Interpretable cloud cover prediction for dynamic solar prediction in low-data environments

Nitipon Pongphaw[1*], Keerati Maneesai[2], Pantong Sanonok[1], Prommin Buaphan[1]

[1] Department of Electrical and Computer Engineering, Faculty of Science and Engineering, Kasetsart University Chalermphrakiat Sakon Nakhon Province Campus, Sakon Nakhon, Thailand

[2] Department of General Science, Faculty of Science and Engineering, Kasetsart University Chalermphrakiat Sakon Nakhon Province Campus, Sakon Nakhon, Thailand

Abstract

Cloud cover strongly influences solar irradiance variability, directly affecting photovoltaic (PV) energy generation. Rapid fluctuations in cloud properties such as type, thickness, and movement can cause unpredictable drops in solar output, challenging grid reliability and energy dispatch. Accurate cloud cover prediction is often hindered by sparse meteorological data, particularly in geographically remote or sensor-deficient regions. To address this, we propose a framework that employs Gaussian Mixture Models (GMMs) to generate physically consistent synthetic meteorological data, augmenting limited training datasets. This approach is applied to tree-based machine learning models, including Random Forest, CatBoost, and XGBoost, with SHAP (SHapley Additive exPlanations) integrated to enhance interpretability. Experimental results show improved accuracy and robustness, with Random Forest achieving a Mean Absolute Error (MAE) of 11.7767 ± 0.1091, Root Mean Square Error (RMSE) of 17.2762 ± 0.2604, and R^2 of 0.8092 ± 0.0043. SHAP analyses reveal more stable feature contributions, particularly for dew point and relative humidity. This framework has significant practical value for solar forecasting in Southeast Asian regions with limited sensor networks, enabling accurate cloud cover prediction, improved grid reliability, and scalable edge deployment for solar energy integration.

Keywords: GMMs-enhanced prediction; synthetic meteorological intelligence; SHAP-driven cloud insight; low-cost AI for cloud prediction

1. Introduction

The increasing integration of solar energy into modern power systems requires highly accurate solar irradiance forecasting, which enhances grid reliability and supports sustainability goals [1].

*Corresponding author: nitipon.p@ku.th
Received: 2025-05-29 Accepted: 2025-12-05 Published: 2025-12-12
Journal of Asian Energy Studies (2025), Vol 9, 245-267, doi:10.24112/jaes.090014

Among the environmental factors affecting PV performance, cloud cover plays a dominant and rapidly changing role, as it directly controls the sunlight reaching the ground. Input data typically come from on-site sensors, with solar irradiance and temperature as primary features [2]. Variations in cloud type, thickness, and movement cause rapid fluctuations in energy output, challenging grid stability, particularly in off-grid and hybrid systems [3, 4].

However, predictive models are often limited by scarce high-quality cloud cover data, especially in remote or sensor-sparse regions, leading to poor generalization and reduced accuracy [5]. State-of-the-art ML architectures, such as deep neural networks or CNN-LSTM hybrids, require large datasets and significant computational resources, making them impractical for resource-constrained edge devices. To address data scarcity, various augmentation techniques have been proposed. SMOTE is widely used for imbalanced classification tasks [6], but it has limitations for continuous multivariate regression because it does not model joint feature distributions, risking the loss of multivariate correlations [6–8]. Simple data imputation is efficient but underestimates variability [7], while GANs and VAEs demand large amounts of data and high computational costs [8].

This study proposes a novel approach using GMMs to generate synthetic meteorological data by explicitly modeling the joint probability distribution of features via Gaussian mixtures, which can handle continuous multivariate data, capture complex dependencies, model multimodal distributions, generate samples from the learned distribution rather than interpolating, require minimal hyperparameter tuning, and are computationally efficient [9]. Additionally, SHAP is integrated to enhance model transparency and interpretability, quantifying the contributions of features such as dew point and relative humidity. This combined framework improves the reliability of cloud cover prediction and, consequently, the accuracy and applicability of solar energy forecasting in data-constrained environments.

2. Literature Review

2.1. Impact of cloud cover on solar irradiance

Cloud cover significantly impacts solar power generation, primarily through its effects on solar irradiance, which is the sunlight received at the Earth's surface and is crucial for PV performance. Variability in cloud cover can lead to fluctuations in solar energy output, necessitating accurate predictions for effective integration into power systems. This overview synthesizes recent findings on the influence of cloud cover on solar energy generation, highlighting key factors and methodologies that have emerged in the literature.

Cloud cover affects solar irradiance by reflecting, scattering, and absorbing sunlight. The complex interplay between cloud properties such as type, thickness, and coverage and solar radiation dynamics is well established [10]. For instance, research indicates that while low clouds can significantly block direct solar radiation, thinner clouds may allow some solar energy to pass through, thus influencing total irradiance levels [11, 12]. As such, understanding cloud characteristics is essential for enhancing solar power prediction. Numerous studies have quantified the effects of cloud cover on solar energy generation. Jadhav et al. noted that cloud cover can account for approximately 20-65% of the attenuation of global horizontal irradiance (GHI), impacting overall solar energy availability [12]. Furthermore, Bando et al. emphasized that different types of clouds, e.g., stratocumulus vs. cumulus, result in varying degrees of solar radiation reduction, which affects power output fluctuations in PV systems [13]. Interannual variability linked to cloud cover has also shown correlations with solar radiation levels, underscoring the importance of cloud data in long-term energy generation models [14].

2.2. Overview of ML for cloud cover prediction

ML models are increasingly used to predict solar irradiance based on cloud cover data, improving energy management in solar power systems. Accurate cloud cover prediction is crucial for enhancing solar power forecasts, as it directly affects PV system irradiance estimation. Various ML approaches, including supervised learning, advanced image processing, and deep learning frameworks, have proven effective in this domain. For instance, Kim et al. demonstrated that ground-based imaging combined with ML techniques, particularly SVM, achieves high accuracy in cloud cover calculations, making it suitable for nowcasting applications, while incorporating detailed image features further improves prediction accuracy [15]. Similarly, Park et al. utilized ML-based cloud cover estimates to predict solar irradiance, showing a clear correlation between predicted cloud cover and measured irradiance [16]. Deo et al. developed CNN-LSTM hybrid models to forecast photosynthetic photon flux density under cloud effects, effectively capturing temporal dynamics essential for short-term solar power forecasting [17]. Other advanced methods, such as K-Means clustering for cloud classification, have also shown promise, with accuracy validated using both satellite and ground-based data [18]. These ML techniques are particularly beneficial in managing the intermittency associated with cloud movements, which can cause rapid decreases in solar power output [16, 19]. Intra-hour forecasts are critical, as even small changes in cloud cover can lead to significant output variations [20], highlighting the need for refined nowcasting techniques during dynamic weather events [20, 21]. Combining satellite imagery with ML models improves the reliability of predictions, and adapting algorithms to various climatic conditions further enhances the efficiency and adaptability of solar power systems to local environmental factors [22, 23].

2.3. Overview of GMMs and handling limited data

GMMs serve as effective statistical tools for modeling complex data distributions, particularly in scenarios with limited data, a common situation in meteorological applications. The foundation of GMMs is the idea that a dataset can be represented as a combination of multiple Gaussian distributions, with each distribution contributing to the overall data pattern. This adaptability makes GMMs applicable across various tasks, from clustering to density estimation, and their multimodal capability enables more comprehensive modeling and meaningful insights even when only small datasets are available [15, 24–26].

In the context of cloud cover prediction, GMMs can infer the underlying data distribution using available features such as relative humidity, ambient temperature, and historical cloud patterns. Their ability to generalize from sparse observations makes them especially advantageous in sensor-deficient or remote regions where meteorological data are scarce. To enhance parameter estimation under these constraints, GMMs employ the Expectation-Maximization (EM) algorithm, an iterative optimization process that refines estimates until convergence is achieved [26–28]. This approach allows for accurate estimation of each Gaussian component while mitigating the risk of overfitting. Overall, the flexibility and robustness of GMMs make them an effective solution for modeling in data-constrained environments, particularly for applications such as solar irradiance forecasting and cloud classification.

2.4. Data scarcity and augmentation techniques

Limited availability of high-quality cloud cover data poses a significant challenge for developing accurate predictive models, particularly in remote or sensor-sparse regions [5]. To address data scarcity, various augmentation techniques have been proposed across different domains. One

widely used method is the Synthetic Minority Over-sampling Technique (SMOTE), originally designed for classification tasks with imbalanced datasets [6]. However, SMOTE has inherent limitations when applied to continuous regression tasks involving multivariate meteorological data. Specifically, it was developed for discrete class boundaries and does not naturally extend to continuous target variables without modification [6,7]. Additionally, SMOTE does not explicitly model the joint probability distribution of features, which may lead to the loss of complex multivariate correlations that are crucial in meteorological systems, where variables such as temperature, humidity, and pressure are interdependent [8].

Alternative approaches, such as simple data imputation, offer computational efficiency but tend to underestimate natural variability, introducing bias into the training process [7]. More advanced generative models, including GANs and VAEs, can produce realistic synthetic data but require large datasets, substantial computational resources, and careful hyperparameter tuning, making them less practical for deployment in data-constrained or edge environments [8].

Given these limitations, GMMs have emerged as a promising alternative. Unlike interpolation-based techniques such as SMOTE, GMMs explicitly model the joint probability distribution of meteorological features using a weighted mixture of Gaussian components. This probabilistic approach naturally handles continuous multivariate data, captures complex dependencies among correlated features, models multimodal distributions, and generates synthetic samples by sampling from the learned distribution rather than interpolating between existing points. GMMs also require minimal hyperparameter tuning and modest computational resources, making them suitable for real-time or edge applications [9]. Table 1 summarizes a comparison of these techniques in terms of their mechanism, physical realism, computational cost, data requirements, and compatibility with cloud physics. The potential of GMMs as both a data augmentation and predictive modeling tool provides a natural bridge to their direct application in cloud cover prediction, as discussed in the next section.

Table 1: *Comparison of data augmentation and generative techniques*

Technique	Computational Cost	Data Requirement	Cloud Physics Compatibility	Reference
SMOTE	Low	Moderate	Limited; for discrete classes	[6,7]
Simple Imputation	Very low	Low	Limited; may bias training	[7]
GANs	High	Large	Moderate; captures correlations	[8]
VAEs	High	Large	Moderate; captures dependencies	[8]
GMMs	Low	Moderate	High; handles continuous multivariate data	[9]

2.5. GMMs for cloud cover prediction

Given the limitations of conventional augmentation techniques, GMMs offer a robust probabilistic framework for cloud cover prediction in scenarios with limited labeled data [29,30]. By modeling the joint distribution of atmospheric variables such as relative humidity, temperature, and historical cloud patterns, GMMs can generate synthetic samples that capture complex dependencies and variability inherent in the atmosphere.

Beyond data augmentation, GMMs support probabilistic imputation for missing values, which is particularly valuable in environmental monitoring, where sensor gaps or cloud obstructions are common [26]. They can also be integrated into broader ML workflows, serving as preprocessing layers or components in ensemble models and neural networks, enhancing feature extraction and overall predictive performance [30].

In summary, GMMs provide a flexible, robust, and interpretable tool for cloud cover prediction under limited data conditions [16, 18, 26]. They facilitate uncertainty quantification, handle heterogeneous spatiotemporal data, and support hybridization with advanced ML methods,

making them suitable for applications in agriculture, aviation, climate modeling [31–33], and solar energy forecasting in low-data or sensor-deficient regions.

2.6. Research gaps and motivation

Despite recent advances in ML techniques for cloud cover prediction, two critical challenges continue to limit practical implementation: data scarcity and model complexity. High-quality, sufficiently large datasets are often unavailable, particularly in remote or developing regions, which reduces predictive accuracy and hinders generalization across different geographical and climatic conditions [34,35]. Most ML models rely heavily on large-scale satellite or ground-based datasets, and in data-scarce environments, these datasets are often incomplete or entirely missing.

In addition, many modern ML models, especially deep neural networks or hybrid architectures, are computationally intensive, requiring substantial processing power, memory, and energy. Such models are not only unsuitable for deployment on resource-constrained edge devices, like microcontrollers, but they also lack interpretability, limiting their practical usability in real-time operational settings [36–38]. Furthermore, few studies provide systematic comparisons of different ML algorithms under consistent experimental settings, making it difficult to evaluate trade-offs between model complexity, prediction accuracy, and deployability [30].

To address these gaps, data synthesis techniques, particularly GMMs, have been explored to augment training datasets. GMMs generate synthetic data that approximates the statistical distribution of real-world meteorological features, enabling models to learn effectively even with limited observations [9]. Alongside data augmentation, there is a growing need for model interpretability, which can be addressed using tools such as SHAP to quantify feature contributions and clarify model behavior under varying atmospheric conditions.

This study is motivated by these limitations and aims to develop efficient, interpretable cloud cover prediction models that balance accuracy with computational simplicity, making them suitable for deployment in edge and resource-constrained environments. By leveraging GMMs for data augmentation and incorporating SHAP for transparency, this work addresses the dual challenges of data scarcity and model complexity. Additionally, we perform a comprehensive comparison of multiple ML algorithms to identify optimal models for practical, real-world scenarios, bridging the gap between experimental research and deployable cloud cover prediction systems.

3. Methodology

This section provides an overview of the proposed methodology, including data preparation, parameter configuration, and preprocessing procedures applied prior to model training and evaluation. Two experimental pipelines were implemented: one incorporating GMMs for synthetic data augmentation and another using only the original dataset without augmentation. The overall workflow is illustrated in Figure 1 (GMMs-augmented pipeline) and Figure 2 (non-GMMs pipeline).

3.1. Data

This study utilizes a high-resolution time-series weather dataset obtained from Kaggle [39], which contains hourly meteorological records spanning from 1980 to 2024. The data file used in this research, Weather_Data_1980_2024 (hourly).csv, offers a comprehensive collection of atmospheric variables suitable for temporal modeling and forecasting applications. For the purpose of this study, a subset of nine key features was selected based on their relevance to short-term and

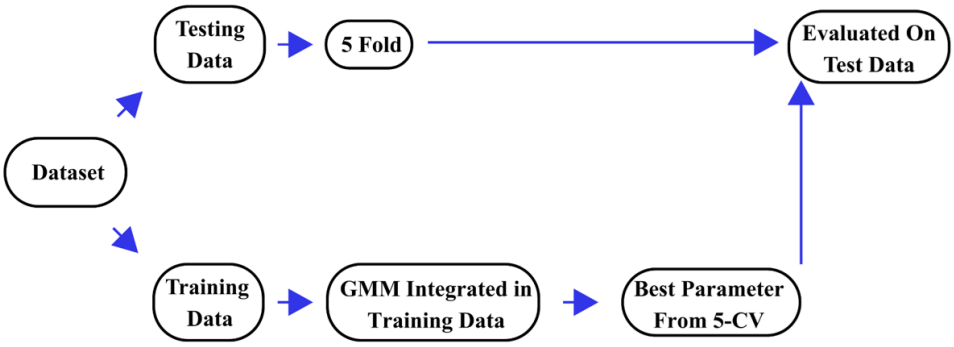

Figure 1: *Flowchart of the methodology using GMMs*

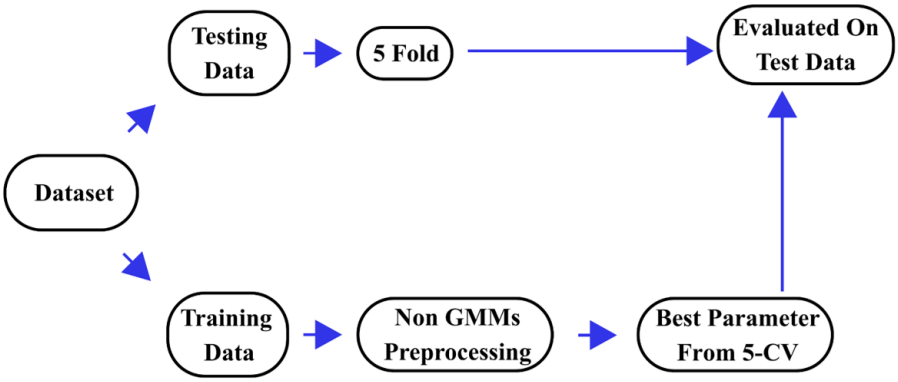

Figure 2: *Flowchart of the methodology not using GMMs*

long-term weather prediction. These features include temperature, relative humidity, dew point, precipitation (mm), rain (mm), mean sea-level pressure (pressure_msl in hPa), surface pressure (hPa), vapor pressure deficit (kPa), and wind speed at 10 meters (km/h). These variables were chosen due to their strong influence on weather dynamics and their frequent use in predictive modeling tasks across climate and environmental sciences.

3.2. Tree based model

This study employs several tree-based algorithms, including decision trees (DT), random forests (RF), extra trees (ET), and gradient boosting (GB) methods such as XGBoost (XGB) and catboost (CB). These models are well-suited for capturing complex nonlinear relationships within data. Ensemble techniques like RF and GB combine multiple DT to enhance predictive accuracy and mitigate overfitting, thereby improving the models' generalization to unseen data.

These algorithms have demonstrated robust performance in environmental modeling and weather forecasting, where accurately representing intricate interactions among meteorological variables is critical. Moreover, tree-based models offer practical advantages for deployment in resource-constrained environments, making them highly applicable for edge AI implementations. Their relatively lightweight architecture and efficient inference enable integration into edge devices

such as IoT sensors and embedded systems, facilitating real-time data processing and prediction without reliance on centralized computing resources [40]. This capability is particularly valuable in remote or distributed settings where low latency and limited connectivity pose challenges for conventional cloud-based solutions.

3.3. Hyperparameter tuning

Hyperparameter optimization was conducted to ensure that all tree-based regression models were fairly and consistently calibrated across both data scenarios, with and without GMM-based synthetic augmentation. GridSearchCV with 5-fold cross-validation was applied to exhaustively evaluate predefined parameter grids for each model. For CatBoost, the parameters iterations $\in \{200, 300\}$ and depth $\in \{5, 10\}$ were tuned, while for Random Forest and Extra Trees, the search focused on n_estimators $\in \{200, 300\}$ and max_depth $\in \{5, 10\}$. The Decision Tree model tuned only max_depth $\in \{5, 10\}$, reflecting its simpler structure, and XGBoost was optimized over n_estimators $\in \{200, 300\}$ and max_depth $\in \{5, 10\}$. These grids were applied uniformly to both datasets to ensure that the resulting best estimators were driven solely by intrinsic model characteristics rather than differences in data availability. Following hyperparameter tuning, the selected best models were evaluated on the independent test set using a 5-fold evaluation procedure, where the test data were partitioned into five subsets, and each subset was used once for evaluation. Importantly, no additional training occurred during this stage; all models evaluated on the test folds were the previously tuned models trained exclusively on the training data. This approach provides a stable estimate of generalization performance, reduces fold-specific variance, and enables consistent benchmarking between models trained on real data only and those trained using real + GMM-generated synthetic samples. By standardizing preprocessing, feature selection, cross-validation structure, and hyperparameter search space across all experimental conditions, the study establishes a rigorous and reproducible methodology for assessing how GMM-based augmentation affects predictive accuracy, robustness, and overall model behavior under real-world data constraints.

3.4. Data preprocessing

A structured data preprocessing pipeline was implemented to address distributional imbalances in the cloud cover variable, which plays a pivotal role in solar irradiance prediction. The raw dataset, comprising hourly meteorological records from 1980 to 2024, was first loaded into a pandas DataFrame. To preserve the statistical characteristics of cloud cover across both training and test subsets, the target variable cloud_cover (%) was discretized into ten quantile-based bins using **qcut**. This approach was selected because quantile discretization creates equal-sized groups, improves the stability of subsequent sampling, and mitigates issues caused by skewed or long-tailed cloud cover distributions. Discretization also enabled the application of stratified sampling, which is particularly beneficial for this dataset because certain cloud cover levels occur far more frequently than others. Without stratification, random splitting would risk under-representing rare but meteorologically meaningful cloud regimes, potentially biasing the models and reducing generalization performance.

A total of 18,000 samples were then randomly selected to balance data diversity with computational efficiency. Stratified sampling based on the binned cloud-cover categories was used to divide the dataset into a training set of 3,000 records and a test set of 15,000 records. This procedure ensured proportional representation of all cloud-cover strata in both subsets, thereby reducing sampling bias, improving fairness, and enhancing model robustness, especially under

data-scarce conditions. After the split, the auxiliary binning column used solely for stratification was removed to prevent information leakage during model training. Finally, the processed training and test sets were exported in CSV format for downstream model development and evaluation.

Non-GMMs: The data preprocessing began with loading the training and testing datasets from CSV files and verifying their dimensions. Invalid records with missing or infinite values were removed, and the features were normalized to ensure a zero mean and unit variance. This structured workflow, summarized in Figure 3, ensures that all models are trained on complete, standardized, and reliable inputs, supporting robust and valid predictive evaluations.

GMMs: GMMs were employed to perform clustering and density estimation on the meteorological dataset as part of the synthetic data generation pipeline. The process began by loading the raw data from a CSV file, including key features such as temperature, relative humidity, dew point, precipitation, rainfall, mean sea-level pressure, surface pressure, vapor pressure deficit, 10-meter wind speed, and cloud cover percentage. To ensure numerical stability during model fitting, missing values (NaNs) were handled using a simple imputation strategy in which each NaN was replaced by the mean of its corresponding feature. The dataset was then normalized using StandardScaler, transforming each feature to have zero mean and unit variance. This standardization ensures comparable feature scales and improves the convergence properties of the GMMs optimization process. This workflow forms the foundation of the GMMs-based augmentation strategy illustrated holistically in Figure 4, which depicts the end-to-end sequence from raw data ingestion through imputation, normalization, and Gaussian mixture modeling. By structuring the preprocessing pipeline in this manner, the study ensures that the fitted GMMs accurately capture the underlying joint distribution of meteorological variables, leading to reliable and physically consistent synthetic sample generation.

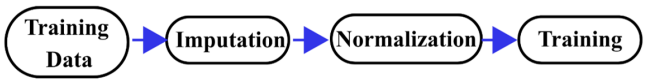

Figure 3: *Preprocessing workflow for non-GMMs datasets*

Figure 4: *Preprocessing workflow for GMMs datasets*

To determine the optimal number of gaussian components, we conducted an extensive model selection procedure by fitting GMMs with varying numbers of components from 1 to 100. For each model, bayesian information criterion (BIC) and akaike information criterion (AIC) scores were computed, serving as penalized likelihood measures to balance model complexity and goodness of fit. The resulting BIC and AIC values were visualized across the component range to identify the optimal cluster number minimizing these criteria. According to this analysis, the number of components corresponding to the lowest BIC and AIC values was selected as the best model configuration for subsequent clustering and synthetic data generation tasks, which yielded AIC = 62 and BIC = 17, as illustrated in Figure 5.

According to Figure 5, to determine the optimal number of components in the GMMs, both BIC and AIC were compared. While BIC reached its minimum at 17 components, AIC attained its lowest value at 62 components, indicating a preference for a more expressive model; therefore, the AIC-optimal solution of 62 components was adopted in this study. Although BIC is known

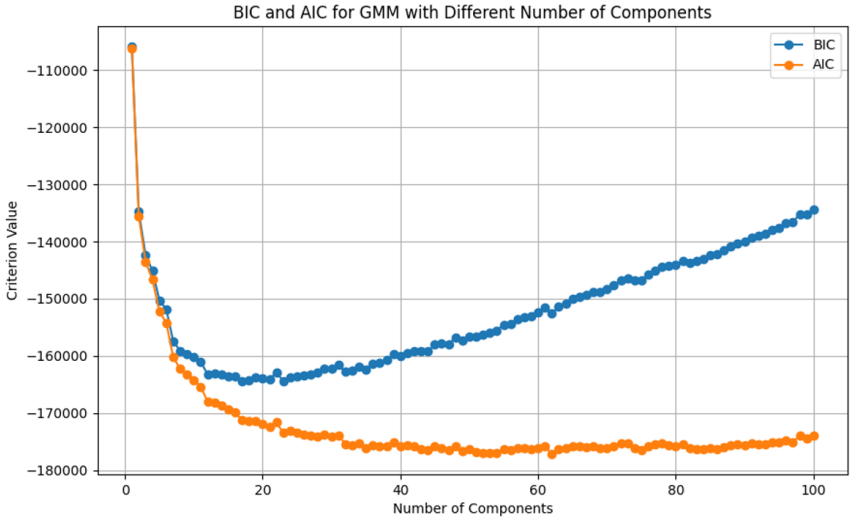

Figure 5: *Comparison between AIC and BIC*

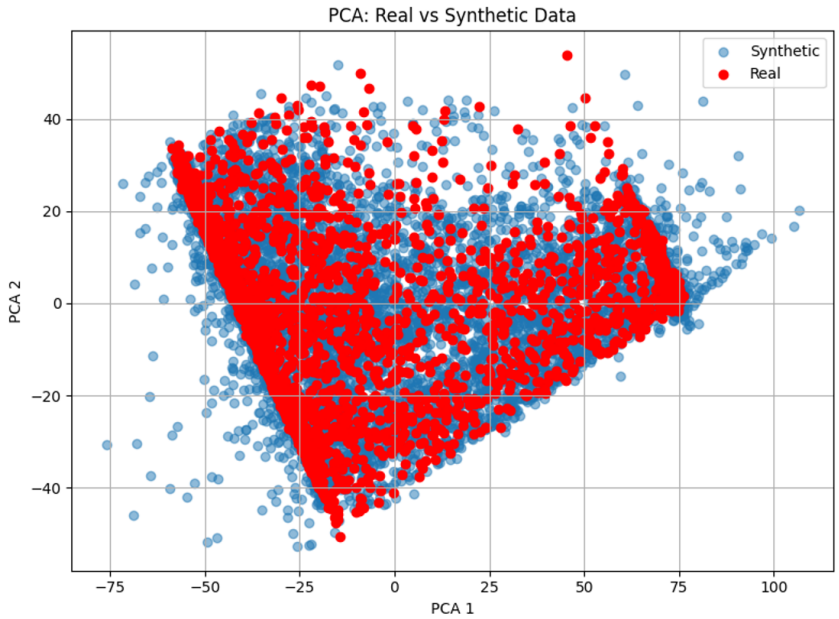

Figure 6: *Comparison of real vs. synthetic data in PCA-reduced space*

for its conservative nature and asymptotic consistency, AIC is often more effective in capturing fine-grained structures and subtle latent patterns, particularly when the primary objective is density estimation or synthetic data generation rather than model parsimony. After selecting the optimal number of Gaussian components, a GMMs was trained using the identified 62 components on the Min-Max normalized and NaN-imputed dataset, and 10,000 synthetic samples were subsequently generated to augment the original data. Principal Component Analysis (PCA) was then applied to reduce the dimensionality of the combined real and synthetic datasets to

two principal components for visualization. As shown in Figure 6, the resulting scatter plot demonstrates substantial overlap between real and synthetic data distributions in the PCA space, indicating that the GMMs successfully captured the underlying structure of the real data and produced representative synthetic samples.

To assess the structural fidelity of the synthetic data relative to the real dataset, PCA was conducted, and the projection onto the first two principal components (PC1 and PC2) was visualized. As shown in Figure 6, both real (red) and synthetic (blue) samples exhibit broadly overlapping distributions in the reduced dimensional space, indicating that the generative model successfully captured the principal variance directions of the real data. The synthetic samples, however, span a wider region along PC1 and PC2, suggesting increased variance or noise, whereas the real data occupy a more compact subspace. This discrepancy may reflect either over-generalization by the generative model or mode-collapse artifacts in regions with sparse real-data support. Although the density of real samples within the central manifold suggests that key distributional modes are preserved, peripheral structures appear underrepresented in the synthetic data.

This compares the distributions of cloud cover (Figure 7), dew point (Figure 8), and relative humidity (Figure 9). The synthetic data broadly resemble the real data, but notable issues remain in terms of physical realism and boundary adherence. For cloud cover and relative humidity, the synthetic distributions are overly smoothed and occasionally produce values outside the valid 0–100% range, unlike the real data, which show sharp peaks and clear boundary behavior. Dew point is modeled more accurately, capturing its bimodal structure and overall density shape, though the tails slightly exceed the plausible range. Overall, the model captures general statistical patterns but struggles with physical constraints and extreme-value behavior, indicating the need for physics-aware constraints and boundary-preserving mechanisms to improve realism.

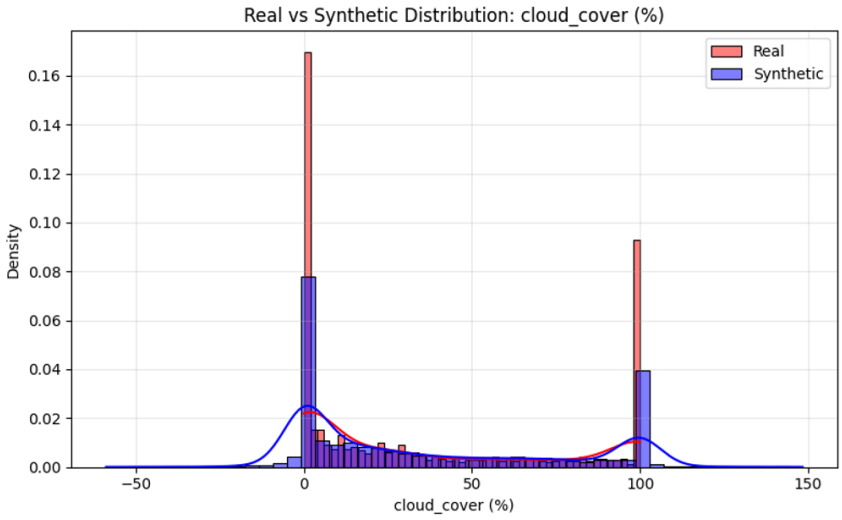

Figure 7: *Distribution of real vs synthetic data in cloud_cover (%)*

For subsequent modeling, the data integration process involved loading the original training data (train.csv) and the GMMs-generated synthetic data (syn_train.csv), importing both into pandas DataFrames, and concatenating them into a unified dataset that was exported as combined.csv. This merged dataset, containing both real and synthetic samples, was then used as the training set, while an independent test set was sourced from test.csv. The feature set consisted of

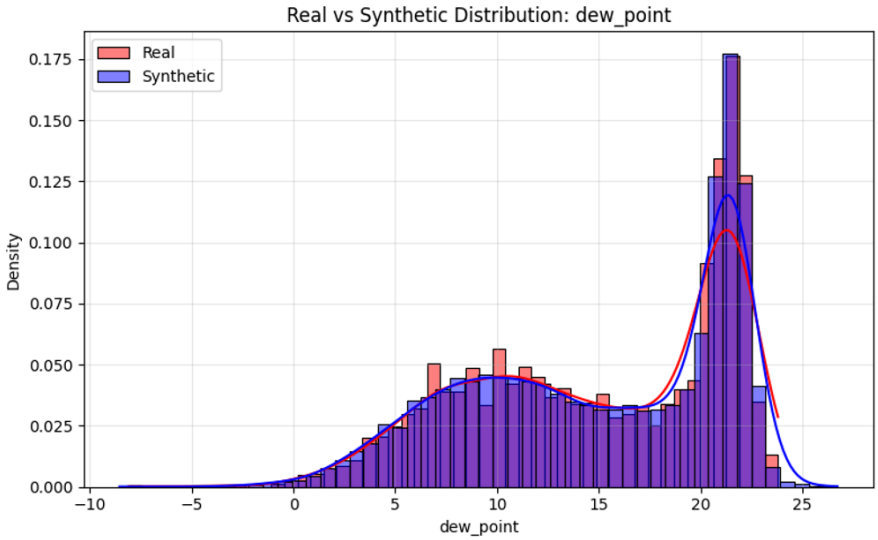

Figure 8: *Distribution of real vs synthetic data in dew_point*

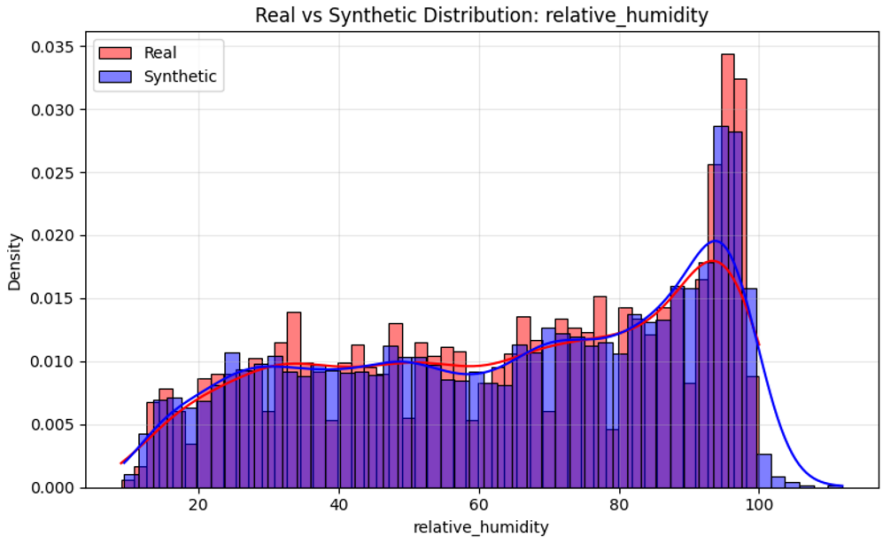

Figure 9: *Distribution of real vs synthetic data in relative_humidity*

meteorological variables known to influence cloud cover, including temperature, relative humidity, dew point, precipitation, rainfall, mean sea-level pressure, surface pressure, vapor pressure deficit, and wind speed at 10 meters, with the target variable defined as cloud-cover percentage. Before training, thorough preprocessing was performed to ensure dataset integrity, with NaN and infinite values systematically identified and removed from both the training and test sets.

3.5. Evaluate metrics

Model performance was evaluated using three common regression metrics: MAE, which measures the average absolute difference between predictions and actual values; R^2, which indicates the

proportion of variance explained by the model; and RMSE, which emphasizes larger errors by squaring the deviations before averaging.

4. Results

The comparison results are presented in Table 2 and Figure 10-14 to clearly illustrate the differences in model performance. Table 2 provides the evaluation metrics, including MAE, RMSE, and R^2, for each model, while Figure 10-14 visualizes the results graphically, allowing for a more intuitive understanding of the trends and the relationship between predicted and actual values. Presenting the results in both tabular and graphical formats enhances the clarity and credibility of the experimental findings.

Table 2: *Comparative performance of CB, XGB, RF, ET, and DT models with and without GMMs-based supervised learning*

Model	GMMs	MAE	RMSE	R^2
CB	Yes	11.9785 ± 0.1054	17.3127 ± 0.2604	0.8084 ± 0.0047
CB	No	12.1090 ± 0.1208	17.4298 ± 0.2493	0.8058 ± 0.0044
CB delta ± std	–	-0.1305 ± 0.1603	-0.1171 ± 0.3606	$+0.0026 \pm 0.0064$
XGB	Yes	12.3204 ± 0.0880	17.9425 ± 0.2356	0.7942 ± 0.0048
XGB	No	12.8699 ± 0.2090	19.1566 ± 0.3479	0.7653 ± 0.0075
XGB delta ± std	–	-0.5495 ± 0.2268	-1.2141 ± 0.4203	$+0.0289 \pm 0.0089$
RF	Yes	11.7767 ± 0.1091	17.2762 ± 0.2604	0.8092 ± 0.0043
RF	No	11.9498 ± 0.1187	17.3730 ± 0.2677	0.8070 ± 0.0048
RF delta ± std	–	-0.1731 ± 0.1610	-0.0968 ± 0.3740	$+0.0022 \pm 0.0064$
ET	Yes	12.1870 ± 0.1161	17.4017 ± 0.2323	0.8064 ± 0.0039
ET	No	12.0513 ± 0.1008	17.2042 ± 0.2332	0.8108 ± 0.0036
ET delta ± std	–	$+0.1357 \pm 0.1540$	$+0.1975 \pm 0.3290$	-0.0044 ± 0.0053
DT	Yes	12.4412 ± 0.1486	19.1079 ± 0.2504	0.7665 ± 0.0050
DT	No	12.6141 ± 0.1736	18.4647 ± 0.2325	0.7820 ± 0.0035
DT delta ± std	–	-0.1729 ± 0.2285	$+0.6432 \pm 0.3419$	-0.0155 ± 0.0061

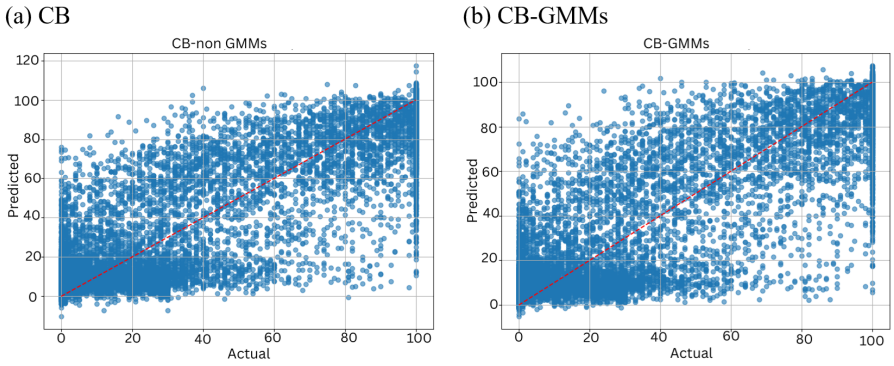

Figure 10: *Performance comparison of five ML models, including: (a) CB, and (b) CB-GMMs under actual value and predicted value*

Table 2 and Figures 10-14 summarize the performance of five ML models (CB, XGB, RF, ET, DT) using MAE, RMSE, and R^2. RF (Figure 12), trained with GMMs-augmented data, achieved the lowest MAE (11.78 ± 0.11) and a high R^2 (0.809 ± 0.004), demonstrating strong accuracy and stability, while CB (Figure 10) and XGB (Figure 11) also improved with GMM augmentation. In contrast, ET (Figure 13) and DT (Figure 14) showed minor gains or worsened performance, reflecting their sensitivity to synthetic data. GMMs augmentation reduced MAE for RF (-0.17), CB (-0.13), and

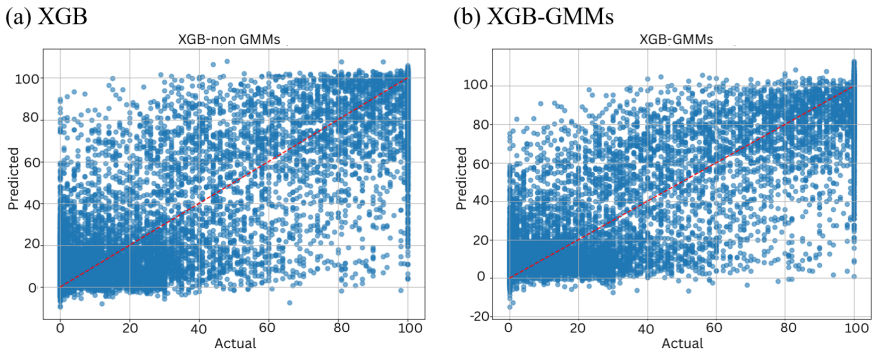

Figure 11: *Performance comparison of five ML models, including: (a) XGB, and (b) under actual value and predicted value*

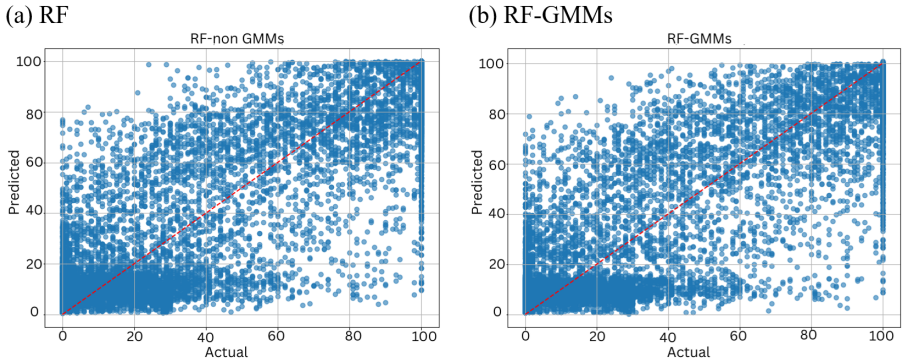

Figure 12: *Performance comparison of five ML models, including: (a) RF, and (b) RF-GMMs under actual value and predicted value*

XGB (-0.55), indicating effective supplementation of the training set without introducing significant noise. RF's superior results are attributed to its bootstrap aggregation and randomized feature selection, which enhance robustness in low-data scenarios and prevent overfitting, explaining its consistent performance with or without synthetic data. These findings directly address the challenge of limited high-quality cloud cover data highlighted in the introduction, showing that RF combined with GMMs augmentation can reliably predict cloud cover in remote or sensor-sparse regions, thus providing practical value for solar energy forecasting in real-world low-data environments.

Notably, DT displayed piecewise constant approximation patterns when trained without GMMs augmentation. This means that the model predicted outputs in discrete steps rather than continuously, creating a "stair-step" effect where all inputs within a certain range received the same predicted value. Such behavior is inherent to DT because it splits the input space into distinct regions, assigning a constant output to each. While simple and interpretable, this step-wise pattern cannot capture smooth trends or gradual changes, often resulting in abrupt predictions. After GMMs integration, these patterns diminished, producing more continuous and stable predictions. This improvement occurs because the GMMs-generated synthetic samples enrich the training set with structurally informative data, allowing the DT to better approximate gradual changes in the underlying distribution.

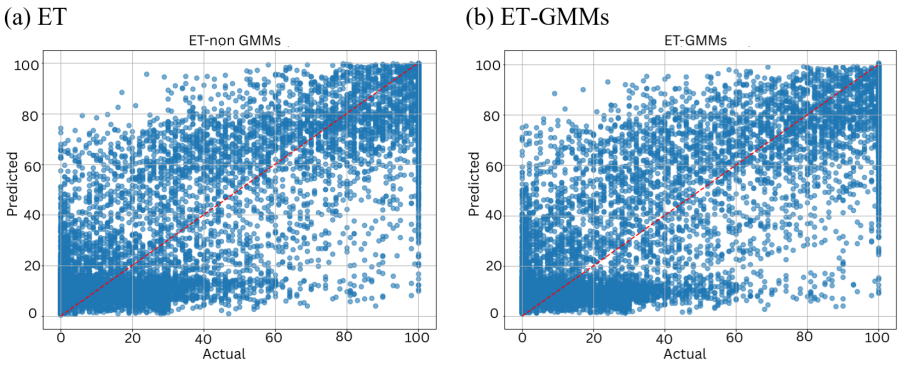

Figure 13: *Performance comparison of five ML models, including: (a) ET, and (b) ET-GMMs under actual value and predicted value*

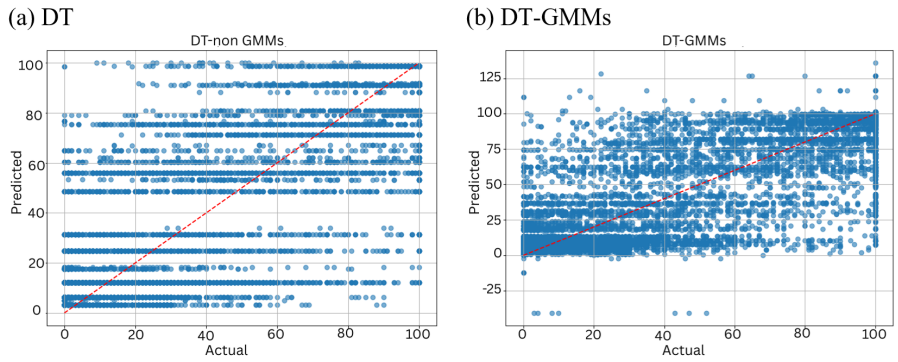

Figure 14: *Performance comparison of five ML models, including: (a) DT, and (b) DT-GMMs under actual value (X-axis) and predicted value (Y-axis)*

4.1. SHAP

The SHAP-based model evaluation enhances our understanding of the model's decision-making process. This interpretability allows us to identify the most influential parameters, plan for feature selection, and develop effective data collection strategies. It also contributes to model transparency, which is essential for building trust in ML systems, as illustrated in Figure 15-19.

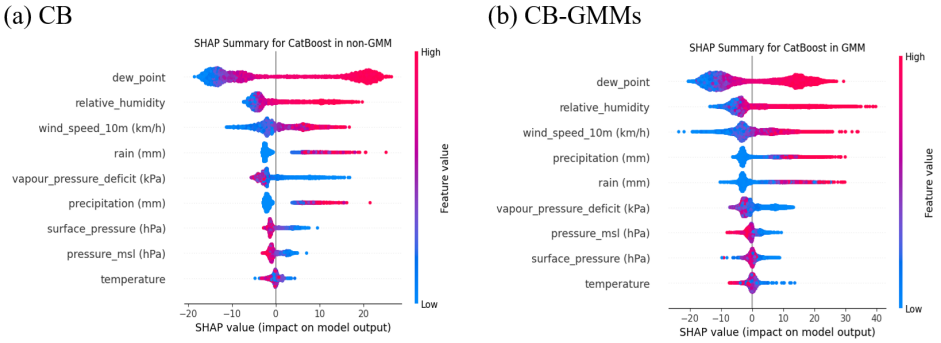

Figure 15: *SHAP summary plots for all models under comparison: (a) CB, and (b) CB-GMMs*

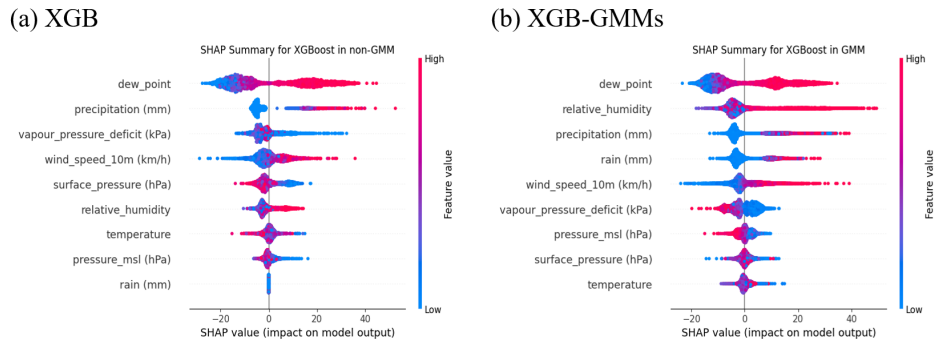

Figure 16: *SHAP summary plots for all models under comparison: (c) XGB, and (d) XGB-GMMs*

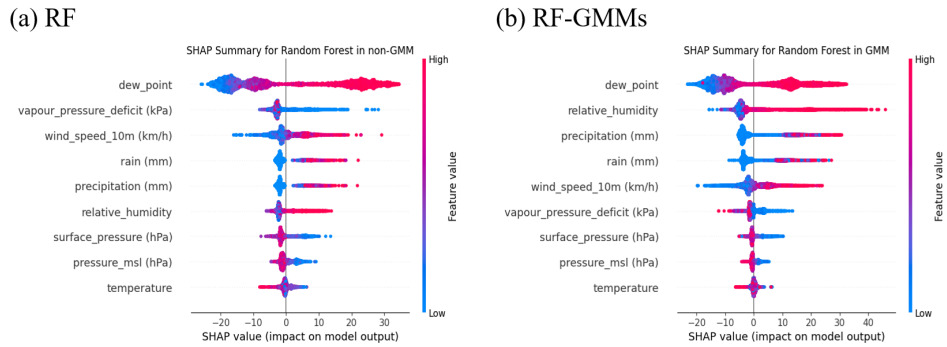

Figure 17: *SHAP summary plots for all models under comparison: (a) RF, and (b) RF-GMMs*

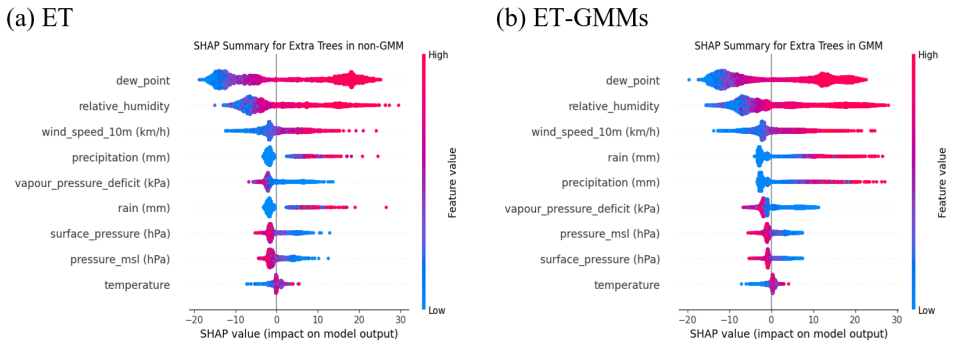

Figure 18: *SHAP summary plots for all models under comparison: (a) ET, and (b) ET-GMMs*

The SHAP analysis was systematically implemented to enhance model transparency and to validate the predictive mechanism against established atmospheric physics principles [41–43]. The resulting feature attributions confirmed that the ML models prioritize meteorological variables directly linked to condensation, saturation, and atmospheric dynamics. This physical grounding provides essential validation for the model's interpretability. Across all datasets, the most influential feature was Dew Point (Td), as shown in Figure 15-19. Its prominence is physically

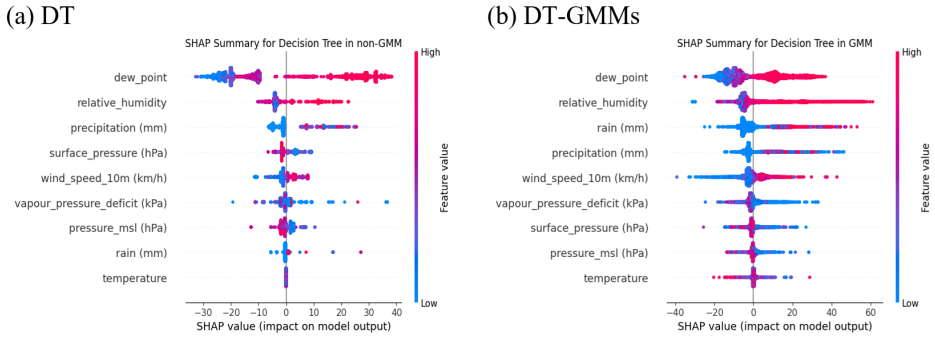

Figure 19: *SHAP summary plots for all models under comparison: (a) DT, and (b) DT-GMMs*

justified because Td directly reflects the temperature at which saturation occurs, derived from the same thermodynamic formulation used:

$$T_d = \frac{b \cdot \gamma(T, RH)}{a - \gamma(T, RH)} \tag{1}$$

This variable represents an absolute moisture indicator, and the closer Td is to T, the greater the likelihood of condensation and cloud formation [42]. In datasets without GMMs-based augmentation, Td exhibited a wide numerical range (–30 to 40), and models responded accordingly. The DT, in particular, relied most heavily on Td, showing a broad and discontinuous SHAP value distribution. These discrete SHAP patterns are consistent with the piecewise-constant approximations of DT (Figure 14), resulting in sharply defined decision boundaries. In contrast, CB displayed the narrowest SHAP spread for Td (–20 to 20), indicating a more balanced reliance on multiple variables and reflecting its high predictive accuracy.

Relative Humidity (RH) emerged as the second most influential feature for several models, including ET, DT, and CB, especially on datasets without GMMs augmentation. Its contributions were predominantly positive, consistent with the physical principle that cloud fraction increases non-linearly as RH approaches the critical saturation threshold [41]. This diagnostic relationship is captured using the same cloud-fraction equation adopted earlier:

$$C = 1 - \sqrt{\frac{1 - \frac{RH}{100}}{1 - RH_{\text{crit}}}} \tag{2}$$

Temperature (T) also contributed meaningfully, although its physical role is indirect; it controls saturation vapor pressure via the Clausius–Clapeyron equation:

$$e_s(T) = e_0 \cdot \exp\left(\frac{L_v}{R_v}\left(\frac{1}{T_0} - \frac{1}{T}\right)\right) \tag{3}$$

The models correctly prioritized the saturation-derived variables Td and RH instead of T alone. Similarly, Vapor Pressure Deficit (VPD), derived coherently from the same thermodynamic framework, showed important patterns aligned with expectations: low VPD corresponds to near-saturation and high cloud likelihood, while high VPD suppresses cloud formation. Other meteorological variables, including pressure and precipitation, also contributed significantly across several models. Their SHAP signatures reflect their known physical roles: pressure regulates large-scale convergence and uplift, while precipitation indicates ongoing microphysical processes within deep cloud fields.

Comparing model behavior with and without GMMs augmentation reveals notable changes, particularly in the DT model. Before augmentation, several features displayed highly discrete, point-wise SHAP patterns, consistent with DT's fragmented decision regions. After augmentation, these SHAP distributions became more continuous, indicating improved generalization and smoother decision boundaries due to richer training samples. Other models also adapted meaningfully as data volume increased. In XGB, RH became the second most influential feature after augmentation and even surpassed Td in SHAP value spread. In CB, Td initially exhibited a narrow SHAP range, but after GMMs-based augmentation, its influence expanded to 40, demonstrating the model's flexibility and enhanced sensitivity to moisture-related dynamics when provided with more diverse samples.

4.2. Statistical test (T-Test)

To evaluate the impact of GMMs-based data augmentation on model performance, we formulated the following hypotheses regarding the MAE. The null hypothesis (H_0) states that the use of GMMs augmentation has no effect on the MAE of the models, implying that the average MAE of models trained with GMMs-augmented data is equal to that of models trained without it. In contrast, the alternative hypothesis (H_1) asserts that GMMs augmentation does affect model performance, meaning that the average MAE differs between the two conditions. To test these hypotheses, a two-tailed t-test was conducted for each model to determine whether the observed differences in MAE were statistically significant. The results, as shown in Table 3.

Table 3: *Statistical significance of performance MAE differences between GMMs and Non-GMMs models*

Model	t-statistic	p-value	Significance
CB	−6.4498	0.00297	Significant
XGB	−8.0901	0.00127	Significant
RF	−7.0911	0.00209	Significant
ET	+10.9777	0.00039	Significant
DT	−3.9880	0.01629	Significant

According to Table 3, the results show that CB (t = –6.4498, p = 0.00297), XGB (t = –8.0901, p = 0.00127), RF (t = –7.0911, p = 0.00209), and DT (t = –3.9880, p = 0.01629) all exhibit significant negative t-statistics, indicating that these models achieve higher performance (lower MAE) when trained with GMMs-augmented data. In contrast, ET (t = 10.9777, p = 0.00039) shows a significant positive t-statistic, meaning that its MAE increases with GMMs augmentation rather than decreases. These findings indicate that, in most cases, the null hypothesis H_0 can be rejected, confirming that GMMs augmentation significantly affects MAE. However, the direction and magnitude of the effect vary across different model architectures, suggesting that while GMMs augmentation generally enhances model performance, certain tree-based ensembles with highly randomized splits may not benefit from synthetic data in the same way. Overall, the results support H_1 for most models, demonstrating that GMMs augmentation is broadly effective but not universally advantageous across all algorithms.

4.3. Boxplot

In addition to metric-based evaluation, the use of boxplots further aids in understanding the stability of the models, as illustrated in Figure 20.

The boxplot in Figure 20 clearly illustrates how GMMs-based augmentation influences MAE across the five evaluated models. In general, models such as Random Forest, Extra Trees, and

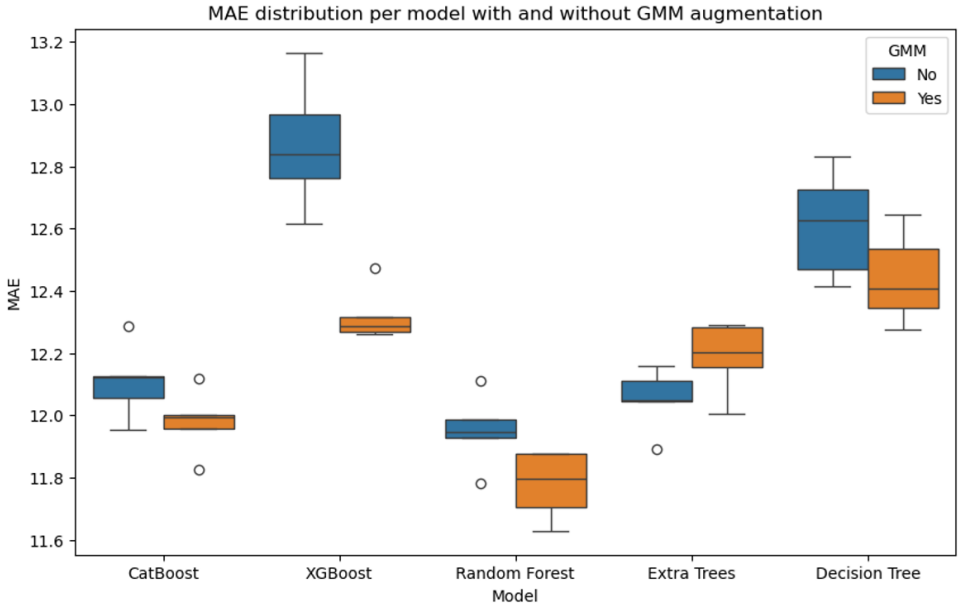

Figure 20: *Impact of GMMs-augmented per fold testing data on models MAE distributions*

Decision Tree show noticeably lower and more stable MAE values when trained with GMMs-augmented data, suggesting improved robustness and reduced variance. For CatBoost and XGBoost, the effect is more modest, but GMMs augmentation still provides slightly tighter distributions, indicating more consistent performance. Overall, the comparison highlights that GMMs augmentation tends to benefit tree-based models by reducing error variability and enhancing predictive performance.

5. Discussion

This study evaluates the effect of GMMs-based synthetic data augmentation on five regression models (CB, XGB, RF, ET, and DT) for weather-related prediction tasks, using MAE, RMSE, and R^2 metrics. Overall, GMMs augmentation led to modest but consistent improvements for most ensemble models. RF achieved the best performance with the lowest MAE (11.7767 ± 0.1091) and high R^2 (0.8092 ± 0.0043), reflecting both accurate predictions and stable variance capture. CB also benefited, showing decreased MAE (-0.1305 ± 0.1603) and improved R^2 with GMMs, indicating balanced reliance on predictive features.

The effect of augmentation was model-dependent. XGB showed notable improvement (MAE delta -0.5495 ± 0.2268), while ET exhibited marginal performance degradation (MAE +0.1357 ± 0.1540, RMSE +0.1975 ± 0.3290), suggesting sensitivity to misaligned synthetic samples. DT consistently underperformed (MAE 12.4412 ± 0.1486, RMSE 19.1079 ± 0.2504, R^2 0.7665 ± 0.0050), reflecting structural limitations as a single-tree model with piecewise constant predictions. These results indicate that simpler or highly randomized models may not fully benefit from synthetic augmentation and require careful calibration.

From an interpretability perspective, SHAP analysis showed dew_point as the most influential feature across all models. Ensemble methods like RF and CB displayed smooth, continuous SHAP distributions, indicating robust and interpretable decision boundaries. DT exhibited

discrete, stepwise SHAP patterns before augmentation; these became more continuous after GMMs integration, suggesting improved generalization even in simpler models.

5.1. Edge AI

We tested deploying the RF model on an ESP32 board, as it demonstrated the most stable performance in our experiments. Although this model has relatively low complexity, it still required parameter tuning to be feasible on the ESP32, specifically by reducing the number of estimators to 100 and the maximum tree depth to 5. The results of this deployment are illustrated in Figure 21.

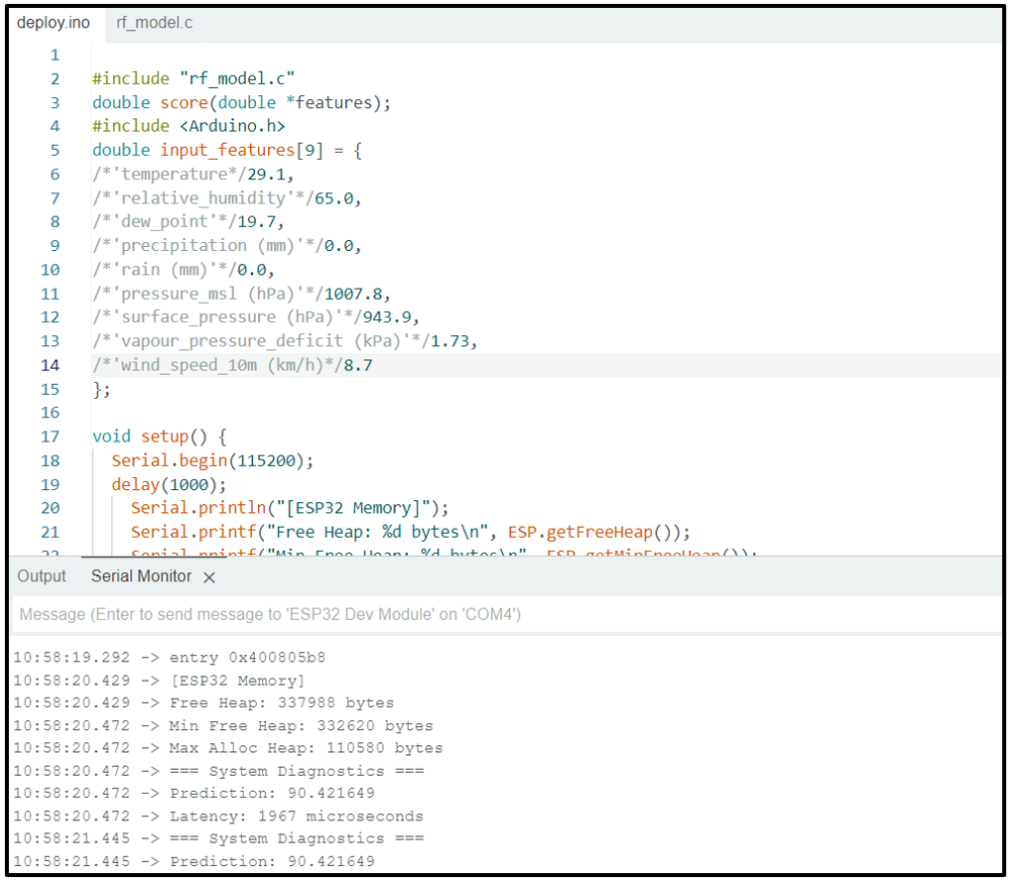

Figure 21: *System diagnostics and inference latency of RF model deployed on ESP32 edge device*

According to Figure 21, the deployment of the RF model on the ESP32 microcontroller demonstrates its suitability for edge computing applications. The ESP32 offers a balanced combination of computational capability and memory resources, with approximately 338 KB of free heap memory available and a minimum free heap consistently remaining above 332 KB during operation. The maximum heap allocation, recorded at around 110 KB, indicates efficient memory management under the device's resource constraints.

RF is particularly advantageous for edge deployment due to its relatively low computational complexity compared to deep learning models, while still maintaining robust performance for

regression tasks. The model's tree-based structure allows efficient inference through simple decision rules, which can be executed with minimal processing power and low latency. Indeed, the observed prediction latency of approximately 1.97 milliseconds per inference confirms that the ESP32 can perform timely predictions suitable for real-time applications. This combination of a manageable memory footprint, low latency, and computational efficiency makes RF a practical choice for edge AI, enabling on-device analytics and reducing the dependency on cloud-based processing. Such edge deployment facilitates faster decision-making, improved data privacy, and reduced network communication overhead.

5.2. Limitations

Despite the promising results, this study has several limitations. The effectiveness of GMMs-based augmentation depends heavily on the distribution and quality of the original dataset; in cases of low variance, high noise, or highly dynamic meteorological conditions, synthetic data may fail to capture rare or extreme patterns, potentially reinforcing biases and reducing predictive accuracy. Simpler models, such as Decision Trees, showed limited improvements, suggesting that augmentation strategies should be tailored to model complexity. Additionally, the evaluation framework, based on standard regression metrics like MAE, RMSE, and R^2, does not fully capture temporal dependencies or extreme weather events, and the interpretability analysis using SHAP may not generalize well to different geographic regions or meteorological conditions.

5.3. Practical integration

The outcomes of this study hold significant practical implications for integrating ML in solar energy prediction and PV system optimization. Accurate prediction of solar power generation relies heavily on precise weather information, particularly when meteorological datasets are incomplete or sparse; a common challenge in many regions of Southeast and South Asia, where sensor networks remain underdeveloped, especially in rural or remote areas. In such contexts, the demonstrated effectiveness of GMMs-based synthetic data augmentation, particularly in enhancing ensemble model performance, offers a robust approach to mitigating data scarcity and variability. By incorporating synthetic samples into the training pipeline, energy providers and grid operators can improve the reliability of solar power predictions without relying solely on extensive ground-based measurements. The enhanced interpretability provided by SHAP analysis further supports transparency and trust in predictions, which is critical for real-time energy dispatch, smart grid integration, and renewable energy planning in developing regions.

6. Conclusion

This study investigated the impact of GMMs-based synthetic data augmentation on the performance and interpretability of five ML models (CB, XGB, RF, ET, and DT) for weather-related regression tasks. Our experiments demonstrated that synthetic data can meaningfully enhance model performance, particularly for ensemble models. The Random Forest trained with GMMs-augmented data achieved the lowest MAE and maintained strong generalization, emerging as the most effective model. SHAP analyses further confirmed that GMMs augmentation improved predictive accuracy and contributed to more stable and meaningful feature attribution, especially in complex models.

These findings underscore the practical relevance of synthetic data augmentation in real-world energy prediction, particularly for solar power forecasting, where environmental variability, most

notably cloud cover, critically affects PV output. GMMs-generated synthetic samples can enrich sparse or inconsistent datasets, improving model reliability and supporting grid integration, energy dispatch, and renewable energy planning. This work can contribute to overcoming barriers to PV deployment by enhancing forecasting accuracy, which is particularly crucial given that high upfront costs, long payback periods, and technical uncertainties remain primary barriers to solar PV adoption across residential, commercial, and institutional sectors in urban contexts [44, 45]. Improved prediction models can help reduce these barriers by demonstrating clearer economic benefits, reducing investment risks, and optimizing system performance to support evidence-based policy decisions. Importantly, the benefits of augmentation are model-dependent, highlighting the need to tailor strategies to specific model architectures.

Looking forward, future research could extend validation across diverse climates and seasons, explore alternative generative approaches such as VAEs or GANs, integrate physical constraints or domain-specific rules, and incorporate temporal modeling techniques like RNNs or sequence-aware architectures. Such directions promise to enhance the robustness, generalizability, and practical applicability of solar energy prediction systems in data-constrained environments.

Declaration of interest: The authors declare no conflicts of interest.

References

[1] Chaudhry A, Chandni. The impact of innovation technology on carbon emissions in India. *Journal of Asian Energy Studies* 2023:7:1-19.

[2] Budhi GS, Tanoto Y, Jovian D, Adipranata R, Raphael C. Solar photovoltaic power output prediction using machine learning-based regressors. *Journal of Asian Energy Studies* 2025:9:111–130.

[3] Anantwar H, Sunda S. Intelligent optimized voltage control for hybrid off-grid power systems. *Journal of Asian Energy Studies* 2023:7:39-47.

[4] Obuseh E, Eyenubo J, Alele J, Okpare A, Oghogho I. A systematic review of barriers to renewable energy integration and adoption. *Journal of Asian Energy Studies* 2025:9:26-45.

[5] Liza ZA, Aktar H, Islam MR. Solar energy development and social sustainability: a case study on the Teknaf solar power plant in Bangladesh. *Journal of Asian Energy Studies* 2020:4(1):1-8.

[6] Mukherjee M, Khushi M. SMOTE-ENC: a novel SMOTE-based method to generate synthetic data for nominal and continuous features. *Applied System Innovation* 2021:4(1):18.

[7] Afrifa-Yamoah E, Mueller UA, Taylor SM, Fisher AJ. Missing data imputation of high-resolution temporal climate time series data. *Meteorological Applications* 2020:27:e1873.

[8] Mumuni A, Mumuni F, Gerrar NK. A survey of synthetic data augmentation methods in machine vision. *Machine Intelligence Research* 2024:21:831-869.

[9] Jouan G, Cuzol A, Monbet V, Monnier G. Gaussian mixture models for clustering and calibration of ensemble weather forecasts. *Discrete & Continuous Dynamical Systems-Series S* 2023:16(2):309-328.

[10] Villefranque N, Hogan RJ. Evidence for the 3D radiative effects of boundary-layer clouds from observations of direct and diffuse surface solar fluxes. *Geophysical Research Letters* 2021:48(14):e2021GL093369.

[11] Gardner AS, Maclean IM, Rodríguez-Muñoz R, Ojanguren AF, Tregenza T. How air temperature and solar radiation impact life history traits in a wild insect. *Ecology and Evolution* 2025:15(3):e71135.

[12] Jadhav AV, Rahul PR, Kumar V, Dumka UC, Bhawar RL. Spatiotemporal assessment of surface solar dimming in India: Impacts of multi-level clouds and atmospheric aerosols. *Climate* 2024:12(4):48.

[13] Bando T, Ito T, Wakisaka H, Miyahara Y, Aizawa T, Harigai T, et al. Statistical analysis of cloud layers and solar irradiations for all seasons in Toyohashi City, Japan. *Renewable Energy and Environmental Sustainability* 2023:8:18.

[14] Hofsteenge MG, Cullen NJ, Conway JP, Reijmer CH, Van Den Broeke MR, Katurji M. Meteorological drivers of melt at two nearby glaciers in the McMurdo Dry Valleys of Antarctica. *Journal of Glaciology* 2024:70:e48.

[15] Kim B, Wan J, Chang K. Twenty-four-hour cloud cover calculation using a ground-based imager with machine learning. *Atmospheric Measurement Techniques* 2021:14(10):6695-6710.

[16] Park S, Kim Y, Ferrier N, Collis S, Sankaran R, Beckman P. Prediction of solar irradiance and photovoltaic solar energy product based on cloud coverage estimation using machine learning methods. *Atmosphere* 2021:12(3):395.

[17] Deo R, Grant R, Webb A, Ghimire S, Igoe D, Downs N, et al. Forecasting solar photosynthetic photon flux density under cloud cover effects: novel predictive model using convolutional neural network integrated with long short-term memory network. *Stochastic Environmental Research and Risk Assessment* 2022:36(10):3183-3220.

[18] Alblooshi MA, Kalathingal SH, Mirza SB, Ridouane FL. Assessment and classification of cloud coverage using K-Means clustering algorithm for the Sentinel-3 LST data: A case study in the Fujairah region. *American Journal of Remote Sensing* 2023 Jul;11(2):32-35.

[19] Dissawa LH, Godaliyadda RI, Ekanayake PB, Agalgaonkar AP, Robinson D, Ekanayake JB, Perera S. Sky Image-Based Localized, Short-Term Solar Irradiance Forecasting for Multiple PV Sites via Cloud Motion Tracking. *International Journal of Photoenergy* 2021:2021(1):9973010.

[20] Harty T, Holmgren W, Lorenzo A, Morzfeld M. Intra-hour cloud index forecasting with data assimilation. *Solar Energy* 2019:185:270-282.

[21] Haputhanthri D, De Silva D, Sierla S, Alahakoon D, Nawaratne R, Jennings A, Vyatkin V. Solar irradiance nowcasting for virtual power plants using multimodal long short-term memory networks. *Frontiers in Energy Research* 2021:9:722212.

[22] Balal A, Jafarabadi Y, Demir A, Igene M, Giesselmann M, Bayne S. Forecasting solar power generation utilizing machine learning models in Lubbock. *Emerging Science Journal* 2023:7(4):1052-1062.

[23] Cha J, Kim M, Lee S, Kim K. Investigation of applicability of impact factors to estimate solar irradiance: comparative analysis using machine learning algorithms. *Applied Sciences* 2021:11(18):8533.

[24] Fricke T, Keay L, Resnikoff S, Tahhan N, Mekountchou I, Paudel P, et al. Improving population-level refractive error monitoring via mixture distributions. *Ophthalmic and Physiological Optics* 2023:43(3):445-453.

[25] López-Lobato AL, Avendaño-Garrido ML. Fitting a Gaussian mixture model through the Gini index. *International Journal of Applied Mathematics and Computer Science* 2021:31(3):487-500.

[26] Mandal N, Sarode T. A framework for cloud cover prediction using machine learning with data imputation. *International Journal of Electrical and Computer Engineering* 2024:14(1):600-607.

[27] Lee J, Lee C. Raining state study using Gaussian mixture model. *International Journal of Advanced Smart Convergence* 2020:2(3):21-25.

[28] Akodad S, Bombrun L, Germain C, Berthoumieu Y. A Gaussian mixture model with multiple tangent planes. European Signal Processing Conference (EUSIPCO) 2023:950-954.

[29] Grundner A, Beucler T, Gentine P, Eyring V. Data-driven equation discovery of a cloud cover parameterization. *Journal of Advances in Modeling Earth Systems* 2024:16(3):e2023MS003763.

[30] Baran Á. Machine learning for total cloud cover prediction. *Neural Computing and Applications* 2021:33:2605-2620.

[31] Kim B, Wan J, Lee Y. Estimation of twenty-four-hour continuous cloud cover using ground-based imager with convolutional neural network. *Atmospheric Measurement Techniques* 2023:16(21):5403-5413.

[32] Grundner A, Beucler T, Gentine P, Iglesias-Suarez F, Giorgetta MA, Eyring V. Deep learning based cloud cover parameterization for ICON. *Journal of Advances in Modeling Earth Systems* 2022:14(12):e2021MS002959.

[33] Wu L, Chen T, Nima C, Wang D, Meng H, Li M, et al. Development of a machine learning forecast model for global horizontal irradiation adapted to Tibet based on visible all-sky imaging. *Remote Sensing* 2023:15(9):2340.

[34] Tscholl S, Tasser E, Tappeiner U, Vigl L. Coupling solar radiation and cloud cover data for enhanced temperature predictions over topographically complex mountain terrain. *International Journal of Climatology* 2021:42(9):4684-4699.

[35] Raj S, Deo R, Sharma E, Prasad R, Dinh T, Salcedo-Sanz S. Atmospheric visibility and cloud ceiling predictions with hybrid IIS-LSTM integrated model: case studies for Fiji's aviation industry. *IEEE Access* 2024:12:72530-72543.

[36] Morcrette C, Cave T, Reid H, da Silva Rodrigues J, Deveney T, Kreusser L, Van Weverberg K, Budd C. Scale-aware parameterization of cloud fraction and condensate for a global atmospheric model machine-learned from coarse-grained kilometer-scale simulations. *Journal of Advances in Modeling Earth Systems* 2025:17(4):e2024MS004651.

[37] Jang I, Kim H, Lee D, Son Y, Kim S. Knowledge transfer for on-device deep reinforcement learning in resource constrained edge computing systems. *IEEE Access* 2020:8:146588-146597.

[38] Li Y, Hong Y. Prediction of football match results based on edge computing and machine learning technology. *International Journal of Mobile Computing and Multimedia Communications* 2022:13(2):1-10.

[39] Dande P. Time series weather dataset. Kaggle 2025. https://www.kaggle.com/datasets/parthdande/timeseries-weather-dataset

[40] Cheung MCJ, Yip WL, Li CY, Xu Z. Integrated self-sustained renewable-energy explorer (iSEE). *Journal of Asian Energy Studies* 2023:7:132-139.

[41] Lu YS, Franke J, Viñas B, Berndt C, Dippold A, Bittihn JC, Ludwig S. Optimization of weather forecasting for cloud cover over the European domain with the ESIAS-Weather system. *Geoscientific Model Development* 2023:16(3):1083-1108.

[42] Rout A, Mishra P, Mohanty UC. Cloud-integrated meteorological parameter prediction by interpretable deep learning. *IEEE Access* 2024;12:22538-22548.

[43] Rudrappa G, Ramya A, Guruprasad HS. Cloud classification and cloud cover estimation using a hybrid deep Kronecker network. *Engineering Applications of Artificial Intelligence* 2025:130:107352.

[44] Mah DN, Wang G, Lo K, Leung MK, Hills P, Lo AY. Barriers and policy enablers for solar photovoltaics (PV) in cities: Perspectives of potential adopters in Hong Kong. *Renewable and Sustainable Energy Reviews* 2018:92:921-936.

[45] Lo K, Mah DN, Wang G, Leung MK, Lo AY, Hills P. Barriers to adopting solar photovoltaic systems in Hong Kong. *Energy & Environment* 2018:29(5):649-663.