Article

Spatiotemporal Downscaling Model for Solar Irradiance Forecast Using Nearest-Neighbor Random Forest and Gaussian Process

Shadrack T. Asiedu , Abhilasha Suvedi , Zongjie Wang , Hossein Moradi Rekabdarkolaee , Timothy M. Hansen *

Posted Date: 11 March 2025

doi: 10.20944/preprints202503.0629.v1

Keywords: distributed energy resources; FERC 2222; nearest-neighbor Gaussian process; nearestneighbor random forest; solar irradiance forecasting; spatiotemporal downscaling

doi:10.20944/preprints202503.0629.v1

Article

Spatiotemporal Downscaling Model for Solar Irradiance Forecast Using Nearest-Neighbor Random Forest and Gaussian Process

Shadrack T. Asiedu [1,†], Abhilasha Suvedi [1,†], Zongjie Wang [2], Hossein Moradi Rekabdarkolaee [3] and Timothy M. Hansen [1,*]

[1] McComish Department of Electrical Engineering and Computer Science, South Dakota State University, Brookings, SD 57007
[2] Department of Electrical and Computer Engineering, University of Connecticut, Storrs, CT 06268
[3] Department of Mathematics and Statistics, South Dakota State University, Brookings, SD 57007
[*] Correspondence: timothy.hansen@sdstate.edu
[†] These authors contributed equally to this work.

Abstract: Accurate solar photovoltaic (PV) capacity estimation requires high-resolution, site-specific solar irradiance data to account for localized variability. However, global datasets, such as the National Solar Radiation Database (NSRDB), provide regional averages that fail to capture the fine-scale fluctuations critical for large-scale grid integration. This limitation is particularly relevant in the context of increasing distributed energy resources (DERs) penetration, such as rooftop PV. Additionally, it is critical to the implementation of Federal Energy Regulatory Commission (FERC) Order 2222, which facilitates DER participation in bulk power markets. To address this challenge, this study evaluates Nearest-Neighbor Random Forest (NNRF) and Nearest-Neighbor Gaussian Process (NNGP) models for spatiotemporal downscaling of global solar irradiance data. By leveraging historical irradiance and meteorological data, these models incorporate spatial, temporal, and feature-based correlations to enhance local irradiance predictions. The NNRF model, a machine learning approach, prioritizes computational efficiency and predictive accuracy, while the NNGP model offers a level of interpretability and prediction uncertainty by numerically quantifying correlations and dependencies in the data. Model validation was conducted using day-ahead predictions. The results showed that the average goodness of fit (GoF) of the NNRF model of 90.61% across all eights sites outperformed the GoF of the NNGP of 85.88%. Additionally, the computational speed of NNRF was 2.5 times faster than the NNGP. Finally, the NNGP displayed polynomial scaling while the NNRF scaled linearly with increasing number of nearest neighbors. These findings highlight the robustness and computational efficiency of NNRF for large-scale solar irradiance downscaling, making it a strong candidate for improving PV capacity estimation and real-time electricity market integration.

Keywords: distributed energy resources; FERC 2222; nearest-neighbor Gaussian process; nearest-neighbor random forest; solar irradiance forecasting; spatiotemporal downscaling

1. Introduction

The ongoing global de-carbonization effort has informed regulatory policies such as the U.S. Federal Energy Regulatory Commission (FERC) Order 2222 [1]. This has given way for increasing integration of renewable energy resources which is reshaping the current power system and electricity market operations [2]. This order facilitates the participation of distributed energy resources (DERs), including rooftop solar photovoltaic (PV) systems, in wholesale power markets, necessitating more accurate and localized solar irradiance forecasts. However, integrating variable renewable energy resources (VREs) into the grid introduces new challenges related to stability, reliability, and dispatch [3]. Unlike conventional power plants, VREs such as solar PV are inherently uncertain with volatile output due to rapid fluctuations in cloud cover, humidity, and other weather conditions. Accurate and high-resolution solar irradiance forecasting are essential for power system operations of DERs. However, global solar irradiance models are typically available at coarse spatial and temporal resolutions.

Advances in power electronic converter (PEC) technology, such as improved dynamic models [4] and virtual inertia solutions [5], address some stability concerns. However, accurate solar PV forecasting remains a critical challenge for grid operators and market participants. Existing approaches for solar PV forecasting primarily rely on numerical weather prediction (NWP) models [6,7], statistical methods and machine learning techniques [8–11], deep learning algorithms [12,13], and hybrid approaches [9,14,15]. Some studies incorporate probabilistic methods to quantify uncertainty [16–18], while others consider the spatial influence of neighboring sites on a target location [12,19]. Although these methods have demonstrated success, many rely on an implicit assumption that global solar irradiance datasets provide accurate site-specific representations. This assumption introduces inaccuracies since global datasets, such as those from the National Solar Radiation Database (NSRDB) [20], provide irradiance values averaged over large spatial scales (e.g., 4 km^2), which fail to capture fine-grained local variability.

A key challenge in solar PV forecasting is the need for high-resolution, site-specific solar irradiance data. Current datasets from sources like NSRDB and Open Meteo [21] cover broad spatial regions, with some reaching resolutions as coarse as 11 km × 11 km. This leads to approximation errors that can undermine accurate capacity estimation and power system operations such as dispatch and balancing. While the ideal solution would involve deploying weather stations at every prospective PV site to obtain direct irradiance measurements, this approach is impractical due to cost, maintenance, and logistical constraints. Furthermore, it would be infeasible for forecasting irradiance at prospective solar PV sites where physical measurements are not yet available. Therefore, there is a pressing need for advanced downscaling techniques that can transform coarse global irradiance data into high-resolution, site-specific estimates suitable for real-time market operations and DER integration.

By leveraging historical data on both global and local solar irradiance measurements, models can be developed to learn the mapping of the global values to the local values. Additionally, by incorporating spatial correlations, the dependencies of neighboring sites could be captured to infer forecast values for sites whose local solar irradiance measurements are unknown. As promising as downscaling is for power system operations, there is a dearth of literature on this method. Most solar irradiance downscaling studies focused on generating high resolution time scale measurements from coarse-grained solar irradiance forecast values [22–26]. Some of the studies that focused on spatial downscaling include [27], where coarse resolution downward shortwave radiation is disaggregated into a 30 meters scale using atmospheric transmittance-based weighting technique. Although the proposed model achieves reliable downscaling results, especially for mountainous areas, the model heavily depends on a satellite derived dataset. Similarly, artificial neural networks (ANNs) were used to downscale weather variables from a 1.2 km spatial resolution to a 240 m resolution [28]. The study, however, used a simulated dataset generated from the Weather Research and Forecasting model in Large Eddy Simulation mode with 240 grid, which may bring to question the practical implementation of their model. The study in [29] adopted the nearest-neighbor Gaussian process (NNGP) to downscale solar irradiance from global resolution to a more fine-grained local resolution. However, the model in [29] rather performed *interpolation* instead of extrapolation (i.e., forecasting) to obtain day-ahead predictions. By evaluating the NNGP model on a temporal point within the training temporal space, the study failed to validate the capability of the NNGP to accurately forecast downscaled irradiance for timepoints beyond the training temporal space.

The main contribution of this paper is the development of an accurate scalable spatiotemporal downscaling model for day-ahead solar irradiance forecast. Our study designs a novel approach to spatiotemporal downscaling of solar irradiance for modeling and forecasting using a nearest-neighbor random forest (NNRF). The NNRF approach is compared to the the performance of NNGP updated from [29] to properly perform forecasting.

doi:10.20944/preprints202503.0629.v1

In the rest of the paper, the methodology is presented in Section 2, which describes the NNGP and NNRF spatiotemporal models. The simulation setup and data collection are presented in Section 3. Section 4 of this paper discusses the results. Finally, conclusions are made in Section 5, alongside future research directions. Additionally, the data pre-processing steps are shown in the appendix.

2. Downscaling Theory and Methods

2.1. Spatiotemporal Models

Time series data, used in many forecasts, contain information about *when* each data element was observed. Spatial data on the other hand reveals information about *where* each data element was collected [30]. Spatiotemporal forecast models, combining these two dependencies, predict target variables by analyzing the dataset in both space and time dimensions [31]. The combination of temporal and spatial features enables the model to capture both the influence of nearby locations and past observations on the prediction outcome. Such integration of spatial components into temporal forecasts is demonstrated in literature to improve forecast accuracy [32]. Mathematically, a spatiotemporal model for predicting a variable y can be expressed as:

$$\mathbf{y_{i,t}} = f(\mathbf{y}_{i,t-1}, \ldots, \mathbf{y}_{i,t-n}, \mathbf{X}_{i,t}, \mathbf{Y}_{\text{neighbors}}, \boldsymbol{\epsilon_{i,t}}), \tag{1}$$

where, at time t and location i, $\boldsymbol{y}_{i,t} \in R^n$ is the target variable, $y_{i,t-1}, \ldots, y_{i,t-n}$ represents n past values (i.e., temporal dependency), $\mathbf{X}_{i,t}$ are exogenous variables (e.g., weather features in the case of solar irradiance forecast), $\mathbf{Y}_{\text{neighbors}}$ are target variable values from neighboring locations (i.e., spatial dependency), and $\boldsymbol{\epsilon_{i,t}} \in R^n$ is the model error term. The spatiotemporal model is then represented as the function f (e.g., statistical, machine learning, deep learning).

2.2. Methods for Downscaling Solar Irradiance

Global solar irradiation data, like those provided by the NSRDB, cover a 4 km by 4 km spatial resolution and mostly in a coarse temporal resolution [33]. The low resolution of the NSRDB poses two significant challenges to the accurate estimation of solar PV output. First, the 30 minute time resolution is not suitable for real-time dispatch of solar PV, where output forecast are needed in a shorter time scale (e.g., 5 minutes for real-time electricity markets). Secondly, due to the erratic variability of weather, the averaged solar irradiance over the 4 km by 4 km space does not reflect the individual site specific measurements, which can contain many different solar PV installations in the 16 km^2 area. Figure 1 shows that the inherent variation between the global solar irradiance and the ground measurements can range from very small negligible deviations to significantly high disparities in a day. Such significant margins could misinform the forecast of solar PV output and lead to inefficient market dispatch or, in the worst case, grid failure and blackouts. These spatiotemporal resolution issues necessitate the development of accurate models to map global irradiance to site specific solar irradiance for better PV capacity estimation and electricity market integration.

doi:10.20944/preprints202503.0629.v1

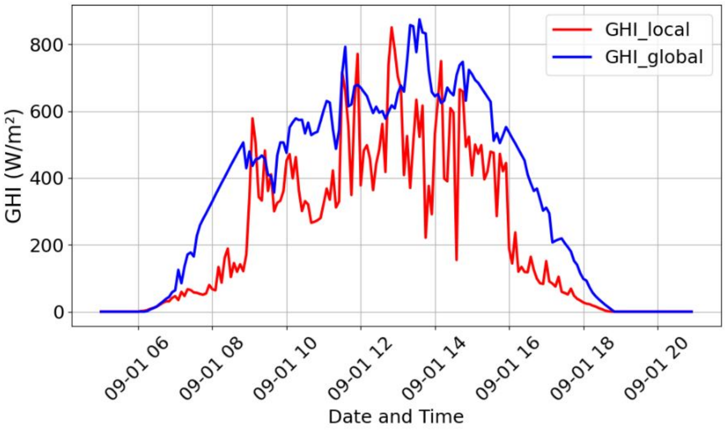


Figure 1. Daily Global and Local Solar Irradiance

To better estimate the ground measurements, the Gaussian copula is used to downscale global to local global horizontal irradiance (GHI) clear sky indices in [24]. The same method is used in [23] to spatiotemporally downscale solar irradiance from hourly resolution to a 15 second resolution. This method models dependencies between variables with potentially different marginal distributions by transforming them into a common Gaussian space using a copula function. The study in [24] is improved by adopting the T-copula method to downscale the GHI clear sky index in [25]. This improved method offered the advantage of better simulation of binary events caused by the movement of clouds of adjustable frequency. Further, a nearest-neighbor (NN) approach is combined with a Gaussian process (GP) in [29] to downscale solar irradiance from the global level to local resolution. The GP simplifies the Gaussian copula method by assuming the marginal distributions of the variables within the dataset to be uniform or Gaussian. Such assumption makes the GP model very suitable for sparse dataset. However, this process still involves the inversion of large covariance matrices, making its application computationally intensive. Based on the concept of the NNGP, our proposed NNRF resolves this challenge by replacing the GP with a random forest (RF) model, eliminating the need to invert a large covariance matrix. Instead, an ensemble of decision trees are used to capture the temporal and spatial relationship between neighboring sites and past observations. We will next describe the underlying theory and application of NNGP and NNRF to spatiotemporal downscaling.

2.3. Nearest Neighbor Gaussian Process (NNGP)

NNGP is an efficient approximation of the traditional GP, designed for large-scale datasets. Traditional GPs involve a covariance matrix of size $n \times n$, where n is the number of training points. This leads to computational complexities of $\mathcal{O}(n^3)$ for matrix inversion and $\mathcal{O}(n^2)$ for storage. NNGP mitigates this limitation by using only the k-nearest neighbors for each data point, resulting in a sparse precision matrix. This approximation reduces the computational complexity from $\mathcal{O}(n^3)$ to $\mathcal{O}(k^3 n)$, while preserving local dependencies. As k is a constant parameter, the scaling of NNGP becomes linear with n.

For a random process $f(x)$, a GP is defined as:

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')), \tag{2}$$

where $m(x)$ is the mean function, and $k(x, x')$ is the covariance kernel function that models the similarity between points x and x'. A common kernel is the squared exponential kernel [34,35], given as:

$$k(x, x') = \sigma^2 \exp\left(-\frac{\|x - x'\|^2}{2\ell^2}\right), \tag{3}$$

where σ^2 is the variance and ℓ is the length scale controlling smoothness. NNGP approximates the full GP by considering only the k nearest neighbors (k-NNs) for each data point. The covariance matrix is sparsified as:

$$K_{\text{NN}}(x_i, x_j) = \begin{cases} k(x_i, x_j), & \text{if } x_j \in \mathcal{N}_i, \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

where $\mathcal{N}_i$ denotes the set of k-NNs of x_i. This informs the computation of the covariance matrices which incorporates spatial, temporal, and feature-based distances. Thus:

$$k(x, x') = \sigma^2 \exp\left(-\frac{d_s^2}{2\ell_s^2}\right) \exp\left(-\frac{d_t^2}{2\ell_t^2}\right) \exp\left(-\frac{d_g^2}{2\ell_g^2}\right), \tag{5}$$

where

- $d_s = \|x_{\text{spatial}} - x'_{\text{spatial}}\|$ is the spatial distance,
- $d_t = |t - t'|$ is the temporal distance, and
- $d_g = \|x_{\text{global features}} - x'_{\text{global features}}\|$ is the feature-based distance.

Using the GP kernel in (5), we compute the covariance matrix between the test point and its neighbors ($K_{\text{test,train}}$), as well as the covariance matrix among the neighboring points themselves ($K_{\text{train,train}}$). The final downscaled prediction is obtained as:

$$\hat{\mathbf{y}} = \mathbf{X}_{\text{test}}^T \hat{\beta} + K_{\text{test, train}} K_{\text{train, train}}^{-1} \left(\mathbf{y}_{\text{train}} - \mathbf{X}_{\text{train}}^T \hat{\beta}\right), \tag{6}$$

where

$$\hat{\beta} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}. \tag{7}$$

In the above equation, X_{test} represents the input features of the target site, while X_{train} denotes the input features of the neighboring sites. The matrix $K_{\text{test, train}}$ is the covariance matrix between the test site and its neighbors, and $K_{\text{train, train}}$ is the the covariance matrix among the neighboring sites. The vector y_{train} contains the target values of the neighboring sites. More generally, X and y denote the input features and target values for all sites, respectively, and $\hat{\beta}$ is the estimated coefficient vector obtained from the linear regression model. Finally, $\hat{y}$ is the predicted downscaled solar irradiance of the target site. The pseudo algorithm for the implementation of this model is described in Algorithm 1.

Algorithm 1 Pseudo Algorithm for Nearest-Neighbor Gaussian Process (NNGP)

```
Require: • Training data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with spatial coordinates $\mathbf{s}_i$ and global features $\mathbf{f}_i$
    • Hyperparameters $\sigma$, $\ell_s$, $\ell_t$, $\ell_g$ (spatial, temporal, and feature length scales, and signal variance)
    • Regularization term $\epsilon$ and number of neighbors $k$
Ensure: Predictions $\hat{y}$ for test points $\mathbf{x}_*$
Step 1: Data Preprocessing
Construct input features $\mathbf{X}$ for training points:
    $\mathbf{X}_i \leftarrow \{t_i, \mathbf{s}_i, \mathbf{f}_i\} \quad \forall i \in \{1, 2, \ldots, n\}$
Step 2: Kernel Function
Define the combined kernel function:
    $K(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \exp\left(-\frac{d_s^2}{2\ell_s^2}\right) \exp\left(-\frac{d_t^2}{2\ell_t^2}\right) \exp\left(-\frac{d_g^2}{2\ell_g^2}\right),$
  where:
    • $d_s = \|\mathbf{s}_i - \mathbf{s}_j\|$ is the spatial distance
    • $d_t = |t_i - t_j|$ is the temporal distance
    • $d_g = \|\mathbf{f}_i - \mathbf{f}_j\|$ is the feature distance
Step 3: Hyperparameter Selection
Select hyperparameters that minimize the validation mean absolute error:
    $\hat{\ell}_s, \hat{\ell}_t, \hat{\ell}_g, \hat{\sigma} = \arg\min_{\ell_s, \ell_t, \ell_g, \sigma} \mathrm{MAE}_{\mathrm{val}}$
Step 4: Nearest Neighbor Selection
for each test point $\mathbf{x}_*$ do
    Compute spatial distances $d_s$, temporal distances $d_t$, and feature distances $d_g$ between $\mathbf{x}_*$ and all training points
    Identify k-NN based on combined distances
end for
Step 5: Covariance Computation
Compute the covariance matrix among neighbors for training points:
    $K_{\text{train, train}} = K(\mathbf{x}_i, \mathbf{x}_j) + \epsilon \mathbf{I},$
  where:
    • $\epsilon \mathbf{I}$ is a penalty term to prevent singularity of the covariance matrix
Compute the covariance vector between test point and neighbors:
    $K_{\text{test, train}} = K(\mathbf{x}_*, \mathbf{x}_j) \quad \forall j \in \text{neighbors}$
Step 6: Prediction
Estimate the coefficients of the linear regression model
    $\hat{\beta} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}$
Predict the output for test point:
    $\hat{\mathbf{y}} = \mathbf{X}_{\text{test}}^T \hat{\beta} + K_{\text{test, train}} K_{\text{train, train}}^{-1} \left(\mathbf{y}_{\text{train}} - \mathbf{X}_{\text{train}}^T \hat{\beta}\right)$
Step 7: Evaluation
Calculate metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Goodness-of-Fit (GoF) for predictions
Step 8: Output
Return trained NNGP predictions for each site and evaluation results
```

2.4. Nearest-Neighbor Random Forest (NNRF)

The RF algorithm is formed by bagging random samples of decision trees to produce an ensemble model that improves the predictions of each tree [36]. RFs are well known for their ability to learn complex non-linear patterns while efficiently handling high-dimensional features and large datasets [37,38]. This makes RF well-suited for scalability as the number of neighbors and data points increases.

An RF model minimizes the Mean Squared Error (MSE) of the ensemble decision trees over the training points such that:

$$\text{MSE} = \frac{1}{N} \sum_{i \in \mathcal{N}_s} \left(y_i - \frac{1}{T} \sum_{t=1}^{T} h_t(\mathbf{f}_i) \right)^2, \tag{8}$$

where T is the total number of trees in the forest and $h_t(\mathbf{f}_i)$ is the prediction from the t-th tree for input $\mathbf{f}_i$. Each decision tree, h_t is trained by recursively splitting the training data to minimize the *impurity*. A common type of impurity is the entropy, which quantifies the amount of uncertainty or information disorder at each splitting node. The RF model seeks to minimize the average entropy of the decision tree as:

$$\min H(S) = -\sum_i P_i(S) log_2 P_i(S), \tag{9}$$

where S represents the set of samples at a particular node in the decision tree, H is the entropy of S, and P_i measures the probability of class i in sample S.

For a given test point $\mathbf{X}_{\text{test}}$, the RF model predicts the normalized local irradiance as:

$$\hat{\mathbf{y}}_{\text{test}} = \frac{1}{T} \sum_{t=1}^{T} h_t(\mathbf{X}_{\text{test}}), \tag{10}$$

where $h_t(\mathbf{X}_{\text{test}})$ is the prediction from the t-th tree for test input $\mathbf{X}_{\text{test}}$.

The NNRF approach enhances traditional RF by incorporating an NN framework, ensuring both spatial and temporal locality in predictions. NNs are selected based on spatial coordinates, allowing the model to capture local dependencies more effectively while maintaining computational efficiency. For a target site s, the k-NNs are determined using the Euclidean distance [39]:

$$d_{\text{spatial}}(s, n) = \|\mathbf{x}_s - \mathbf{x}_n\|_2, \tag{11}$$

where $\mathbf{x}_s$ and $\mathbf{x}_n$ are the spatial coordinates of the target site s and the n-th neighbor, respectively. The k-NNs are stored in the set:

$$\mathcal{N}_s = \{n_1, n_2, \ldots, n_k\}.$$

For a target site s with spatial coordinates $\mathbf{x}_s$, the k-NNs are identified as:

$$\mathcal{N}_s = \arg\min_{\mathcal{S}'} \sum_{i \in \mathcal{S}'} \|\mathbf{x}_s - \mathbf{x}_i\|_2, \tag{12}$$

where $|\mathcal{S}'| = k$ and $\mathcal{S}$ is the set of all sites. For each site s, the training data consists of its feature vectors in addition to the global irradiance of the neighboring sites. The target variable for each site is the local irradiance of that particular test site. The pseudo algorithm for the implementation of NNRF is described in Algorithm 2.

Algorithm 2 Pseudo Algorithm for Nearest-Neighbor Random Forest (NNRF)

```
Require: Training data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, spatial coordinates $\mathbf{s}_i$, number of neighbors $k$, features $\mathcal{F}$,
    target variable $y$
Ensure: Trained random forest model for predicting $y$
Step 1: Preprocessing
Combine data from all sites into a single dataset $\mathcal{D}$
Normalize feature values using MinMaxScaler
Step 2: Nearest-Neighbor Identification
for each site $s$ in spatial coordinates $\mathbf{s}_i$ do
    Compute spatial distances $d_s$ between site $s$ and all other sites
    Identify k-NNs based on $d_s$
end for
Step 3: Training Data Preparation
for each site $s$ do
    Extract training data $\mathcal{D}_s$ for site $s$ from $\mathcal{D}$
    for each neighbor $s_{\text{neighbor}}$ of site $s$ do
        Extract neighbor's data $\mathcal{D}_{s_{\text{neighbor}}}$
        Align neighbor's data by datetime with $\mathcal{D}_s$
        Add neighbor's GHI as a new feature: $GHI_{\text{neighbor}}$
    end for
end for
Step 4: Random Forest Training
for each site $s$ do
    Define features $\mathcal{F}$, including $GHI_{\text{neighbor}}$ for each neighbor
    Split data into $\mathbf{X}_{\text{train}}$ (features) and $mathbfy_{\text{train}}$ (target variable $GHI\_local$)
    Train Random Forest Regressor on $\mathbf{X}_{\text{train}}$ and $\mathbf{y}_{\text{train}}$ with $n_{\text{estimators}}$ trees
end for
Step 5: Prediction
for each test point $\mathbf{x}_*$ at site $s$ do
    Align neighbor data by datetime and add $GHI_{\text{neighbor}}$ as features
    Predict $\hat{\mathbf{y}}$ using trained random forest model
end for
Step 6: Evaluation
Calculate metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and
    Goodness-of-Fit (GoF) for predictions
Step 7: Output
Return trained NNRF predictions for each site and evaluation results
```

3. Simulation Setup

In this study, we trained the NNGP and NNRF models on datasets from eight different sites, and compared their performances in terms of accuracy and computational speed. The datasets were first cleaned of outliers and missing data. Details of the preprocessing steps are provided in the appendix. The temporal features were engineered to capture the periodic patterns in the dataset. The Euclidean distances between the sites were computed to determine the nearest neighboring sites. The NNGP model was first built using the squared exponential kernel, followed by the replacement of the GP with RF to form the NNRF model. Both models were developed in Python version 3.11.7. Major packages used in the code implementation include scikit-learn, mainly for the machine learning modeling, and geopandas for the spatial analysis of the sites' coordinates. Both models were executed on a local desktop equipped with an Intel® Core™ processor featuring 16 CPUs, each operating at a clock speed of 2.0 GHz. Additionally, the system was supported by 16 GB of RAM. The rest of the simulation setup describes the dataset used in the study, the implementation details of NNGP and NNRF, scalability test of the models, and the evaluation metrics.

3.1. Data Collection

Texas has significant solar energy potential and deployed solar resources, with annual average global tilt solar radiation ranging from 4.76 kWh/m^2/day to 6.58 kWh/m^2/day across the state [40]. Its abundant solar resources, combined with the unique characteristics of the isolated ERCOT grid, make it an ideal region for this study. Eight sites across Texas were selected for analysis, with their locations and coordinates listed in Table 1, and their spatial distribution illustrated in Figure 2. In this study, Sites 2 and 7 share the same global 4 km^2 grid space, while each other site has a distinct global region.

Table 1. Selected Sites and their Locations

Name	Site	Latitude	Longitude
3-Door, Ovalo	1	32.154290	-99.81891
Camp David Dallas	2	32.776600	-96.79706
Hermitage Austin	3	30.26726	-97.74281
Kruckenland San Antonio	4	29.5213	-98.5101
Misty Morn Cedar Park	5	30.49450	-97.82257
Roof Station Houston	6	29.7554	-95.37575
Weather Toy Dallas	7	32.78709	-96.79605
Wiley, Leander	8	30.58262	-97.86712

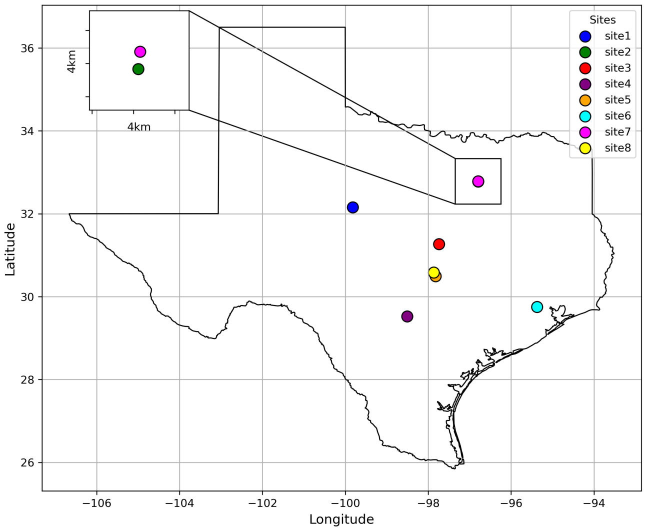


Figure 2. Spatial representation of selected sites in Texas. The inset provides a zoomed-in view of Site 2 and Site 7 located in the same 4km by 4km global space.

The spatial distribution of the selected sites, shown in Figure 2, indicates that most of the sites are sparsely distributed across the region. However, certain locations—specifically Sites 3, 4, 5, and 8 form a relatively close cluster, while Sites 2 and 7 share the same global space as shown by the inset plot in Figure 2. Understanding this spatial arrangement is essential for accurately capturing the influence of neighboring sites on local conditions. Table 2 presents the ground distances between the sites in km, further illustrating their relative proximity and separation.

Table 2. Distances between sites (in km)

Site	1	2	3	4	5	6	7	8
1	0.00	291.81	219.76	318.32	264.60	500.59	292.17	254.71
2	291.81	0.00	190.04	396.97	271.69	362.07	1.17	264.13
3	219.76	190.04	0.00	207.62	86.26	282.26	191.11	77.05
4	318.32	396.97	207.62	0.00	126.86	304.03	398.07	133.25
5	264.60	271.69	86.26	126.86	0.00	249.26	272.81	10.69
6	500.59	362.07	282.26	304.03	249.26	0.00	363.12	256.55
7	292.17	1.17	191.11	398.07	272.81	363.12	0.00	265.25
8	254.71	264.13	77.05	133.25	10.69	256.55	265.25	0.00

The global dataset for each site was obtained from the NSRDB website [20], while the local dataset was sourced from Ambient Weather [41]. The global dataset consists of 5-minute interval data on solar irradiance, temperature, pressure, wind speed, and dew point, spanning the period from January 1, 2022, to December 31, 2022. During the same period and with the same temporal resolution, ground-measured solar irradiance data was collected as the local irradiance dataset. The dataset for each site has 105,120 data points, enough to effectively train a machine learning model.

Figures 3 and 4 present the annual profiles of global and local irradiance, respectively, for Site 1 in 2022. The plots show significant differences between the global and local irradiance. This is confirmed by the residual plot in Figure 5, which shows the numerical deviations of the global and local irradiance at each time point for Site 1. The global irradiance, being forecast values by NSRDB, do not accurately represent the ground measurements. In the same manner, the datasets of the other seven sites have similar distribution and deviation between the global and local irradiance. This affirms the need to downscale the forecast global irradiance to a more accurate site specific values.

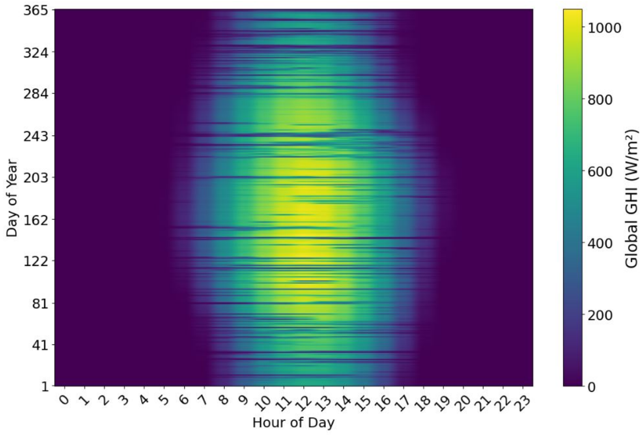


Figure 3. Heatmap of Global Irradiance Profile for Site 1 (2022)

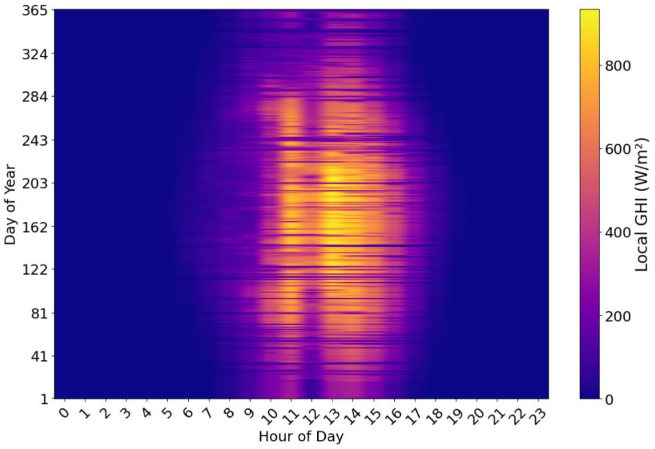


Figure 4. Heatmap of Local Irradiance Profile for Site 1 (2022)

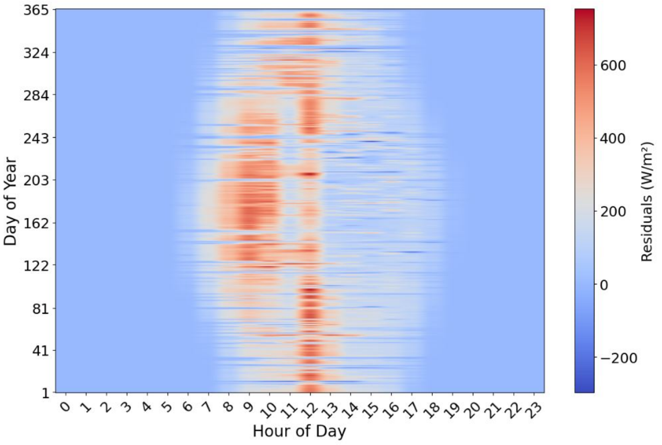


Figure 5. Heatmap of Deviation Between Global and Local Solar Irradiance for Site 1 (2022)

3.1.1. Feature Engineering

Solar irradiance is influenced by temporal and meteorological factors. Therefore, critical temporal features such as the hour of the day and day of the year were encoded using sine and cosine transformations to capture cyclical trends accurately. This cyclical encoding ensures continuity across periodic boundaries, such as midnight to midnight or December to January transitions. Capturing these cyclical temporal variations is crucial in reducing predictive errors, particularly during transitional periods with rapid changes in solar angles. Meteorological parameters such as temperature, atmospheric pressure, wind speed, and dew point from the global dataset were incorporated to reflect weather-dependent variability in irradiance. Integrating meteorological data provides essential contextual information which improves the model's ability to generalize under diverse weather conditions and thus enhances its robustness. The `Hour_sin` and `Hour_cos` features were computed using equations:

$$\text{Hour\_sin} = \sin\left(\frac{2\pi \cdot \text{Hour}}{24}\right), \tag{13}$$

$$\text{Hour\_cos} = \cos\left(\frac{2\pi \cdot \text{Hour}}{24}\right). \tag{14}$$

Similarly, the `DayOfYear_sin` and `DayOfYear_cos` features were computed as:

$$\text{DayOfYear\_sin} = \sin\left(\frac{2\pi \cdot \text{DayOfYear}}{365}\right), \tag{15}$$

$$\text{DayOfYear\_cos} = \cos\left(\frac{2\pi \cdot \text{DayOfYear}}{365}\right). \tag{16}$$

Additionally, meteorological factors such as temperature, pressure, wind speed, and dew point were included to capture the impact of weather conditions on solar irradiance. These features, together with the global solar irradiance were scaled using *scikit-learn's MinMax Scaler*.

3.2. NNGP Implementation

The NNGP model was implemented to predict site-specific GHI values using a combination of spatial, temporal, and feature-based distances. For each site, its k-NNs were identified based on Euclidean distance in the spatial domain, with geographic coordinates serving as inputs. By focusing solely on these neighbors, the NNGP framework effectively localized the GP. This significantly reduced the dimensionality of the covariance matrices. Temporal dependencies were incorporated by considering the time difference between the current observation and historical data from neighboring sites. Temporal distances were calculated in minutes to ensure fine-grained temporal resolution suitable for solar irradiance modeling. Furthermore, feature-based distances were included to account for meteorological conditions such as temperature, pressure, wind speed, and dew point. These feature-based distances were computed in the feature space to capture similarities and differences in environmental conditions across sites.

The squared exponential kernel in (5), used for the computation of the covariances incorporated spatial, temporal, and feature-based distances, with separate length-scale parameters (ℓ_s, ℓ_t, ℓ_g, σ) for each distance type. Larger values of these parameters implied distance neighbors and observations had significant influence on the prediction outcome. These parameters, including the variance (σ), were tuned to balance their influence on the predictions. The final values of the hyperparamters used for the results in this paper are presented in Table 3.

Table 3. Values of Hyperparameters for the NNGP model

Hyperparameter	Value
`num_neighbors` (k)	3
σ	2.0
ℓ_s	0.1
ℓ_t	288 (1 day)
ℓ_g	0.5
penalty term	0.2

To enhance the predictive capacity of the model, cyclical encoding of time was applied to represent the periodic nature of hours and days of the year. The covariance matrix among training points ($K_{\text{train, train}}$) was regularized for numerical stability during inversion. Regularization was achieved by adding a small offset penalty term to the training covariance matrix to prevent singularities. Predictions were made using the GP regression formula provided in (6). This was achieved by combining the computed covariance matrices to scale the residuals between the neighbors local GHI and their predicted values from the linear regression model computed in (7). The resulting scaled residuals, known as the Gaussian correction term, was added to the linear regression estimated values of the target site to obtained the final downscaled GHI.

3.3. NNRF Implementation

Similar to the NNGP model, the NNRF model was developed to predict site-specific GHI values by leveraging spatial, temporal, and meteorological information. For each site, the model utilized its k-NNs, determined using Euclidean distances computed from the geographic coordinates of the sites. A k value of 3 was used in this model. This nearest-neighbor framework enabled the model to incorporate local spatial relationships, ensuring that predictions were informed by the most relevant neighboring data. Temporal dependencies were integrated by incorporating the encoded *Hour* and *Hour* as part of the input features. This was done to ensure that training data reflected consistent diurnal and seasonal patterns. Additionally, meteorological features such as temperature, pressure, wind speed, and dew point were included to capture the environmental factors influencing solar irradiance variability.

The global GHI values of neighboring sites were also included as additional features to provide spatial context for each prediction. These neighbor features were aligned temporally to ensure consistency between the target site and its neighbors. The RF model, comprising an ensemble of decision trees, was trained using these features to predict the local GHI values. The model was initialized with 150 estimators to balance accuracy and computational efficiency. The ensemble approach enabled the RF model to handle complex non-linear relationships between input features and the target variable.

3.4. Scalability Test

Additionally, we investigated into how well both models scale in terms of accuracy and computational speed with varying numbers of k-NNs. We achieved this by adding seven more sites to the existing ones to make fifteen, and retrained the models on a new dataset. To ensure adequate evaluation of the model's scaling performance, the new training dataset were collected from from January 1 to December 30, 2023, different from the former training dataset. Similar to earlier simulation, the dataset for December 31, 2023 was excluded to validate the accuracy of the scaled models. The k-NNs were varied from 2 to 14 for each of the sites. As a result of the increased data size, the scalability test was performed on one compute node of the South Dakota State University's Innovator's Cluster equipped with 2 Intel Xeon Gold 6342 CPU with 48 cores. Each processor operated at a clock speed of 2.80 GHz and supported by a 256 GB RAM.

3.5. Evaluation of Models

After the initial training of the two models on the dataset from January 1, 2022 to December 30, 2022, the models were validated by making day-ahead predictions of the downscaled local GHI for all eight sites for December 31, 2022. The evaluation of the models performance were achieved using Mean Absolute Error (MAE) and the Goodness of Fit (GoF). The MAE was computed as:

$$\mathrm{MAE} = \frac{1}{\mathrm{N}} \sum_{i=1}^{n} |y_i - \hat{y}_i|, \tag{17}$$

where y_i is the actual or true value of the i-th observation, $\hat{y}_i$ is the predicted value corresponding to the i-th observation, and N denotes the sample size in the testing set. To measure how well the estimated models capture the patterns of the dataset, the GoF is computed using the normalized Root Mean Squared Error (NRMSE), such that:

$$\mathrm{GoF}(Percent) = (1 - \mathrm{NRMSE}) * 100, \tag{18}$$

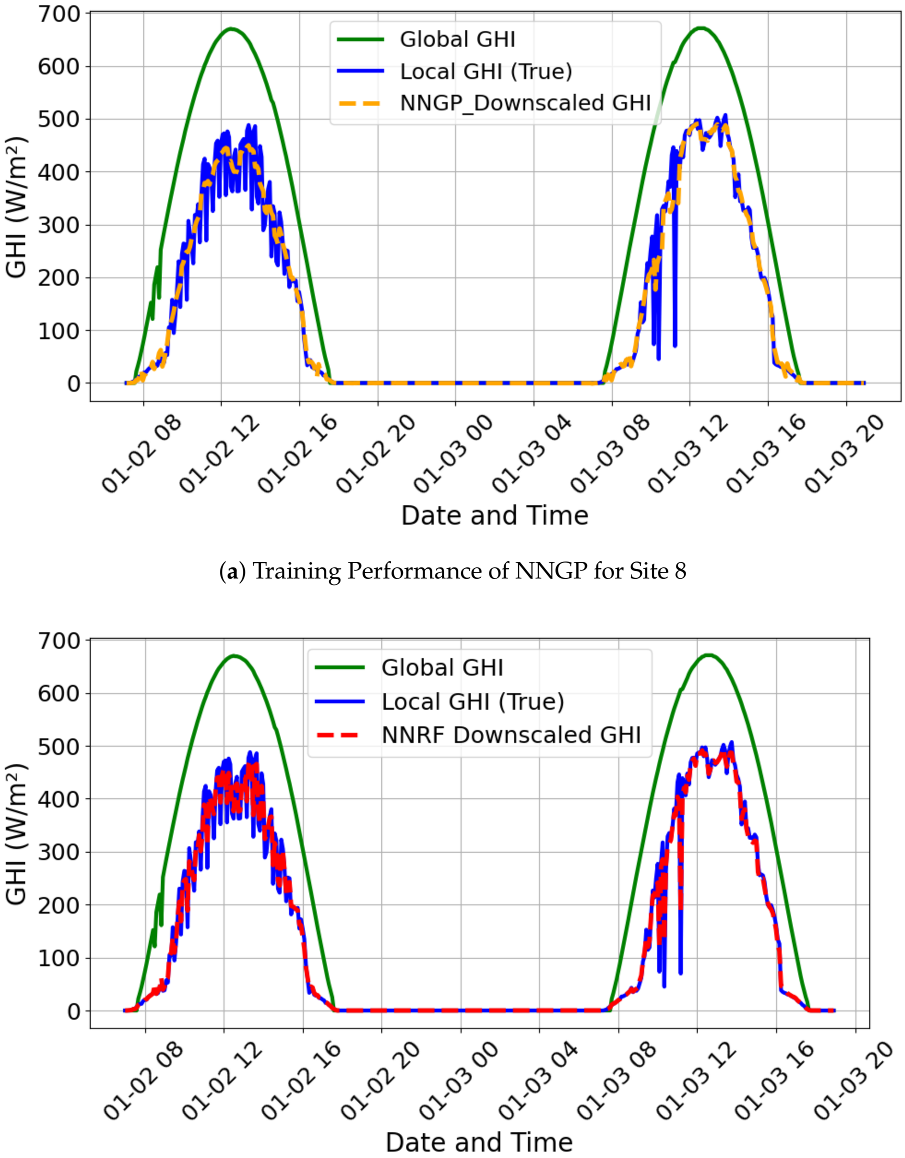
where NRMSE is given by:

$$\mathrm{NRMSE} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{y_{\max} - y_{\min}}. \tag{19}$$

4. Results and Discussion

4.1. Training Performance of NNRF and NNGP

Both the NNGP and NNRF perform very well during the training phase, however the NNRF shows slightly superior accuracy over NNGP. The performance of NNRF is consistently higher than NNGP across all the eight sites, as could be seen from Table 4. The NNRF records an average GoF of 97.63% across all sites, compared to 95.87% of the NNGP. This superior performance of the NNRF model can be attributed to RF's ability to learn non-linear patterns in the dataset, in contrast to the NNGP inherent linearity. Furthermore, averaging the predictions of several decision trees makes the RF model robust to noise and outliers. This makes it comparatively more suitable for non-smooth data such as solar irradiance. Visual displays of the training performance of both models for Site 8 are shown in Figure 6a,b.



(**a**) Training Performance of NNGP for Site 8

(**b**) Training Performance of NNRF for Site 8

Figure 6. Training performance comparison for Site 8

A careful observation of the two plots reveals how the NNRF model accurately captures the variabilities of the local irradiance better than the NNGP model. However, even though the RF model gives some partial interpretability in terms of the number of decision tree estimators and tree depth, the GP offers a better interpretability in terms of the correlation and influence of neighboring sites on

doi:10.20944/preprints202503.0629.v1

predictions. The interpretability of NNRF can be further explored using explainable AI (xA), such as SHapley Additive exPlanations (SHAP) values [42]. The GP's hyperparameters such as the temporal, spatial, and feature length scales, and the variance give visibility of how far in time space or feature space the model looks for correlations. Similarly, the covariance matrices structure reveals how points influence each other's predictions.

This improved interpretability of the NNGP however, comes with increased computational cost. The NNGP took **1004.19** seconds to train on the eight sites, compared to **407.63** seconds of the NNRF model. By avoiding the computation and inversion of covariance matrices, the NNRF model speeds up computational speed by **2.5** times more than the NNGP model. While these results are promising, validation is required to confirm how well they generalize, and scale with increasing number of neighboring sites, which is discussed later in the results.

Table 4. Comparison of NNRF and NNGP Training Performance

Site	NNRF			NNGP		
	MAE	NRMSE	GoF (%)	MAE	NRMSE	GoF (%)
1	13.6480	0.0278	97.22	19.8886	0.0479	95.21
2	8.1693	0.0178	98.22	12.7651	0.0303	96.97
3	13.5003	0.0256	97.44	20.5084	0.0446	95.54
4	11.9800	0.0271	97.29	19.0394	0.0475	95.25
5	12.0539	0.0245	97.55	20.7114	0.0441	95.59
6	13.6236	0.0250	97.50	20.4696	0.0432	95.68
7	5.9689	0.0183	98.17	9.2296	0.0316	96.84
8	12.4045	0.0235	97.65	19.6315	0.0410	95.90

4.2. Validation Performance of NNRF and NNGP

The final evaluation of the NNRF was done by predicting the downscaled local solar irradiance for the next day to be used in day-ahead power system operations. The dataset for December 31, 2022, used for this validation was excluded from the training. The features of the global data and the global solar irradiance of the selected neighboring sites were used as input variables for the predictions. The day-ahead forecast was made for each site in 5 minutes time step, and compared to the actual local solar irradiance of each site. The plots of the predicted downscaled solar irradiance for Sites 1, 2, 7, and 8 are presented in Figure 7. The orange line represents the NNGP's predictions and the red depicts the predictions of the NNRF. Table 5 displays the model's evaluation results for each site.

Once again, the NNRF performed better across all eight sites with an average GoF of 90.61%, compared to 86.31% of the NNGP model. This is confirmed by the plots in Figure 7, where the NNRF's predictions closely follow the trajectory of the local GHI better than the NNGP model. The boxplots in Figure 8 show the residuals of the actual and predicted values. While each model is approximately mean 0, the variance of the NNGP is much larger than the NNRF (which appears in the other error metrics). Additionally, there are larger outliers predicted by NNGP, such as the -400 point in Site 1. All these observations confirm the superior performance of the NNRF model over the NNGP model.

Aside from these observations, the NNGP predictions show an interesting trend where it performed better for sites with less rapid variation in the local irradiance profile. Likewise, its performance is relatively lower for sites with rapid local irradiance variation, which deviates more from the shape of the global solar irradiance. For instance, as can be seen in Figure 7, Site 1 and Site 7, which recorded the lowest GoF, also display higher deviations from the profile of the global irradiance. This is can been seen from Table 6, where they show relatively lower correlation values of 0.7635 and 0.5797 respectively. Similarly, with the exception of Site 3, it is observed that the more the shape of the local and global irradiance are highly correlated the higher the NNRF prediction's accuracy. This implies that generally, the alignment of the shape profiles of the global and local irradiance has a high impact on the accuracy of prediction than the value of their variance. The MAE and the NRMSE also follow similar pattern, where Site 7, whose profile's shape deviates most from that of the global irradiance, recorded the highest prediction error. In sharp contrast, Site 8 records the minimum prediction error due its high correlation value of 0.9707 between the global and local GHI.

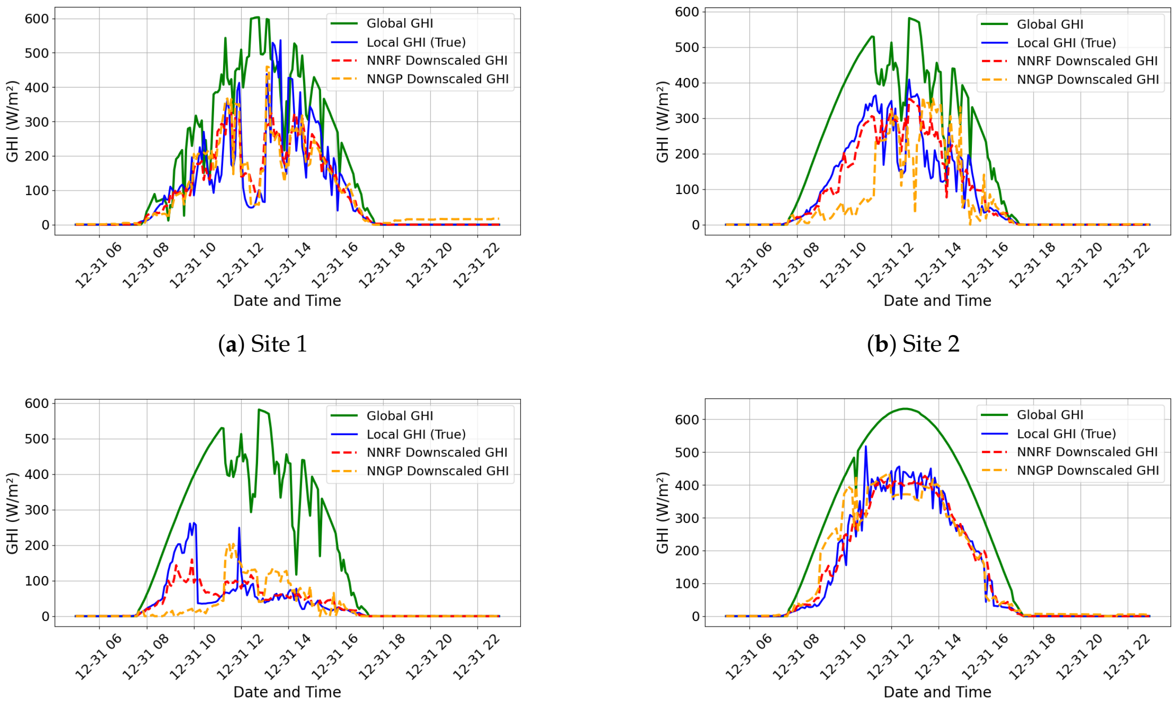
This high correlation makes the global solar irradiance a good predictor of the local solar irradiance. However, when the local irradiance differs significantly from the global profile due to localized weather conditions such as cloud cover, terrain shading, or microclimate effects, the global irradiance features are unable to fully explain these local variations. The departure of Site 3 from this trend could be as the result of lower influence of the other features such as wind, temperature, pressure, and dew point on the predictions.

Table 5. Validation Results of NNRF and NNGP for Day-Ahead Forecast

Site	NNRF			NNGP		
	MAE	NRMSE	GoF (%)	MAE	NRMSE	GoF (%)
1	32.4447	0.1162	88.38	36.4148	0.1197	88.03
2	22.1226	0.0986	90.14	50.8843	0.2216	77.84
3	35.8097	0.1115	88.85	43.9536	0.1385	86.15
4	21.4388	0.0728	92.72	25.2055	0.0835	91.65
5	26.0342	0.1047	89.53	29.8221	0.1187	88.13
6	18.8950	0.0748	92.52	28.3873	0.1185	88.15
7	14.4113	0.1209	87.91	29.4329	0.2304	76.96
8	16.3520	0.0594	94.06	25.8150	0.0946	90.54

Table 6. Correlations Between Global and Local Solar Irradiance

Site	1	2	3	4	5	6	7	8
Correlation	0.7635	0.9346	0.9543	0.9226	0.6487	0.9686	0.5797	0.9707



(**a**) Site 1

(**b**) Site 2

(**c**) Site 7

(**d**) Site 8

Figure 7. Day-Ahead GHI Prediction at Different Sites. Each panel shows the prediction for a specific site: (**a**) Site 1, (**b**) Site 2, (**c**) Site 7, (**d**) Site 8.

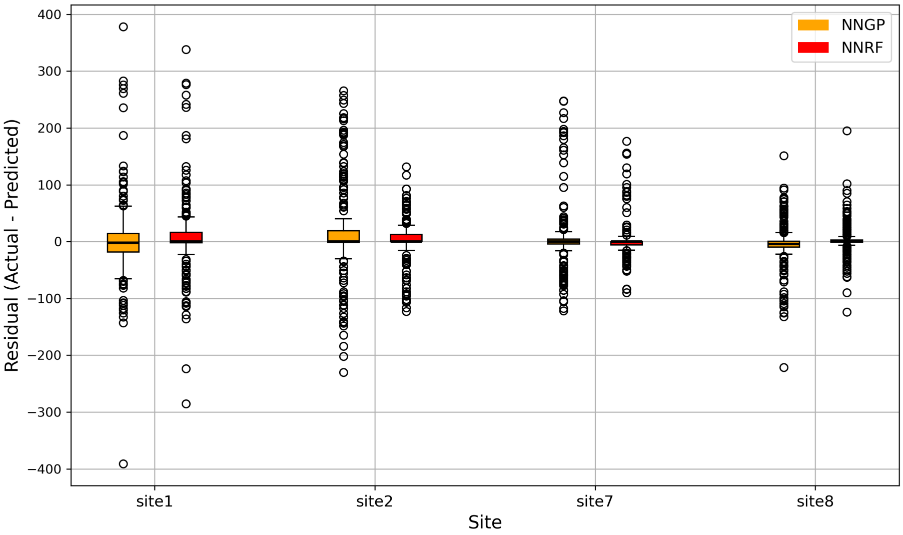
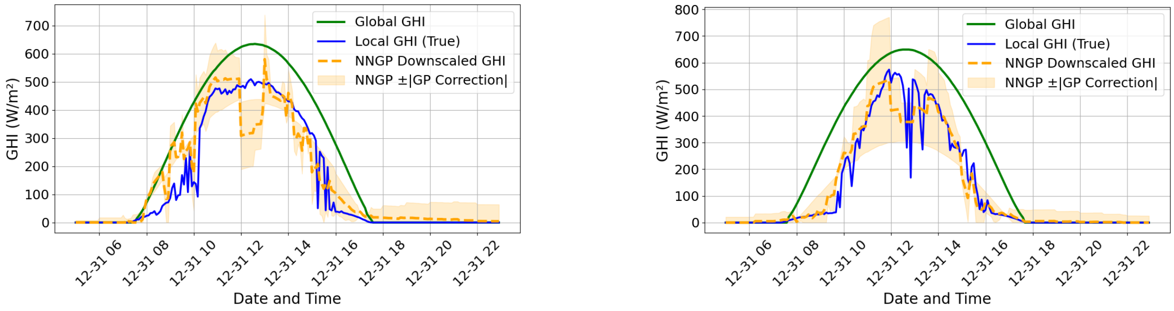


Figure 8. Boxplots of NNGP and NNRF Prediction Residuals for Sites 1, 2, 7, and 8.

While the models performances are satisfactory, the variability in solar irradiance demands quantification of predictions uncertainty to improve grid planning. The NNRF model, though robust and more accurate, produces point estimates only. This makes it unable to give any insight into the prediction tolerance for effective planning. The NNGP model, however displays high strength in this regard. By using the values of the Gaussian corrections as variance, upper and lower bound predictions values could be generated by addition and subtraction with the mean values. These upper and lower bounds give some insight into the accompanying uncertainties of the predicted values. The results of the NNGP's predictions for Site 3 and Site 4 are shown in Figure 9. This makes NNGP model very suitable for probabilistic forecasts. Based on these results, the decision to choose NNRF or NNGP does not depend solely on factors such as accuracy, computational speed and intepretabilities, but also, whether point forecasts or probabilistic forecasts are desirable.



(**a**) Site 3

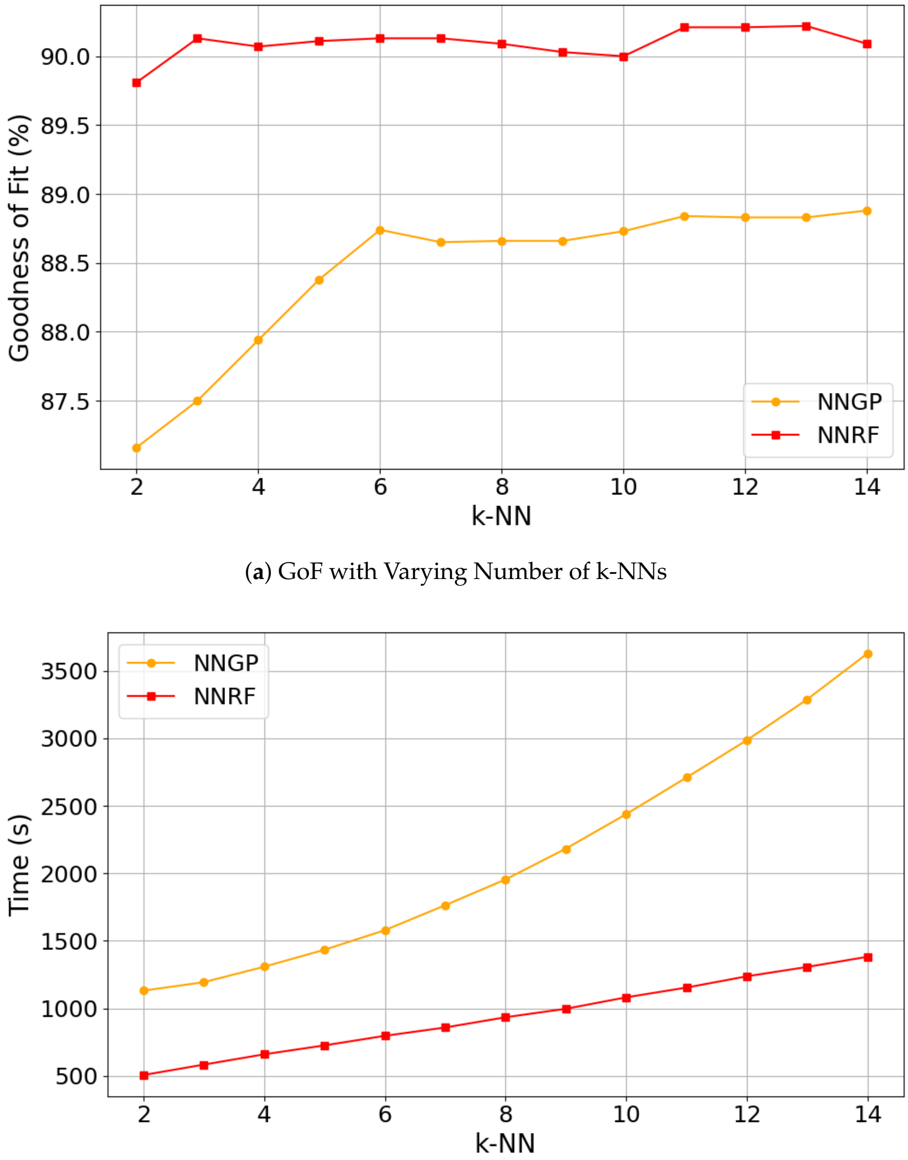
(**b**) Site 4

Figure 9. NNGP's Day-Ahead GHI Predictions with Gaussian Uncertainty for (**a**) Site 3, and (**b**) Site 4

4.3. Scalability of the NNGP and NNRF Models

The scalability results as presented in Figure 10 reveal how the NNRF consistently performs better than the NNGP model for different number of k-NNs. These accuracy values are the GoF of the day-ahead predictions of the NNGP and NNRF for December 31, 2023. It can been seen from Figure 10(**a**) that the NNRF requires as few as 3 k-NNs to achieve very good accuracy. Beyond 3 kNNs, the NNRF model does not seem to record any significant improvement in prediction accuracy. Conversely, the NNGP accuracy's scales linearly from 2 to 6 k-NNs, beyond which the accuracy shows very minimal improvement. If we select 3 k-NNs and 6 k-NNs as the local optimum values for the NNRF and NNGP respectively, the corresponding accuracy are 90.13% for NNRF and 88.74% for NNGP. The equivalent computational time is 582.92 seconds for the NNRF and 1578.60 seconds for the

NNGP. This clearly shows that besides requiring fewer number of k-NNs to achieve a locally optimum accuracy, the NNRF also takes comparably less computational time (in this case, almost three times as fast as the NNGP model). Additionally, based on the results presented in Figure 10(**b**), the NNRF scales linearly, whiles the NNGP displays a polynomial scaling with increasing number of k-NNs. This implies the NNRF responds to a step increase in the number of k-NNs with a scalar multiple of the computational time, while the NNGP responds with an exponential increase in the computational time. This further proves the effectiveness and efficiency of the NNRF model for large scale solar irradiance downscaling tasks, making it more suitable for real-time grid integration of DERs.



(**a**) GoF with Varying Number of k-NNs

(**b**) Training Time with Varying Number of k-NNs

Figure 10. Scalability Performance of NNGP and NNRF

5. Conclusions

This study evaluates the performance of the NNRF and NNGP models in downscaling solar irradiance to localized measurements. The NNRF model performs better than the NNGP model in terms of accuracy and computational time during both training and validation. The NNRF recorded an average validation accuracy of 90.61% which outperformed the 85.88% recorded by the NNGP model. The NNRF model also improved computational speed by 2.5 times over the NNGP model. Although the NNGP lags behind in terms of accuracy and computational speed, it shows strength in terms of interpretability and prediction uncertainty quantification. Its hyperparameters such as the variance (σ), temporal (ℓ_t), spatial (ℓ_s) and feature-based (ℓ_g) scale parameters, give insight into how far in time and space past observations and neighboring sites influence the prediction outcome. This makes the NNGP more suitable for probabilistic estimates which demands more transparency of the modeling process.

Further, a scalability tests which measured both models performance and computational speed with varying number of k-NNs from 2 to 14 showed a linear scaling for NNRF and a polynomial scaling for NNGP in terms of computational time. Similarly, the NNRF achieved a local optimal accuracy with 3 kNNs whiles the NNGP took 6 k-NNs to obtain a local optimal accuracy. Irrespective of the increased number of k-NNs, the NNGP still could not outperform the NNRF, which had a scaling accuracy of 90.13% compared to 88.74% of the NNGP. These findings proves the superiority of the NNRF for large scale solar irradiance downscaling tasks.

These findings could be very useful in the implementation of the FERC order 2222, by enabling real time accurate estimation of solar PV capacity for day ahead dispatch. Additionally, due the improved accuracy and computational speed of the NNRF model, its application can be extended to the real-time electricity market with accurate PV estimates on a 5 minute rolling basis. Finally, the use of downscaled solar irradiance forecasts, reduces the level of solar PV generation uncertainties with its associated reserves requirements and instability issues in power dispatch planning.

Author Contributions: Conceptualization: S.T.A., H.M.R., T.M.H.; methodology: S.T.A., A.S., H.M.R., T.M.H.; software: S.T.A., A.S.; validation: S.T.A., A.S., T.M.H.; formal analysis: S.T.A., A.S.; data curation: S.T.A., A.S.; writing—original draft preparation: S.T.A., A.S., T.M.H.; writing—review and editing: Z.W., H.M.R., T.M.H.; visualization: S.T.A., A.S.; supervision: Z.W., H.M.R., T.M.H.; funding acquisition: T.M.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the SDSU Richard McComish Ph.D. Fellowship in Energy Infrastructure, the U.S. Department of Energy Office of Science, Office of Electricity Microgrid R&D Program, and Office of Energy Efficiency and Renewable Energy Solar Energy Technology Office under the EPSCoR grant number DE-SC0020281, and the U.S. National Science Foundation grant no. 2316400.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset utilized in this study, along with the corresponding code, results, and visualizations are publicly accessible on GitHub. Interested readers can access these resources at the following link: https://github.com/ShadrackTAsiedu/Downscaling-NNGP-NNRF.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ANN	Artificial Neural Network
CPU	Central Processing Unit
DER	Distributed Energy Resource
ERCOT	Electric Reliability Council of Texas
FERC	Federal Energy Regulatory Commission
GB	Giga Bytes
GHI	Global Horizontal Irradiance
GHz	Giga-Hertz
GoF	Goodness of Fit
GP	Gaussian Process
MAE	Mean Absolute Error
NN	Nearest-Neighbor

NNGP	Nearest-Neighbor Gaussian Process
NNRF	Nearest-Neighbor Random Forest
NRMSE	Normalized Root Mean Squared Error
NSRDB	National Solar Radiation Database
NWP	Numerical Weather Prediction
PEC	Power Electronic Converter
PV	Photovoltaic
RAM	Random Access Memory
RF	Random Forest
RMSE	Root Mean Squared Error
VREs	Variable Renewable Energy Resources

Appendix A. Data Processing

The pre-processing steps include handling raw data inconsistencies, ensuring uniform temporal resolution, addressing missing data, and engineering features critical for downstream modeling. Below, we detail each stage of the data pre-processing pipeline.

Appendix A.1. Raw Data Cleaning

The raw irradiance data from eight sites were provided as individual CSV files. Each dataset contained a timestamp and solar irradiance measurements, with inconsistencies in date formats and row order. To address these challenges, each dataset was loaded, and the `Date` column was converted to a standardized `datetime` format for uniformity. The rows were reversed to ensure chronological order, starting from the earliest date, and the indices were reset to sequential integers starting from 1. The column containing solar irradiance data, originally named inconsistently across files, was renamed to `GHI_local` for uniformity. The cleaned datasets were stored in a dictionary, with site names as keys for efficient access.

Appendix A.2. Temporal Alignment and Missing Data Handling

Global solar irradiance data are often recorded at irregular intervals, leading to temporal misalignment and missing values. For accurate spatiotemporal modeling, a complete time series with a 5-minute resolution was generated for the entire year (2022), yielding a total of 105,120 expected timestamps. Each site's dataset was reindexed to align with the expected date range, and missing timestamps were identified by comparing the reindexed dataset to the complete time series. Missing irradiance values in the local irradiance data were imputed using a trend-based approach. For each missing value, data from the same time of day in the previous or subsequent valid days were averaged over a 7-day window. If sufficient historical data were unavailable, forward imputation was used as a fallback. Additionally, the cleaned datasets were capped at a maximum irradiance value of $1200W/m^2$ to avoid outliers resulting from sensor anomalies.

Appendix A.3. Data Validation and Visualization

To verify the integrity of the cleaned data, temporal trends in solar irradiance were visualized for each site. The local irradiance datasets were plotted over time for the entire year, revealing daily and seasonal variability across the sites. Each site's dataset was inspected for remaining inconsistencies, such as abrupt changes in irradiance values, which could indicate potential anomalies. The annual local solar irradiance for Site 1, before and after cleaning and filling missing data are shown in Figure A1 and Figure A2.

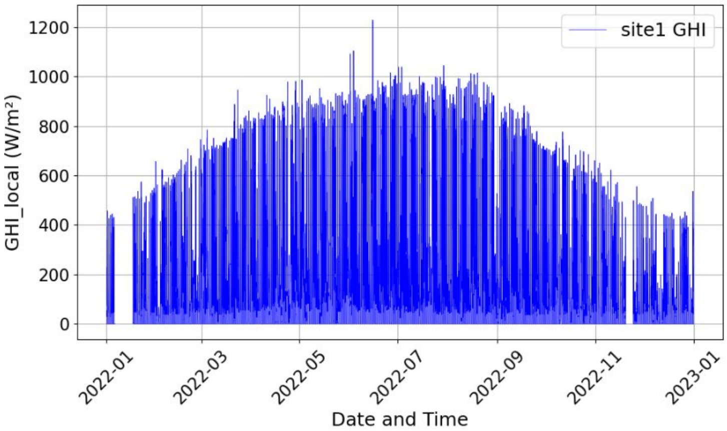


Figure A1. Raw Data on Local Solar Irradiance for Site 1 (2022)

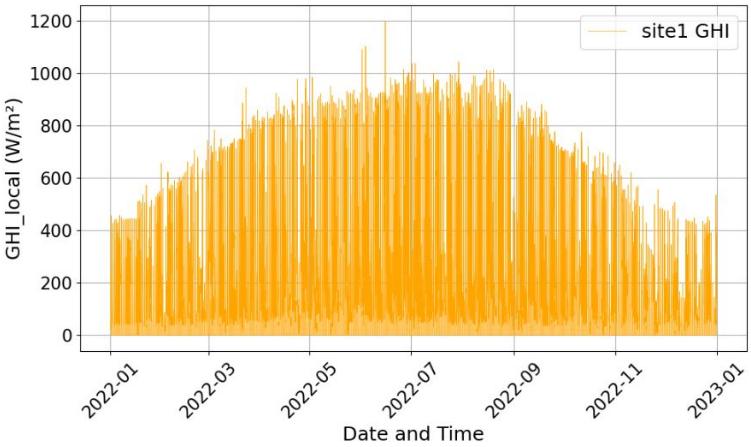


Figure A2. Cleaned Data on Local Solar Irradiance for Site 1 (2022)

References

1. National Renewable Energy Laboratory (NREL). Technical Report NREL/TP-5C00-80166. Technical report, U.S. Department of Energy Office of Energy Efficiency & Renewable Energy, Golden, Colorado, 2021. Contract No. DE-AC36-08GO28308. Operated by the Alliance for Sustainable Energy, LLC.
2. Eldridge, B.C.; Somani, A. Impact of FERC Order 2222 on DER Participation Rules in US Electricity Markets. Technical report, Pacific Northwest National Laboratory (PNNL), Richland, WA (United States), 2022.
3. Garg, S.; Tyagi, S. A Comprehensive Review on Opportunities and Challenges of Grid Integration of Renewable Energy Resources. In Proceedings of the 2024 Second International Conference on Smart Technologies for Power and Renewable Energy (SPECon), 2024, pp. 1–6. https://doi.org/10.1109/SPECon61254.2024.10537518.
4. Subedi, S.; Poudel, B.; Aslami, P.; Fourney, R.; Rekabdarkolaee, H.M.; Tonkoski, R.; Hansen, T.M. Automated Data–Driven Model Extraction and Validation of Inverter Dynamics with Grid Support Function. *e-Prime - Advances in Electrical Engineering, Electronics and Energy* **2023**, *6*, 100365. https://doi.org/https://doi.org/10.1016/j.prime.2023.100365.
5. Tamrakar, U.; Shrestha, D.; Maharjan, M.; Bhattarai, B.P.; Hansen, T.M.; Tonkoski, R. Virtual Inertia: Current Trends and Future Directions. *Applied Sciences* **2017**, *7*. https://doi.org/10.3390/app7070654.
6. Tiwari, S.; Sabzehgar, R.; Rasouli, M. Short Term Solar Irradiance Forecast Using Numerical Weather Prediction (NWP) with Gradient Boost Regression. In Proceedings of the 2018 9th IEEE International Symposium on Power Electronics for Distributed Generation Systems (PEDG), 2018, pp. 1–8. https://doi.org/10.1109/PEDG.2018.8447751.
7. Dou, W.; Wang, K.; Shan, S.; Zhang, K.; Wei, H. Day-Ahead Correction of Numerical Weather Prediction Solar Irradiance Forecasts Based on Similar Day Analysis. In Proceedings of the 2023 35th Chinese Control and Decision Conference (CCDC), 2023, pp. 2554–2559. https://doi.org/10.1109/CCDC58219.2023.10327305.

doi:10.20944/preprints202503.0629.v1

8. De Giorgi, M.G.; Congedo, P.M.; Malvoni, M. Photovoltaic power forecasting using statistical methods: impact of weather data. *IET Science, Measurement and Technology* **2014**, *8*, 90–97. https://doi.org/10.1049/iet-smt.2013.0135.
9. Asiedu, S.T.; Nyarko, F.K.; Boahen, S.; Effah, F.B.; Asaaga, B.A. Machine learning forecasting of solar PV production using single and hybrid models over different time horizons. *Heliyon* **2024**, *10*, e28898. doi: 10.1016/j.heliyon.2024.e28898, https://doi.org/10.1016/j.heliyon.2024.e28898.
10. Gayathry, V.; Kaliyaperumal, D.; Salkuti, S.R. Seasonal solar irradiance forecasting using artificial intelligence techniques with uncertainty analysis. *Scientific Reports* **2024**, *14*, 17945. https://doi.org/10.1038/s41598-024-68531-3.
11. Garg, S.; Agrawal, A.; Goyal, S.; Verma, K. Day Ahead Solar Irradiance Forecasting using Different Statistical Techniques. In Proceedings of the 2020 IEEE International Conference on Power Electronics, Drives and Energy Systems (PEDES), 2020, pp. 1–4. https://doi.org/10.1109/PEDES49360.2020.9379907.
12. Ziyabari, S.; Du, L.; Biswas, S. A Spatio-temporal Hybrid Deep Learning Architecture for Short-term Solar Irradiance Forecasting. In Proceedings of the 2020 47th IEEE Photovoltaic Specialists Conference (PVSC), 2020, pp. 0833–0838. https://doi.org/10.1109/PVSC45281.2020.9300789.
13. Hari, N.G.; G, J. Solar Irradiance Prediction using Deep Learning-Based Approaches. In Proceedings of the 2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2022, pp. 1–5. https://doi.org/10.1109/CSDE56538.2022.10089282.
14. Kartini, U.T.; Hariyati.; Aribowo, W.; Wardani, A.L. Development Hybrid Model Deep Learning Neural Network (DL-NN) For Probabilistic Forecasting Solar Irradiance on Solar Cells To Improve Economics Value Added. In Proceedings of the 2022 Fifth International Conference on Vocational Education and Electrical Engineering (ICVEE), 2022, pp. 151–156. https://doi.org/10.1109/ICVEE57061.2022.9930352.
15. Zafar, R.; Chung, I.Y. Day-Ahead Solar Irradiance Forecasting using a Hybrid Weather-Based Attention BiLSTM Approach for Power System Operation Scheduling. In Proceedings of the 2023 IEEE PES Innovative Smart Grid Technologies Europe (ISGT EUROPE), 2023, pp. 1–5. https://doi.org/10.1109/ISGTEUROPE56780.2023.10408404.
16. Panamtash, H.; Zhou, Q. Coherent Probabilistic Solar Power Forecasting. In Proceedings of the 2018 IEEE International Conference on Probabilistic Methods Applied to Power Systems (PMAPS), 2018, pp. 1–6. https://doi.org/10.1109/PMAPS.2018.8440483.
17. Kakimoto, M.; Endoh, Y.; Shin, H.; Ikeda, R.; Kusaka, H. Probabilistic Solar Irradiance Forecasting by Conditioning Joint Probability Method and Its Application to Electric Power Trading. *IEEE Transactions on Sustainable Energy* **2019**, *10*, 983–993. https://doi.org/10.1109/TSTE.2018.2858777.
18. Opoku, K.; Lucemo, S.; Zhou Sun, Q.; Dimitrovski, A. A Bayesian Approach to Probabilistic Solar Irradiance Forecasting. In Proceedings of the 2022 North American Power Symposium (NAPS), 2022, pp. 1–6. https://doi.org/10.1109/NAPS56150.2022.10012253.
19. Ziyabari, S.; Du, L.; Biswas, S.K. Multibranch Attentive Gated ResNet for Short-Term Spatio-Temporal Solar Irradiance Forecasting. *IEEE Transactions on Industry Applications* **2022**, *58*, 28–38. https://doi.org/10.1109/TIA.2021.3130852.
20. National Renewable Energy Laboratory (NREL). National Solar Radiation Database (NSRDB), 2024. Accessed: 2024-12-02.
21. Zippenfenig, P. Open-Meteo.com Weather API, 2023. Accessed: 2025-01-22.
22. Yang, J.; Sengupta, M.; Bailey, M.; Nychka, D.; Habte, A.; Bandyopadhyay, S.; Xie, Y. Bias Correction and Statistical Downscaling of Future Solar Irradiance Projections Using the NSRDB. In Proceedings of the 2024 IEEE 52nd Photovoltaic Specialist Conference (PVSC), 2024, pp. 793–795. https://doi.org/10.1109/PVSC57443.2024.10748789.
23. Widén, J.; Munkhammar, J. Spatio-Temporal Downscaling of Hourly Solar Irradiance Data Using Gaussian Copulas. In Proceedings of the 2019 IEEE 46th Photovoltaic Specialists Conference (PVSC), 2019, pp. 3172–3178.
24. Huang, J.; Perez, M.; Perez, R.; Yang, D.; Keelin, P.; Hoff, T. Nonparametric Temporal Downscaling of GHI Clear-sky Indices using Gaussian Copula. In Proceedings of the 2022 IEEE 49th Photovoltaics Specialists Conference (PVSC), 2022, pp. 0654–0657.

25. Huang, J.; Perez, M. Temporal Downscaling of GHI Clear-Sky Indices Using T-Copula. In Proceedings of the 2023 IEEE 50th Photovoltaic Specialists Conference (PVSC), 2023, pp. 1–3. https://doi.org/10.1109/PVSC48320.2023.10359613.
26. Munkhammar, J.; Widén, J. Downscaling global, beam and diffuse horizontal irradiance based on hour resolution global horizontal irradiance using Markov mixture distribution modeling. In Proceedings of the 21st Wind & Solar Integration Workshop (WIW 2022), 2022, Vol. 2022, pp. 662–667. https://doi.org/10.1049/icp.2022.2838.
27. Wang, W.; Yin, G.; Zhao, W.; Wen, F.; Yu, D. Spatial Downscaling of MSG Downward Shortwave Radiation Product Under Clear-Sky Condition. *IEEE Transactions on Geoscience and Remote Sensing* **2020**, *58*, 3264–3272. https://doi.org/10.1109/TGRS.2019.2951699.
28. Di Paola, F.; Cimini, D.; De Natale, M.P.; Gallucci, D.; Geraldi, E.; Gentile, S.; Genzano, N.; Larosa, S.; Nilo, S.T.; Ricciardelli, E.; et al. Weather Forecast Downscaling for Applications in Smart Agriculture and Precision Farming using Artificial Neural Networks. In Proceedings of the IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium, 2023, pp. 2791–2794. https://doi.org/10.1109/IGARSS52108.2023.10283439.
29. Wright, T.; Aslami, P.; Dentlinger, M.P.; Hansen, T.M.; Kimn, J.H.; Andrade, F.; Rekabdarkolaee, H.M. Nearest-Neighbor Gaussian Process to Downscale Solar Forecasting at the Grid-Edge for Increased Situational Awareness. In Proceedings of the 2023 North American Power Symposium (NAPS), 2023, pp. 1–6. https://doi.org/10.1109/NAPS58826.2023.10318558.
30. Berkani, S.; Guermah, B.; Zakroum, M.; Ghogho, M. Spatio-Temporal Forecasting: A Survey of Data-Driven Models Using Exogenous Data. *IEEE Access* **2023**, *11*, 75191–75214. https://doi.org/10.1109/ACCESS.2023.3282545.
31. Xu, L.; Chen, N.; Chen, Z.; Zhang, C.; Yu, H. Spatiotemporal forecasting in earth system science: Methods, uncertainties, predictability and future directions. *Earth-Science Reviews* **2021**, *222*, 103828. https://doi.org/https://doi.org/10.1016/j.earscirev.2021.103828.
32. Tascikaraoglu, A.; Sanandaji, B.M.; Chicco, G.; Cocina, V.; Spertino, F.; Erdinc, O.; Paterakis, N.G.; Catalão, J.P. Compressive Spatio-Temporal Forecasting of Meteorological Quantities and Photovoltaic Power. *IEEE Transactions on Sustainable Energy* **2016**, *7*, 1295–1305. https://doi.org/10.1109/TSTE.2016.2544929.
33. Buster, G.; Rossol, M.; Maclaurin, G.; Xie, Y.; Sengupta, M. A physical downscaling algorithm for the generation of high-resolution spatiotemporal solar irradiance data. *Solar Energy* **2021**, *216*, 508–517. https://doi.org/https://doi.org/10.1016/j.solener.2021.01.036.
34. Ulapane, N.; Thiyagarajan, K.; Kodagoda, S. Hyper-Parameter Initialization for Squared Exponential Kernel-based Gaussian Process Regression. In Proceedings of the 2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA), 2020, pp. 1154–1159. https://doi.org/10.1109/ICIEA48937.2020.9248120.
35. Rasmussen, C.E.; Williams, C.K.I. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006.
36. Cutler, A.; Cutler, D.; Stevens, J., Random Forests; 2011; Vol. 45, pp. 157–176. https://doi.org/10.1007/978-1-4419-9326-7_5.
37. Ahmed, H.; Nandi, A.K., Decision Trees and Random Forests. In *Condition Monitoring with Vibration Signals: Compressive Sampling and Learning Algorithms for Rotating Machines*; 2019; pp. 199–224. https://doi.org/10.1002/9781119544678.ch10.
38. Ramasamy, S.; Kantharaju, H.C.; Bindu Madhavi, N.; Haripriya, M.P., 8 Meta-learning through ensemble approach: bagging, boosting, and random forest strategies. In *Toward Artificial General Intelligence: Deep Learning, Neural Networks, Generative AI*; 2024; pp. 167–188.
39. Putro Dirgantoro, G.; Soeleman, M.A.; Supriyanto, C. Smoothing Weight Distance to Solve Euclidean Distance Measurement Problems in K-Nearest Neighbor Algorithm. In Proceedings of the 2021 IEEE 5th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), 2021, pp. 294–298. https://doi.org/10.1109/ICITISEE53823.2021.9655820.
40. Kjellsson, J.; Webber, M. The Energy-Water Nexus: Spatially-Resolved Analysis of the Potential for Desalinating Brackish Groundwater by Use of Solar Energy. *Resources* **2015**, *4*, 476–489. https://doi.org/10.3390/resources4030476.

41. Ambient Weather. Ambient Weather Data and Tools, 2024. Accessed: 2024-12-02.
42. Das, A.; Rad, P. Opportunities and challenges in explainable artificial intelligence (xAI): A survey. *arXiv preprint arXiv:2006.11371* **2020**.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.