question, more complicated relations were applied during the modeling stage and tested to find out whether the results of the forecast of GHI the next day are better than the simpler relations.

Cross-Regional Solar Irradiance Forecasting in Iraq Using Explainable Machine Learning: Evidence from Najaf, Basra, and Erbil

Table 8. Main correlated predictors by city.

City	Strongest predictors of next-day GHI
Basra	Seasonal component, same-day GHI, GHI lag 1, surface pressure, maximum temperature, humidity
Najaf	Seasonal component, same-day GHI, GHI lag 1, surface pressure, temperature, humidity
Erbil	Seasonal component, same-day GHI, GHI lag 1, maximum temperature, humidity, pressure

At city level, the main predictor groups are broadly similar. Same-day GHI, lagged GHI, seasonal indicators, temperature, humidity, and pressure appear among the leading variables in all three locations. The difference lies mainly in the strength of seasonality and irradiance continuity: Erbil is more strongly influenced by seasonal variation, whereas Najaf and Basra exhibit more stable short-term continuity.

3.4 Forecasting Performance Using Machine Learning

The models were then used to predict the GHI for the following day in the three cities. Each model used the same predictor set: solar, meteorological, temporal, lagged, and rolling variables. The year 2025 was treated as the unseen test year. For comparison, a persistence benchmark was also included, in which next-day GHI was assumed to be equal to same-day GHI.

Table 9. Overall performance of GHI forecasts for the next day during the 2025 testing period.

Model	MAE	RMSE	MAPE (%)	R ²
Persistence benchmark	0.538	0.889	15.416	0.789
Linear Regression	0.486	0.723	14.974	0.860
XG Boost	0.492	0.743	15.497	0.852
Light GBM	0.491	0.754	15.259	0.848
Random Forest	0.506	0.754	15.876	0.848
ANN	0.536	0.781	16.648	0.837
SVR	0.496	0.802	16.679	0.828

All the models performed better than persistence for both the root mean square error (RMSE) and coefficient of determination (R²); thereby adding meteorology, time, and history of irradiance parameters to the same day assumption was useful. Linear regression produced the lowest overall error in the 2025 test year (MAE = 0.486 kWh/m²/day, RMSE = 0.723 kWh/m²/day, MAPE = 14.974%, R² = 0.860).

There were minor differences among the best models. XG Boost reached RMSE = 0.743 kWh/m²/day and R² = 0.852, while Light GBM reached RMSE = 0.754 kWh/m²/day and R² = 0.848. These are very similar to the linear regression values which indicates that a high proportion of the predictable "structure" from the next day's irradiance had been obtained from the seasonal and lagged features. More complex nonlinear structures were not able to provide a clear advantage on this daily data set. The current effect is that for regional studies, accuracy, interpretability and simplicity are all important factors to consider when selecting a prediction method.

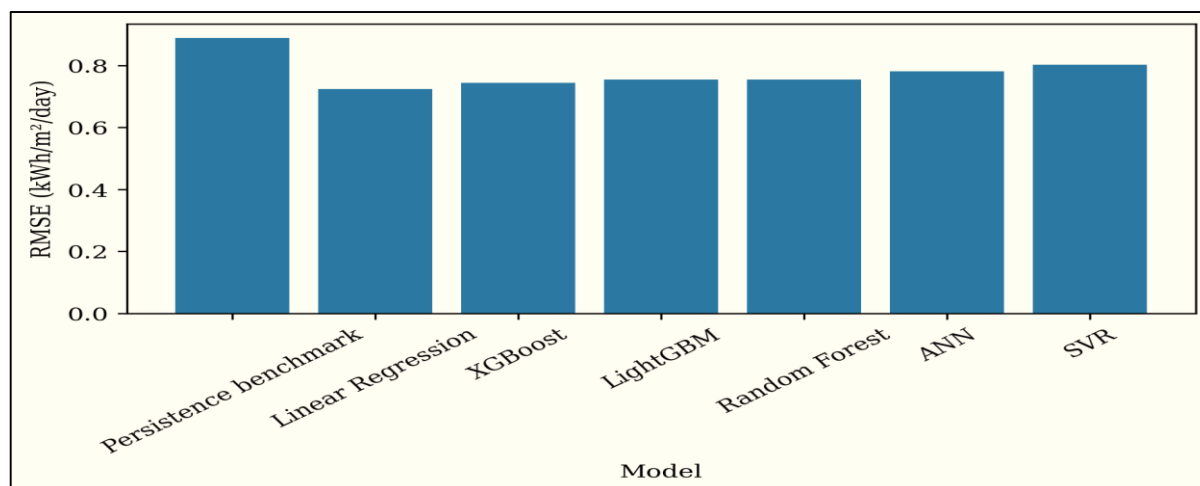


Figure 3. Overall RMSE comparison across the evaluated forecasting models on the unseen 2025 test year.

Cross-Regional Solar Irradiance Forecasting in Iraq Using Explainable Machine Learning: Evidence from Najaf, Basra, and Erbil

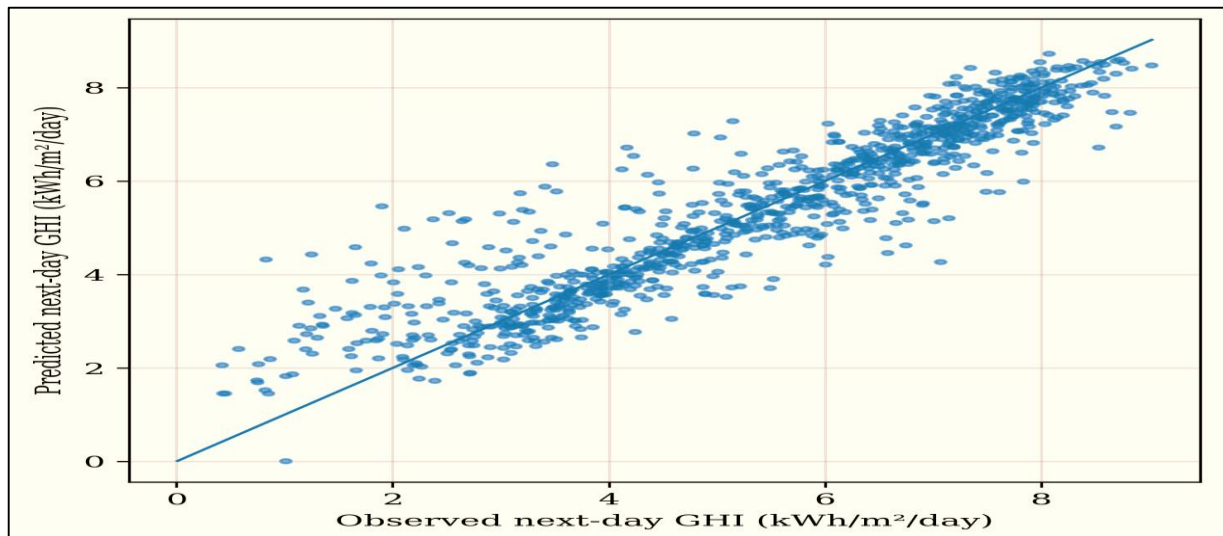


Figure 4. Predicted versus observed next-day GHI for the lowest-error model over the 2025 test year.

The statistical significance of these differences was examined using a Diebold–Mariano test with squared-error loss for the 2025 test year (Table 10). Linear regression significantly outperformed persistence (DM = -4.261, $p < 0.001$). Its difference from XGBoost, however, was marginal and did not reach the 0.05 significance level (DM = -1.920, $p = 0.055$). Differences from LightGBM, Random Forest, ANN, and SVR were statistically significant but modest. The most cautious conclusion is therefore that linear regression achieved the lowest error, but not decisively relative to XGBoost; rather, a simple transparent model remained competitive when leakage-controlled temporal and lagged features were available.

Table 10. Diebold–Mariano test comparing Linear Regression with the other models (squared-error loss, 2025 testing period).

Comparison	Mean loss differential (LR – comparator)	DM statistic	p-value	Interpretation
Linear Regression vs Persistence	-0.2669	-4.261	0.000022	LR significantly outperformed persistence
Linear Regression vs XG Boost	-0.0296	-1.920	0.055	Not significant at 0.05; marginal only
Linear Regression vs Light GBM	-0.0456	-2.097	0.036	Significant but modest
Linear Regression vs Random Forest	-0.0460	-2.525	0.0117	Significant
Linear Regression vs ANN	-0.0864	-3.798	0.000154	Significant
Linear Regression vs SVR	-0.1206	-2.441	0.0148	Significant

A walk-forward check, using 2023, 2024, and 2025 as successive test years, was carried out to assess robustness (Table 11). Both linear regression and XGBoost showed very similar root mean square error (RMSE) values (0.787 and 0.792 kWh/m²/day), and all models outperformed the continuity model (RMSE = 0.958 kWh/m²/day). This confirms that the similarity between linear and tree-based models was not limited to a single year.

Table 11. Root mean squared error (RMSE, kWh/m²/day) of the base models from walk-forward evaluation using 2023, 2024, and 2025 as successive test years.

Model	RMSE 2023	RMSE 2024	RMSE 2025	Mean RMSE
Linear Regression	0.846	0.792	0.723	0.787
XGBoost	0.851	0.783	0.742	0.792
LightGBM	0.865	0.785	0.749	0.800
Random Forest	0.863	0.801	0.756	0.807
Persistence	1.014	0.972	0.889	0.958

Cross-Regional Solar Irradiance Forecasting in Iraq Using Explainable Machine Learning: Evidence from Najaf, Basra, and Erbil

3.5 City-Level Forecasting Performance

To assess regional consistency, the 2025 test results were separated by city. Table 12 shows the least-error model for the cities of Basra, Najaf, and Erbil.

Table 12. Best performance for city-level forecasts during the 2025 testing period.

City	Best model	MAE	RMSE	MAPE (%)	R ²
Basra	Linear Regression	0.447	0.684	12.559	0.855
Najaf	Linear Regression	0.429	0.643	12.736	0.873
Erbil	Linear Regression	0.583	0.829	19.628	0.852

The accuracy of the forecasts varied across regions. The lowest root mean square error (RMSE) was obtained in Najaf, followed by Basra, whereas the highest error was recorded in Erbil. This ranking is consistent with the descriptive and seasonal analyses, as Erbil exhibited stronger seasonal variability and heavier rainfall.

The small error in the case of Najaf and Basra regions, can be attributed to the stability of the pattern of solar radiation in central and southern Iraq in the test period. Erbil has a high relative meaning absolute error (MAPE) suggesting that forecasting in this area is difficult due to variable, humid and seasonal conditions.

Despite these regional differences, the coefficient of determination (R²) remained above 0.85 in all cities. This indicates that the model explained most of the daily variation in GHI across Iraq's climatic zones.

3.6 Feature Significance and Interpretable Machine Learning Analysis

After model training, explanatory analysis was carried out using tree-based importance scores. Random Forest, XGBoost, and LightGBM were examined to identify the variables most strongly associated with next-day GHI. Their importance values were interpreted together with the descriptive, seasonal, and correlation analyses.

Table 13. The best predictive indicators were identified by tree-based machine learning models.

Rank	LightGBM	XGBoost	Random Forest
1	Same-day GHI	7-day rolling GHI	Same-day GHI
2	Wind speeds at 10 m	Same-day GHI	Seasonal component
3	Surface pressure	Seasonal component	7-day rolling GHI
4	Precipitation	3-day rolling GHI	Precipitation
5	Seasonal component	GHI lag 2 days	GHI lag 2 days
6	Direct normal irradiance	7-day rolling diffuse irradiance	7-day rolling diffuse irradiance
7	Seasonal sine component	GHI lag 1 day	3-day rolling GHI
8	Relative humidity	Surface pressure	Surface pressure
9	Diffuse irradiance	Precipitation	Wind speeds at 10 m
10	Diffuse irradiance lag 7 days	3-day rolling diffuse irradiance	GHI lag 3 days

In all three of the tree-based models, solar-radiation variables ranked high. The same day GHI and rolling irradiance terms were consistently recorded as the strongest predictors, highlighting the importance of irradiance persistence in short-term forecasting. Seasonal variables were also significant, particularly in XGBoost and Random Forest, reflecting the high difference between summer and winter across the three cities.

Other meteorological variables such as wind speed, surface pressure, precipitation, relative humidity, and diffuse irradiance also played a significant role, suggesting that irradiance history alone was not the only factor influencing the next day's GHI, but that atmospheric conditions that impact transmission and cloud formation are also important.

Care must be taken with the results, as the importance of features does not imply that they cause the model's response. Therefore, the importance scores were covered along with the descriptive, seasonal and correlation analysis.

Cross-Regional Solar Irradiance Forecasting in Iraq Using Explainable Machine Learning: Evidence from Najaf, Basra, and Erbil

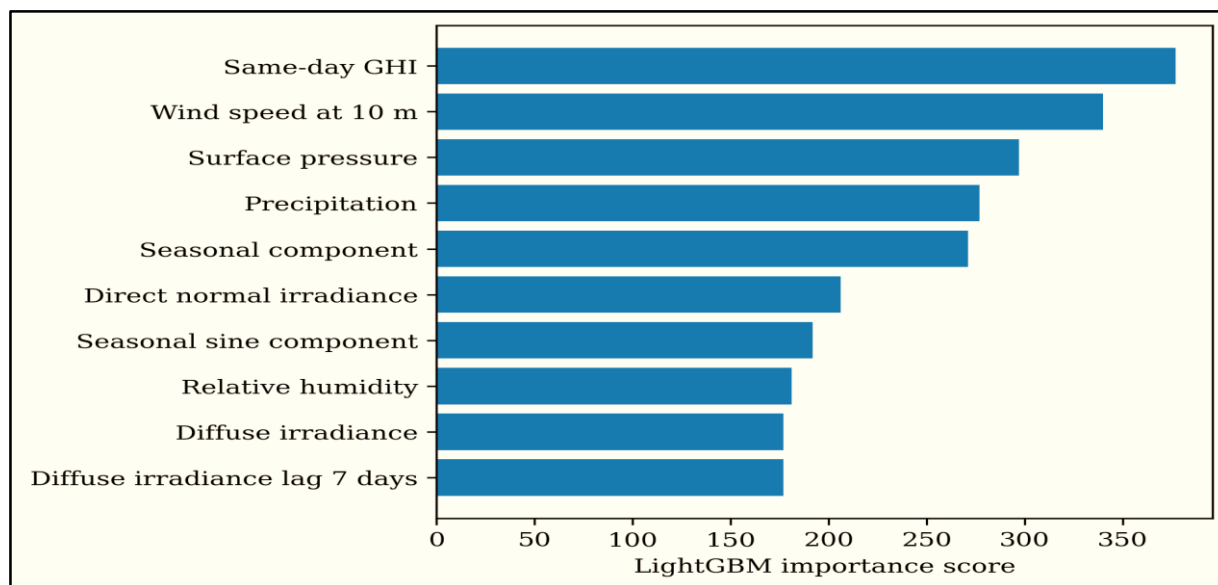


Figure 5. The most important findings regarding the significance of Light GBM features for the next day GHI are presented as a supplementary explanation.

Because linear regression yielded the lowest overall error, standardized coefficients from that model were examined for final interpretation (Table 14). Air temperature, the seasonal cosine term, surface pressure, lagged temperatures, same-day irradiance, rolling irradiance, and rolling diffuse irradiance were among the dominant coefficients. Owing to cross-correlations among these predictors, these coefficients were interpreted as model-level associations rather than direct causal effects.

Table 14. Standard linear regression coefficients for the most influential factors in predicting the GHI for the following day

Predictor	Standardized coefficient
Air temperature (T2M)	-1.857
Seasonal cosine term (cos_doy)	-1.157
Surface pressure (PS)	0.998
Temperature, lag 1 day (T2M_lag1)	0.919
Maximum temperature (T2M_MAX)	0.914
7-day rolling GHI mean (GHI_roll7)	0.904
Same-day GHI (GHI_t)	0.683
Surface pressure, lag 1 day (PS_lag1)	-0.640
Temperature, lag 2 days (T2M_lag2)	-0.415
7-day rolling pressure (PS_roll7)	-0.400
7-day rolling diffuse irradiance (DIFF_roll7)	-0.394

3.7 Discussion of Prediction Behavior Across Regions

Good future forecasts were achieved for all three regions with the best model giving $R^2 = 0.860$ in the unseen test year of 2025. The selected predictors thus accounted for most of the variation seen on a daily basis.

The linear regression method's performance was seen as being dependent on the data structure and not as being superior in some way. More complex nonlinear structures did not lead to significantly improved accuracy on the predictable signal, as this was already achieved by the three variables of seasonal regularity, irradiance continuity and engineered temporal features described above. A simple transparent model was able to stay competitive with the best-performing ensemble models for this reason.

The main regional differences were climatic. Najaf with the minimum errors followed by Basra while Erbil had maximum errors due to greater seasonality, rainfall and less stable conditions of irradiance.

These results show that the one approach wouldn't work well for all regions of Iraq without taking into account regional climatic data. Regional assessment is consequently very crucial for accurate solar-forecasting tools that can be used in PV planning, grid operation and in renewable-energy management in Iraq.

3.8 Practical Implications for Solar PV Planning

The findings can have immediate applications in photovoltaic planning in the Iraqi context. For the city of Basra and Najaf, the daily weather forecast and the irradiance history were enough, while the forecasts for Erbil would be better off with a city specific

Cross-Regional Solar Irradiance Forecasting in Iraq Using Explainable Machine Learning: Evidence from Najaf, Basra, and Erbil

model or more atmospheric data like cloud indices, aerosol optical depth or the satellite derived cloud cover. The proposed framework can then be used for day-ahead PV forecasting, battery scheduling, hybrid system operation and short-term PV dispatch. The regional error pattern is itself an operational signal: more uncertainty is likely in northern Iraq, and the simpler forecasting structures may be sufficient in the more stable conditions in the center and south.

4. CONCLUSIONS

Day-ahead forecasting of global horizontal irradiance (GHI) was developed for the towns of Najaf, Basra and Erbil based on the daily NASA POWER data for the 2020–2025 period. The analysis was aimed at investigating regional forecasting behavior when leakage is imposed and to benchmark the machine-learning models with a persistence model.

Spatial and seasonal differences were clearly discernible. Erbil had the lowest annual mean and the strongest year-to-year fluctuation in GHI while Basra had the highest mean daily GHI and Najaf had the highest mean annual GHI.

Reasonable day-ahead accuracy was achieved using solar, meteorological, temporal, lagged, and rolling predictors. Despite its non-statistically significant advantage over XG Boost, linear regression had the lowest error in the 2025 test year (MAE = 0.486 kWh/m²/day, RMSE = 0.723 kWh/m²/day, MAPE = 14.974%, R² = 0.860). Practical equivalence was the most conservative result as after adding leakage-controlled temporal features, the simple transparent model could compete with more complex ensemble methods.

At city level, the best performance was obtained in Najaf, followed by Basra, while Erbil produced the largest errors. The weaker performance in Erbil was consistent with its stronger seasonal cycle, higher rainfall, and less stable irradiance regime.

Same-day GHI, lagged and rolling irradiance terms, seasonal indicators, surface pressure, temperature, and humidity were identified as the strongest predictors. Taken together, the results suggest that leakage-controlled and interpretable machine-learning methods can provide a practical basis for day-ahead solar forecasting, photovoltaic planning, and renewable-energy decision-making in Iraq.

REFERENCES

- 1) Diagne, M., David, M., Lauret, P., Boland, J., & Schmutz, N. (2013). Review of solar irradiance forecasting methods and a proposition for small-scale insular grids. *Renewable and Sustainable Energy Reviews*, 27, 65-76. <https://doi.org/10.1016/j.rser.2013.06.042>
- 2) Inman, R. H., Pedro, H. T. C., & Coimbra, C. F. M. (2013). Solar forecasting methods for renewable energy integration. *Progress in Energy and Combustion Science*, 39(6), 535-576. <https://doi.org/10.1016/j.pecs.2013.06.002>
- 3) Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., Martinez-de-Pison, F. J., & Antonanzas-Torres, F. (2016). Review of photovoltaic power forecasting. *Solar Energy*, 136, 78-111. <https://doi.org/10.1016/j.solener.2016.06.069>
- 4) Voyant, C., Notton, G., Kalogirou, S., Nivet, M. L., Paoli, C., Motte, F., & Fouilloy, A. (2017). Machine learning methods for solar radiation forecasting: A review. *Renewable Energy*, 105, 569-582. <https://doi.org/10.1016/j.renene.2016.12.095>
- 5) Barbieri, F., Rajakaruna, S., & Ghosh, A. (2017). Very short-term photovoltaic power forecasting with cloud modeling: A review. *Renewable and Sustainable Energy Reviews*, 75, 242-263. <https://doi.org/10.1016/j.rser.2016.10.068>
- 6) Yagli, G. M., Yang, D., & Srinivasan, D. (2019). Automatic hourly solar forecasting using machine learning models. *Renewable and Sustainable Energy Reviews*, 105, 487-498. <https://doi.org/10.1016/j.rser.2019.02.006>
- 7) Al-Maliki, L. A., Al-Mamoori, S. K., El Tawil, K., Al-Ansari, N., & Comair, F. G. (2025). Assessing the accuracy of NASA Power meteorological data in Iraq. *Tikrit Journal of Engineering Sciences*, 32(4), Article 2163. <https://doi.org/10.25130/tjes.32.4.12>
- 8) Khaleel, M. H., Al-Rukabie, J. S. A., Al-Jiboori, M. H., & Al-Ramahy, Z. A. (2025). Relationships between daily solar irradiance and maximum temperature in Iraq. *Journal of Agrometeorology*, 27(1), 67-72. <https://doi.org/10.54386/jam.v27i1.2708>
- 9) Al-Ghezi, M. K., Mahmoud, B. K., Alnasser, T. M., & Chaichan, M. T. (2022). A comparative study of regression models and meteorological parameters to estimate the global solar radiation on a horizontal surface for Baghdad City, Iraq. *International Journal of Renewable Energy Development*, 11(1), 71-81. <https://doi.org/10.14710/ijred.2022.38493>
- 10) Hussain, M. T., & Mahdi, E. J. (2018). Assessment of solar photovoltaic potential in Iraq. *Journal of Physics: Conference Series*, 1032, 012007. <https://doi.org/10.1088/1742-6596/1032/1/012007>
- 11) Beck, H. E., Zimmermann, N. E., McVicar, T. R., Vergopolan, N., Berg, A., & Wood, E. F. (2018). Present and future Köppen-Geiger climate classification maps at 1-km resolution. *Scientific Data*, 5, 180214. <https://doi.org/10.1038/sdata.2018.214>
- 12) Ahmad, T., & Chen, H. (2020). A review on machine learning forecasting growth trends and their real-time applications in different energy systems. *Sustainable Cities and Society*, 54, 102010. <https://doi.org/10.1016/j.scs.2019.102010>
- 13) Ahmed, R., Sreeram, V., Mishra, Y., & Arif, M. D. (2020). A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization. *Renewable and Sustainable Energy Reviews*, 124, 109792. <https://doi.org/10.1016/j.rser.2020.109792>

Cross-Regional Solar Irradiance Forecasting in Iraq Using Explainable Machine Learning: Evidence from Najaf, Basra, and Erbil

- 14) Mellit, A., Massi Pavan, A., Oglari, E., Leva, S., & Lughi, V. (2020). Advanced methods for photovoltaic output power forecasting: A review. *Applied Sciences*, 10(2), 487. <https://doi.org/10.3390/app10020487>
- 15) Iheanetu, K. J. (2022). Solar photovoltaic power forecasting: A review. *Sustainability*, 14(24), 17005. <https://doi.org/10.3390/su142417005>
- 16) Yang, D., Wang, W., Gueymard, C. A., Hong, T., Kleissl, J., Huang, J., Perez, M. J., Perez, R., Bright, J. M., Xia, X., van der Meer, D., & Peters, I. M. (2022). A review of solar forecasting, its dependence on atmospheric sciences and implications for grid integration: Towards carbon neutrality. *Renewable and Sustainable Energy Reviews*, 161, 112348. <https://doi.org/10.1016/j.rser.2022.112348>
- 17) Chodakowska, E., Nazarko, J., & Nazarko, L. (2024). Solar radiation forecasting: A systematic meta-review of methods and trends. *Energies*, 17(13), 3156. <https://doi.org/10.3390/en17133156>
- 18) Chu, Y., Wang, Y., Yang, D., Chen, S., & Li, M. (2024). A review of distributed solar forecasting with remote sensing and deep learning. *Renewable and Sustainable Energy Reviews*, 198, 114391. <https://doi.org/10.1016/j.rser.2024.114391>
- 19) Ağbulut, U., Gürel, A. E., & Biçen, Y. (2021). Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison. *Renewable and Sustainable Energy Reviews*, 135, 110114. <https://doi.org/10.1016/j.rser.2020.110114>
- 20) Huang, L., Kang, J., Wan, M., Fang, L., Zhang, C., & Zeng, Z. (2021). Solar radiation prediction using different machine learning algorithms and implications for extreme climate events. *Frontiers in Earth Science*, 9, 596860. <https://doi.org/10.3389/feart.2021.596860>
- 21) Noor Sami Razzaq Najjar. Machine Learning Techniques for Enhanced Prediction of Diabetes Mellitus.; *Al-Harf Journal*., No. 25., Vol. 2, May 2026
- 22) Jebli, I., Belouadha, F. Z., Kabbaj, M. I., & Tilioua, A. (2021). Prediction of solar energy guided by Pearson correlation using machine learning. *Energy*, 224, 120109. <https://doi.org/10.1016/j.energy.2021.120109>
- 23) Feng, Y., Gong, D., Zhang, Q., Jiang, S., Zhao, L., & Cui, N. (2019). Evaluation of temperature-based machine learning and empirical models for predicting daily global solar radiation. *Energy Conversion and Management*, 198, 111780. <https://doi.org/10.1016/j.enconman.2019.111780>
- 24) Qiu, R., Liu, C., Cui, N., Gao, Y., Li, L., Wu, Z., Jiang, S., & Hu, M. (2022). Generalized extreme gradient boosting model for predicting daily global solar radiation for locations without historical data. *Energy Conversion and Management*, 258, 115488. <https://doi.org/10.1016/j.enconman.2022.115488>
- 25) Nadeem, T. B., Ali, S. U., Asif, M., & Suberi, H. K. (2024). Forecasting daily solar radiation: An evaluation and comparison of machine learning algorithms. *AIP Advances*, 14(7), 075010. <https://doi.org/10.1063/5.0211723>
- 26) Muhanad Tarish Abed. (2025). The Reality of Educational Technology Use Among Arabic Language Teachers in Intermediate Schools in Babylon Governorate., *Al-Harf Journal*., No. 24 ., Vol. 3, August 2025
- 27) Imam, A. A., Abusorrah, A., Seedahmed, M. M. A., & Marzband, M. (2024). Accurate forecasting of global horizontal irradiance in Saudi Arabia: A comparative study of machine learning predictive models and feature selection techniques. *Mathematics*, 12(16), 2600. <https://doi.org/10.3390/math12162600>
- 28) Noura, H. N., Allal, Z., Salman, O., & Chahine, K. (2025). Efficient machine learning-based prediction of solar irradiance using multi-site data. *Future Internet*, 17(8), 336. <https://doi.org/10.3390/fi17080336>
- 29) Demir, V. (2025). Evaluation of solar radiation prediction models using AI: A performance comparison in the high-potential region of Konya, Türkiye. *Atmosphere*, 16(4), 398. <https://doi.org/10.3390/atmos16040398>
- 30) Taha, A., Makeen, P., & Nazih, N. (2025). Short-term and long-term solar irradiance forecasting with advanced machine learning techniques in Zafarana, Egypt. *Scientific Reports*, 15, 39553. <https://doi.org/10.1038/s41598-025-24853-4>
- 31) Yadav, A. K., Kumar, R., Wang, M., Fekete, G., & Singh, T. (2025). Comparative analysis of daily global solar radiation prediction using deep learning models inputted with stochastic variables. *Scientific Reports*, 15, 10786. <https://doi.org/10.1038/s41598-025-95281-7>
- 32) Bakır, H., et al. (2024). Prediction of daily global solar radiation in different climatic regions using machine learning and optimization algorithms. *Environmental Science and Pollution Research*. <https://doi.org/10.1007/s11356-024-33785-x>
- 33) Lara-Benitez, P., Carranza-Garcia, M., Luna-Romera, J. M., & Riquelme, J. C. (2023). Short-term solar irradiance forecasting in streaming with deep learning. *Neurocomputing*, 546, 126312. <https://doi.org/10.1016/j.neucom.2023.126312>
- 34) Ribeiro, R., & Fanzeres, B. (2024). Identifying representative days of solar irradiance and wind speed in Brazil using machine learning techniques. *Energy and AI*, 15, 100320. <https://doi.org/10.1016/j.egyai.2023.100320>
- 35) Haider Hasan Ali Aluqabi. (2025). The impact of Using Educational Trees and Posters on Improving Students' Skills in Understanding Arabic Language Lessons: an applied study., *Al-Harf Journal*., No. 24., Vol. 1, March 2025

Cross-Regional Solar Irradiance Forecasting in Iraq Using Explainable Machine Learning: Evidence from Najaf, Basra, and Erbil

- 36) Al-lahham, A., Theeb, O., Elalem, K., Alshaw, T. A., & Alshebeili, S. A. (2020). Sky imager-based forecast of solar irradiance using machine learning. *Electronics*, 9(10), 1700. <https://doi.org/10.3390/electronics9101700>
- 37) Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>
- 38) Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273-297. <https://doi.org/10.1007/BF00994018>
- 39) Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14, 199-222. <https://doi.org/10.1023/B:STCO.0000035301.49549.88>
- 40) Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). <https://doi.org/10.1145/2939672.2939785>
- 41) Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, 30, 3146-3154.
- 42) NASA POWER Project. (n.d.). POWER Data Access Viewer. NASA Langley Research Center. <https://power.larc.nasa.gov/data-access-viewer/>
- 43) NASA POWER Project. (n.d.). POWER Parameter Dictionary. NASA Langley Research Center. <https://power.larc.nasa.gov/docs/tutorials/parameters/>
- 44) Zhang, T., Chandler, W. S., Hoell, J. M., Westberg, D., Whitlock, C. H., & Stackhouse, P. W. (2009). A global perspective on renewable energy resources: NASA's Prediction of Worldwide Energy Resources (POWER) Project. In *Proceedings of ISES World Congress 2007* (pp. 2636-2640). Springer. https://doi.org/10.1007/978-3-540-75997-3_532
- 45) Sparks, A. H. (2018). Nasa power: A NASA POWER global meteorology, surface solar energy and climatology data client for R. *Journal of Open Source Software*, 3(30), 1035. <https://doi.org/10.21105/joss.01035>
- 46) Sayago, S., Ovando, G., Almorox, J., & Bocco, M. (2020). Daily solar radiation from NASA-POWER product: Assessing its accuracy considering atmospheric transparency. *International Journal of Remote Sensing*, 41(3), 897-910. <https://doi.org/10.1080/01431161.2019.1650986>
- 47) Quansah, A. D., Dogbey, F., Asilevi, P. J., Boakye, P., Darkwah, L., Oduro-Kwarteng, S., Sokama-Neuyam, Y. A., & Mensah, P. (2022). Assessment of solar radiation resource from the NASA-POWER reanalysis products for tropical climates in Ghana towards clean energy application. *Scientific Reports*, 12, 10684. <https://doi.org/10.1038/s41598-022-14126-9>
- 48) Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 30, 4765-4774.
- 49) Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2, 56-67. <https://doi.org/10.1038/s42256-019-0138-9>
- 50) Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
- 51) Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences* (3rd ed.). Academic Press.
- 52) Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd ed.). OTexts.