Systematic Evaluation of Spatial Interpolation Methods for Reconstructing Missing Solar Irradiance Data for Energy Applications

Lilla Barancsuk
HUN-REN Centre for Energy Research
Budapest, Hungary
lilla.barancsuk@ek.hun-ren.hu

Paolo Lazzeroni
Dipartimento Energia "Galileo Ferraris"
Politecnico di Torino, Torino, Italy
paolo.lazzeroni@polito.it

***Abstract*—This paper presents a systematic comparison of spatial interpolation methods for reconstructing missing solar irradiance data on quasi-regular latitude–longitude grids. Using PVGIS-derived datasets, we evaluate interpolation performance under realistic missing-data scenarios, including (M1) random removal, (M2) contiguous gaps, and (M3) regular-grid thinning. A reproducible evaluation pipeline is used to apply k-nearest neighbors, inverse distance weighting, and analytical grid-based methods (linear, nearest) across multiple regions, orientations, and hyperparameter settings. Reconstruction accuracy is assessed using standard pointwise error metrics. Across all scenarios, K-Nearest Neighbors consistently outperforms alternative methods, demonstrating lower reconstruction errors and higher explanatory power even at increased levels of data removal. The results highlight method-dependent performance trends and trade-offs under different missing-data patterns. The results provide practical guidance for interpolation method selection in solar resource assessment and renewable energy applications.**

***Index Terms*—renewable energy, solar irradiance, spatial interpolation, missing data reconstruction, testing protocol, KNN**

I. Introduction

Solar irradiance data are a key input for renewable energy applications such as photovoltaic (PV) power forecasting, energy system operation, and PV system design and operation [1], [2]. These data are obtained from satellite or ground measurement networks but are often incomplete due to sensor failures and communication outages [3], as well as limited station density due to economic constraints [4]. In addition, many operational tasks require irradiance estimates at locations or spatial resolutions that are not directly observed. Consequently, spatial interpolation and gap-filling of solar irradiance data are essential in both research [5] and energy system operation [6]. A wide range of methods has been proposed, including classical spatial interpolation techniques such as inverse distance weighting (IDW), and splines [2], as well as multivariate imputation and regression-based approaches [4]. More recently, machine learning models [7]–[9] have been applied to irradiance reconstruction, while ensemble methods such as random forests, kriging Gaussian process regression are also used for spatial interpolation in environmental applications [1]. Despite this methodological diversity, existing studies typically focus on specific methods or datasets, often using inconsistent experimental setups and missing-data assumptions. As a result, their findings are difficult to compare and generalize. In contrast, this work provides a structured and reproducible evaluation framework that enables consistent comparison across methods under controlled and application-relevant missing-data scenarios [10]. Using quasi-regular latitude–longitude grids derived from the PVGIS database [11], we construct missing-data scenarios that reflect common operational conditions, including random removal, spatially contiguous gaps, and regular-grid thinning (Fig. 1). Multiple interpolation methods are evaluated under identical conditions and systematic hyperparameter variations, enabling a fair comparison. The main contributions of this work are:

- A standardized comparison of widely used spatial interpolation methods for solar irradiance gap-filling under consistent experimental conditions.
- A systematic evaluation of interpolation performance under realistic spatial missing-data scenarios, including random removal, spatially contiguous gaps, and structured thinning.
- An assessment framework that isolates method performance differences under identical datasets, preprocessing steps, and hyperparameter settings.
- Practical guidance for selecting interpolation methods in solar energy and energy system applications, based on observed performance across different missing-data scenarios.

The code and data used in this study will be made available upon request.

II. Ground truth and missing data

A. Dataset retrieval and preprocessing

To compare the interpolation methodologies mentioned in the Introduction, an irradiance dataset was purposefully created to replicate a measurement campaign. The dataset for this work was created using data from the PVGIS data source [11]. Specifically, the Piedmont region in Italy was selected for the analysis. The region has been rastered on a grid where the

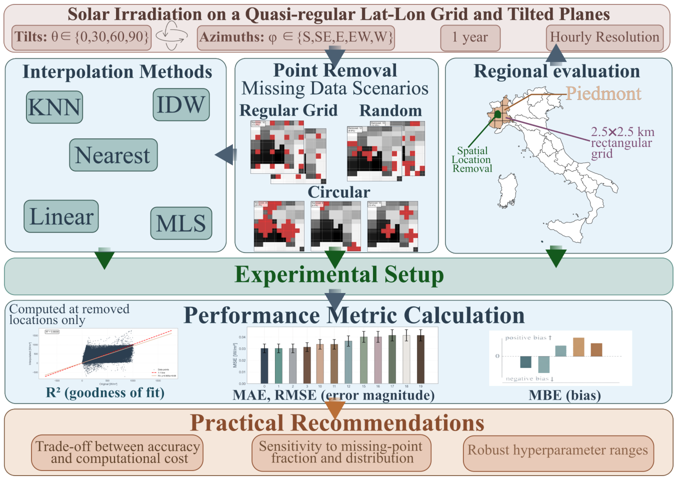

Fig. 1. Overview of the experimental workflow. Missing-data scenarios are generated from PVGIS data, reconstructed using spatial interpolation methods, and evaluated using pointwise error metrics.

distance between two centroids of each couple of adjacent raster cells was fixed at $2.5\,\mathrm{km}$ [12]. As a consequence, quasi regular-grids were obtained due to reprojection effects of the rasters to a geographic Coordinate Reference System (CRS). With the longitude and latitude data of the centroids, solar irradiance data were retrieved from PVGIS-SARAH3 database [13] through the available API [14] by considering the time horizon 2005–2023, a time resolution of one hour, and the tilt and azimuth angles of $\theta = \{0°, 30°, 60°, 90°\}$ and $\phi = \{-60°, -30°, 0°, 30°, 60°\}$, respectively. At this stage, a Typical Meteorological Year (TMY) was generated for each combination of tilt-azimuth angles for each centroid by adopting the methodology of the ISO standard [15] to select the months to form the reference year. Finally, the TMY time series were assigned to virtual stations located at the centroid coordinates. To avoid biasing both interpolation and error metrics, we discard night-time samples for which irradiance is constantly zero (between 21:00 and 04:00). We regard the dataset as a 5D spatio-temporal grid where dimensions correspond to: latitude and longitude (the 2D spatial index is denoted by $\mathbf{x}$), time (t), tilt (θ) and azimuth (ϕ). To simplify notation, we introduce the composite index as:

$$\mathbf{q} := (\mathbf{x}, t, \theta, \phi), \tag{1}$$

where $\mathbf{x} \in \mathcal{L}$ denotes a removed spatial location, $t \in \mathcal{T}$ a daylight time step, $\theta \in \Sigma_{\text{tilt}}$ a tilt angle, and $\phi \in \Sigma_{\text{az}}$ an azimuth angle. Let $\mathcal{Q} = \mathcal{L} \times \mathcal{T} \times \Sigma_{\text{tilt}} \times \Sigma_{\text{az}}$ be the set of all evaluated samples, with cardinality $|\mathcal{Q}| = N$. Observed and reconstructed irradiance values are denoted by $z(\mathbf{q})$ and $\hat{z}(\mathbf{q})$, respectively. Although the data are defined on a 5D grid, all interpolation methods operate on two-dimensional latitude–longitude lattices. A *removed location* is defined as a spatial location $\mathbf{x}$ for which all corresponding tilt, azimuth, and temporal values are jointly removed.

B. Point removal strategies

Missing data in solar irradiance measurements arise from infrastructural and operational failures, resulting in distinct spatial missing-data patterns (scenarios (M1)–(M3)). Because these patterns induce different spatial structures, interpolation performance depends on the underlying missing-data mechanism [16]. To capture this, three complementary missing-data generation strategies are defined. Spatial interpolation is evaluated under controlled conditions by masking a target fraction $p \in \{5\%, 15\%, 25\%\}$ of station locations. For each configuration, the unmasked dataset is retained as ground truth and identical masks are applied across all methods to ensure fair comparison. Examples of the removal masks are seen on Fig. 2. The removal strategies probe robustness under different spatial distributions, with emphasis on spatial clustering. In addition to increasing error with higher removal fractions, the effect of contiguous gap size is analyzed. For each scenario, a single realization of the removal mask is used across all methods, resulting in deterministic results for each configuration. Variability across multiple realizations is not considered and will be addressed in future work.

(M1) Random station removal: This scenario models unstructured and spatially scattered data loss, such as isolated sensor malfunctions, transient calibration errors, or quality-control filtering that removes individual stations without inducing spatially coherent gaps [17]. We uniformly sample R stations without replacement and remove all corresponding measurements (for all orientations and time steps). This baseline represents unstructured outages and provides a reference for comparison.

(M2) Contiguous-gap removal via non-overlapping circles: This scenario represents spatially contiguous outages caused by localized operational failures, such as sensor damage, shared communication or power interruptions, or temporary shutdowns affecting one or more neighboring stations, which result in continuous spatial gaps that must be reconstructed from surrounding observations [18]. To evaluate interpolation performance under continuous missing regions, we generate k non-overlapping circular gaps, with $k \in \{2, 3, 4\}$ fixed per experiment. For a given removal fraction p, the circle radius r is estimated from (p, k) based on local station density so that the expected number of removed stations equals R. Circle centers are selected uniformly at random under a non-overlap constraint, and all stations within radius r are removed. Due to irregular region boundaries, the achieved removal fraction may deviate from p. The radius r is therefore iteratively adjusted until exactly R stations are removed. This procedure produces contiguous gaps with increasing sizes as k decreases.

(M3) Regular-grid removal: This scenario represents systematic spatial thinning, as encountered in sparse measurement networks or when irradiance data are required at a coarser or more uniform spatial resolution [16]. We generate a regular pattern of missing locations by removing stations according to a spatial lattice with spacing q. The spacing is estimated

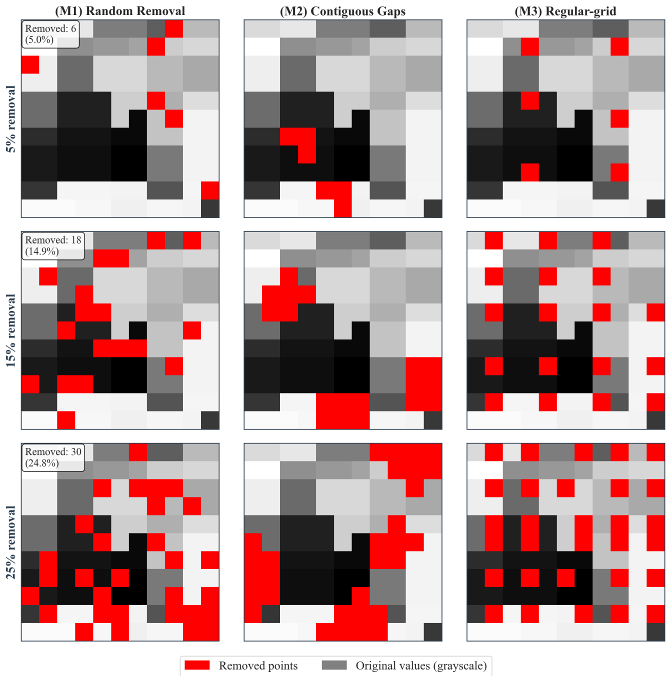

Fig. 2. Missing-data scenarios (columns): (M1) random removal, (M2) contiguous gaps, and (M3) regular-grid removal. Rows correspond to removed-point fractions ($p \in \{5\%, 15\%, 25\%\}$). A $5\,\text{km} \times 5\,\text{km}$ subregion of the Piedmont dataset. Due to the limited number of grid points, some gaps are malformed for (M2). In the $25\,\%$ (M3) case, additional points are clustered along a single row to meet the target removal fraction.

from the target removal fraction p using the heuristic

$$p \approx \frac{1}{q^2}, \tag{2}$$

which corresponds to removing approximately one station per q^2 neighborhood. The lattice is overlaid on the station coordinates, and the nearest station to each grid node is removed, producing an approximately uniform subsampling pattern. Due to irregular station placement, the achieved removal fraction may deviate from p; the grid is therefore iteratively refined until the target number of removed stations is reached, while preserving spatial regularity.

III. Missing Data Reconstruction and Evaluation

This section describes the interpolation methods and evaluation procedure. The workflow consists of generating missing-data scenarios, applying interpolation methods, and evaluating reconstruction accuracy against ground truth. To assess robustness under varying data removal scenarios, we evaluate a set of interpolation methods with different assumptions on spatial smoothness, locality, and functional form, enabling comparison between grid-based and adaptive approaches. All methods operate purely in the spatial domain $\mathbf{x}$ and are applied independently on fixed time, tilt, and azimuth slices (t, θ, ϕ), avoiding coupling across dimensions. Empirically selected hyperparameter-combinations are tested for the K-Nearest Neighbors (KNN), IDW and Moving Least Squares (MLS) methods using an exhaustive search. We denote hyperparameter combinations as $C1$–$C6$ in the order of evaluation (see Tab. I). In the following sections, we denote the distance-based weighting function with $w(\cdot)$ and the local neighborhood of observations as $\mathcal{N}(\mathbf{x})$.

A. Reconstruction methods

This section describes the interpolation methods and evaluation procedure used in this study. The workflow consists of generating missing-data scenarios, applying spatial interpolation methods, and evaluating reconstruction accuracy against ground truth observations.

Analytical Grid-Based Interpolation: Analytical grid-based interpolation methods reconstruct missing values using deterministic schemes defined on a structured spatial grid. In this work, we consider nearest-neighbor interpolation (assigning the value of the closest grid point), and linear interpolation (bilinear interpolation on the grid). Higher-order methods such as cubic, spline, polynomial, and Akima interpolation are not considered, as they tend to be unstable and to over-smooth the field when large data gaps are present [19]. These methods assume a quasi-regular latitude–longitude grid and exploit its topology to achieve high computational efficiency and full reproducibility, making them well suited to gridded products such as PVGIS. The low-order schemes considered here are particularly robust for large missing regions and provide fast, deterministic baselines for comparison [19].

K-Nearest Neighbors: KNN estimates the value at a missing location $\mathbf{x}$ using the k spatially closest grid points. The interpolated value at a missing location $\mathbf{x}$ is computed as:

$$\hat{z}(\mathbf{x}) = \frac{\sum_{i \in \mathcal{N}_K(\mathbf{x})} w(d(\mathbf{x}, \mathbf{x}_i))\; z(\mathbf{x}_i)}{\sum_{i \in \mathcal{N}_K(\mathbf{x})} w(d(\mathbf{x}, \mathbf{x}_i))}, \quad w(d) = \frac{1}{(d + \varepsilon)^{\alpha}}, \tag{3}$$

where $\mathcal{N}_K(\mathbf{x})$ denotes the set of K nearest neighbors. $d(\cdot,\cdot)$ is the Euclidean distance, α controls distance decay, and ε is a small stabilizer [20]. In our implementation, neighbor search is performed once per missing-point mask using a KD-tree built on the valid coordinate set. For each missing location, we cache the indices and distances of its K nearest neighbors. Distances are computed using Euclidean distance. Three neighborhood sizes were tested with $K \in \{2, 3, \cdots, 7\}$.

Inverse Distance Weighting: The interpolated value at location $\mathbf{x}$ is computed as a distance-weighted average of nearby observations:

$$\hat{z}(\mathbf{x}) = \frac{\sum_{i \in \mathcal{N}_M(\mathbf{x})} z(\mathbf{x}_i)\, d(\mathbf{x}, \mathbf{x}_i)^{-s}}{\sum_{i \in \mathcal{N}_M(\mathbf{x})} d(\mathbf{x}, \mathbf{x}_i)^{-s}}, \tag{4}$$

where $s > 0$ controls the rate of distance decay, with larger values of s increasing the influence of closer observations. IDW takes spatial proximity into account, is simple to implement, and is computationally efficient, making it a widely used baseline in spatial interpolation [21]. The method does not impose global smoothness assumptions, thus may produce

TABLE I
HYPERPARAMETER COMBINATIONS FOR EACH METHOD, EVALUATED VIA EXHAUSTIVE SEARCH. PARAMETER VALUES WERE SELECTED EMPIRICALLY.

Method	$C1$	$C2$	$C3$	$C4$	$C5$	$C6$
KNN (K)	2	3	4	5	6	7
IDW (M, s)	2,2	2,3	3,2	3,3	4,2	4,3
MLS (L, w)	30,gauss	30,inv	50,gauss	50,inv	80,gauss	80,inv
Linear / Nearest	–	–	–	–	–	–

locally varying surfaces depending on point density. Neighbor search is restricted to a fixed number of nearest valid points ($M \in \{2, 3, 4\}$) and distance-decay parameters $s \in \{2, 3\}$.

Moving Least Squares: Moving Least Squares interpolation estimates the value at a target location by fitting a local polynomial model to nearby observations using a weighted least-squares criterion. For a target point $\mathbf{x}$, the interpolated value is defined as:

$$\hat{z}(\mathbf{x}) = \mathbf{p}(\mathbf{x})^\top \boldsymbol{\beta}(\mathbf{x}), \tag{5}$$

where $\mathbf{p}(\mathbf{x})$ is the polynomial basis and $\boldsymbol{\beta}(\mathbf{x})$ are the locally fitted coefficients obtained by minimizing:

$$\boldsymbol{\beta}(\mathbf{x}) = \arg\min_{\boldsymbol{\beta}} \sum_{i \in \mathcal{N}_L(\mathbf{x})} w(d(\mathbf{x}, \mathbf{x}_i); \sigma) \left[z(\mathbf{x}_i) - \mathbf{p}(\mathbf{x}_i)^\top \boldsymbol{\beta} \right]^2. \tag{6}$$

where $w(\cdot)$ is a distance-based weighting function controlling the influence of neighboring observations. MLS provides smooth reconstruction that preserves local trends. Four hyperparameter configurations were considered, all using a first-order local polynomial, but varying the number of neighbors and the weighting function. Specifically, configurations included 30, 50, and 80 neighbors (L), inverse-distance or Gaussian weighting kernels (abbreviated as inv and gauss, respectively), and a fixed smoothing parameter $\sigma = 0.5$.

Although KNN, IDW, and MLS all rely on local neighborhoods, they differ in the way spatial information is used. KNN defines locality by neighbor rank and performs local averaging without an explicit distance scale. IDW extends this concept by introducing a continuous distance-decay model, assigning lower influence to more distant observations. MLS departs from averaging altogether by fitting a local polynomial model, enabling the reconstruction of smooth spatial trends. In contrast, analytical grid-based methods do not operate on point geometry but exploit the fixed grid topology, producing deterministic results.

B. Evaluation metrics

Interpolation performance is evaluated using standard pointwise error metrics computed exclusively at spatial locations that were removed, using the original observations as ground truth. For each experimental configuration, errors are averaged across all tilt angles, azimuths, daylight time steps, and removed spatial locations. All metrics are computed consistently across methods using identical missing-data masks.

Mean Absolute Error (MAE): MAE measures the average magnitude of reconstruction errors, treating all deviations equally providing typical absolute error.

$$\mathrm{MAE} = \frac{1}{N} \sum_{\mathbf{q} \in \mathcal{Q}} \left| \hat{z}(\mathbf{q}) - z(\mathbf{q}) \right|. \tag{7}$$

Root Mean Squared Error (RMSE): RMSE penalizes large errors more strongly than MAE and is sensitive to outliers.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{\mathbf{q} \in \mathcal{Q}} \left(\hat{z}(\mathbf{q}) - z(\mathbf{q}) \right)^2}. \tag{8}$$

Mean Bias Error (MBE): MBE quantifies systematic overestimation (positive values) or underestimation (negative values).

$$\mathrm{MBE} = \frac{1}{N} \sum_{\mathbf{q} \in \mathcal{Q}} \left(\hat{z}(\mathbf{q}) - z(\mathbf{q}) \right). \tag{9}$$

Coefficient of Determination (R^2): The coefficient of determination measures the proportion of variance in the observed data explained by the reconstruction, with values in $(-\infty, 1]$, $R^2 = 1$ indicating perfect agreement.

$$R^2 = 1 - \frac{\sum_{\mathbf{q} \in \mathcal{Q}} \left(\hat{z}(\mathbf{q}) - z(\mathbf{q}) \right)^2}{\sum_{\mathbf{q} \in \mathcal{Q}} \left(z(\mathbf{q}) - \bar{z} \right)^2}, \tag{10}$$

where $\bar{z} = \frac{1}{N} \sum_{\mathbf{q} \in \mathcal{Q}} z(\mathbf{q})$ denotes the mean observed irradiance over all evaluated samples.

Mean Absolute Percentage Error (MAPE): MAPE measures the relative reconstruction error as a percentage of the observed values:

$$\mathrm{MAPE} = \frac{100}{N} \sum_{\mathbf{q} \in \mathcal{Q}} \left| \frac{\hat{z}(\mathbf{q}) - z(\mathbf{q})}{z(\mathbf{q})} \right|. \tag{11}$$

To avoid numerical instability and inflated errors at low irradiance values, MAPE is computed only for samples where $z(\mathbf{q}) > 10\,\mathrm{W/m^2}$.

IV. RESULTS AND DISCUSSION

Experimental results are summarized in Fig. 3 and Tab. II. Error bars reflect variability across spatial locations, time steps, and orientations. For each scenario, a single deterministic missing-data realization is used, enabling controlled comparison between methods but not capturing variability across different random masks. This limitation will be addressed in future work through multiple realizations and statistical analysis.

Across all scenarios (M1)–(M3), KNN achieves the lowest errors, with MAE increasing from $47\,\mathrm{W/m^2}$ to $63\,\mathrm{W/m^2}$ and MAPE around $50\,\%$. All metrics show consistent trends. Other methods (MLS, IDW, linear, nearest) exhibit similar but higher errors (MAE $95\,\mathrm{W/m^2}$–$110\,\mathrm{W/m^2}$) with stable ranking. Performance depends on the missing-data scenario (see Fig. 4). For random removal (M1), error remains stable as removal increases, but the optimal KNN configuration shifts from

TABLE II
MAE ($\mathrm{W/m^2}$) BY INTERPOLATION METHOD, MISSING-DATA GENERATION STRATEGIES, AND REMOVED-POINT FRACTION.

Model	(M1) Random removal			(M2) Contiguous gaps									(M3) Regular-grid removal		
				$k=2$			$k=3$			$k=4$					
	5 %	15 %	25 %	5 %	15 %	25 %	5 %	15 %	25 %	5 %	15 %	25 %	5 %	15 %	25 %
KNN	38.35	42.20	40.90	60.63	62.68	75.51	47.77	68.93	71.62	40.65	61.02	74.95	48.65	51.77	51.85
MLS	96.67	96.89	96.70	98.84	100.51	129.91	94.86	99.77	106.95	97.37	106.10	106.66	96.81	97.50	97.99
Linear	102.70	102.57	102.92	95.21	102.54	98.85	101.90	97.95	97.71	103.47	100.25	99.35	100.64	101.07	101.48
IDW	105.02	105.32	105.67	110.76	117.41	117.77	109.10	114.77	116.65	111.35	112.91	114.80	105.79	105.50	106.43
Nearest	106.89	107.10	107.54	102.31	108.94	106.33	106.90	105.49	105.74	108.59	106.70	106.32	104.54	105.50	105.85

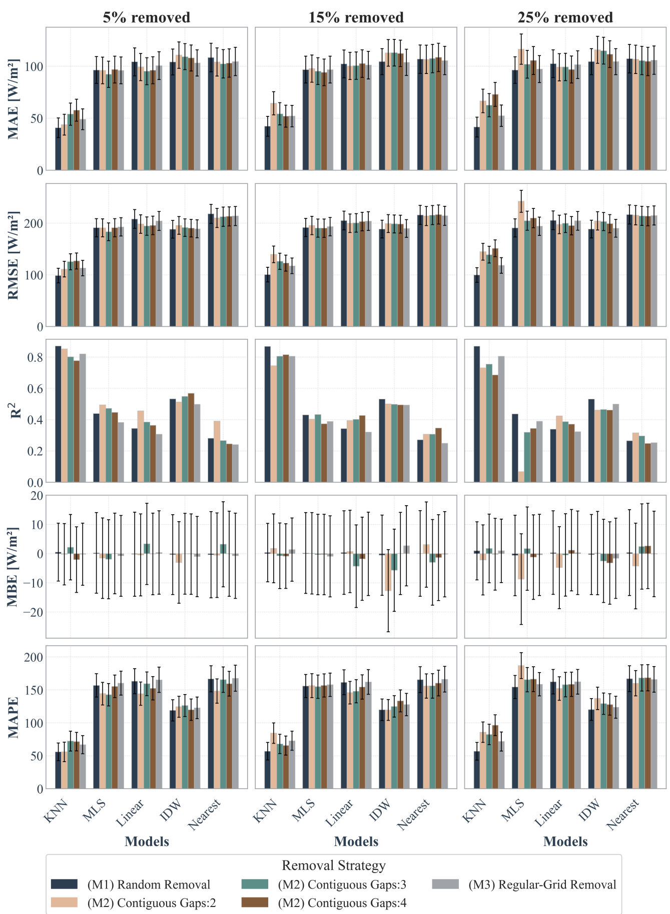

Fig. 3. Reconstruction error under different missing-data fractions and missing-data generation strategies. Columns correspond to pointwise MAE, RMSE, MBE, R^2, and MAPE. The bars report mean values across hyperparameter configurations with error bars indicating variability across location, azimuth and tilt.

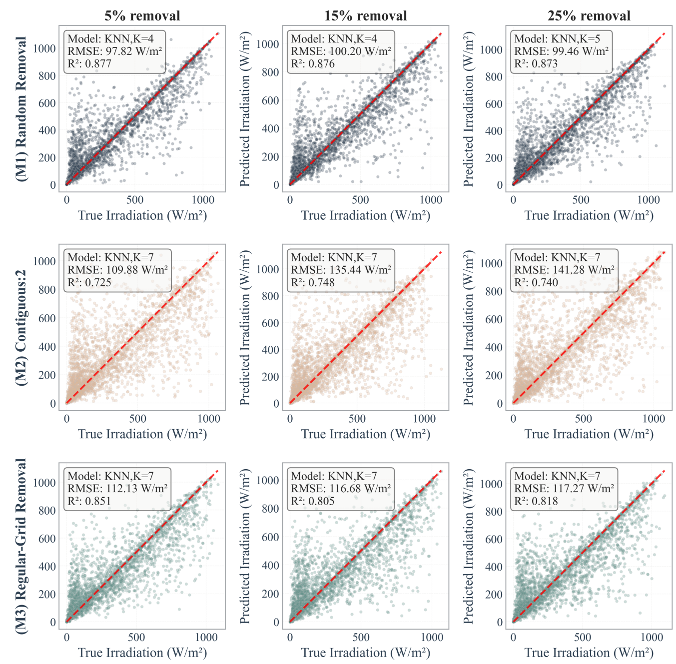

Fig. 4. Regression analysis results for selected removal types and best performing models for a subset of 5000 points

smaller to larger neighborhoods ($K = 4 \rightarrow 7$). For regular-grid removal (M3), errors increase gradually. Contiguous-gap removal (M2) leads to the strongest degradation, consistently favoring larger neighborhoods ($K = 7$). These results indicate strong spatial correlation in irradiance data: reconstruction remains reliable when nearby observations are available, but degrades as the distance to known points increases. Runtime of the methods was measured on a standard workstation (Intel i7-6920HQ, 48 GB RAM) for a representative scenario (5% removal, 2 contiguous gaps). KNN is both accurate and computationally efficient (10.2 s), comparable to IDW (11.4 s) and nearest-neighbor (11.1 s). In contrast, MLS (46.8 s) and linear interpolation (47.3 s) are significantly slower in the current implementation, despite their analytical formulation. From an application perspective, KNN is recommended for all missing data scenarios, with larger neighborhoods improving performance in for contiguous and regular-grid gaps. Linear and nearest-neighbor methods are fast baselines with larger errors, while MLS and IDW enable smoother reconstructions but are more sensitive to hyperparameters (see Fig. 5). For IDW, larger weighting exponents improve accuracy, while increasing the MLS neighborhood size reduces error.

V. Conclusion

This paper presents a comparison of widely used spatial interpolation methods for reconstructing missing solar irradiance data on quasi-regular latitude–longitude grids using

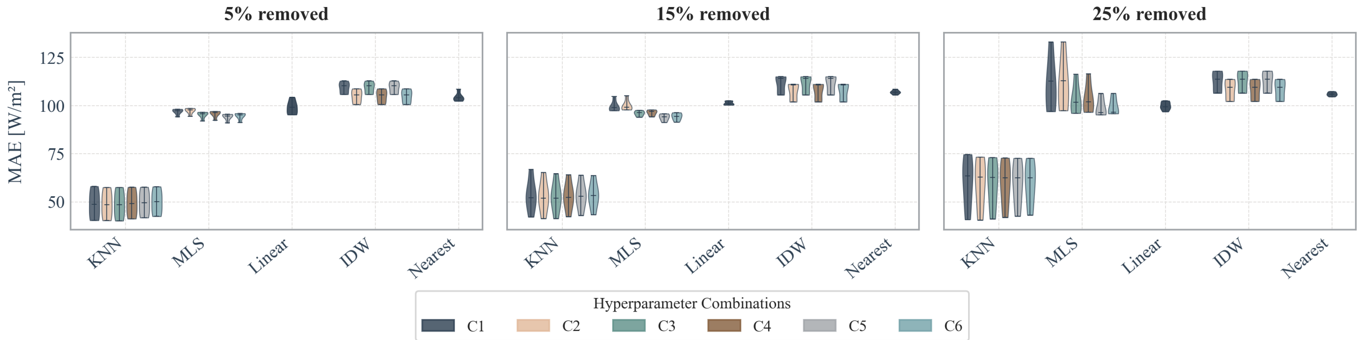

Fig. 5. Distributions of MAE across hyperparameter combinations ($C1 - C6$) for different interpolation methods and missing-point fractions (see Tab. I). Error distributions are shown for each method and configuration.

PVGIS-derived datasets. Interpolation performance was evaluated under controlled missing-data scenarios. Three missing-data generation strategies—random removal, spatially contiguous gaps, and regular-grid thinning—were considered with removal fractions between $5\,\%$ and $25\,\%$, and performance was assessed using standard error metrics. According to the findings, KNN consistently achieved the most accurate reconstructions, maintaining superior performance as data sparsity increased. The geometry of missing data significantly influence interpolation accuracy, however the fraction has a smaller impact on it. This study is limited to spatial location removal and a set of classical interpolation methods. Future work will extend the framework to orientation-dependent gaps, advanced approaches (e.g., geostatistical and machine learning methods), and multiple missing-data realizations to enable statistical evaluation (mean and standard deviation).

References

[1] D. Jeong, A. St-Hilaire, Y. Gratton, C. Bélanger, and C. Saad, "A guideline to select an estimation model of daily global solar radiation between geostatistical interpolation and stochastic simulation approaches," *Renewable Energy*, vol. 103, pp. 70–80, 2017.

[2] M. S. Pessanha, L. M. dos Santos, G. Lyra, A. O. Lima, G. B. Lyra, and J. L. de Souza, "Interpolation methods applied to the spatialisation of monthly solar irradiation in a region of complex terrain in the state of rio de janeiro in the southeast of brazil," *Modeling Earth Systems and Environment*, pp. 1–14, 2020.

[3] A. Dobos, J. S. Stein, D. Riley, and M. A. Mikofski, "Evaluation of Time-Series Gap-Filling Methods for Solar Irradiance Applications," National Renewable Energy Laboratory (NREL), Tech. Rep. NREL/TP-5D00-80730, 2021, technical Report. [Online]. Available: https://www.nrel.gov/docs/fy21osti/80730.pdf

[4] C. Crespo-Turrado, M. Meizoso-López, F. Lasheras, B. A. Rodríguez-Gómez, J. Calvo-Rolle, and F. J. D. C. Juez, "Missing data imputation of solar radiation data under different atmospheric conditions," *Sensors*, vol. 14, pp. 20 382–20 399, 2014.

[5] C. M. Machado, M. C. Brito, T. Santos, and J. A. Tenedório, "GIS-Based Methods and Tools for the Analysis of Solar Potential: A Review," *Applied Sciences*, vol. 9, no. 9, p. 1960, 2019.

[6] J. Ruiz-Arias, "Mean-preserving interpolation with splines for solar radiation modeling," *Solar Energy*, 2022.

[7] B. Jiao, Y. Su, Q. Li, V. Manara, and M. Wild, "An integrated and homogenized global surface solar radiation dataset and its reconstruction based on a convolutional neural network approach," *Earth System Science Data*, 2023.

[8] L. Benavides Cesar, M.-A. Manso-Callejo, and C.-I. Cira, "Bert for missing data imputation in solar irradiance time series," in *ITISE 2023*, 2023.

[9] R. Girimurugan, P. Selvaraju, P. Jeevanandam, M. Vadivukarassi, S. Subhashini, N. Selvam, S. H. Ahammad, S. Mayakannan, and S. K. Vaithilingam, "Application of deep learning to the prediction of solar irradiance through missing data," *International Journal of Photoenergy*, 2023.

[10] J. G. G. Pedreros, S. Lagüela-Lopez, and M. Rodríguez Martín, "Spatial models of solar and terrestrial radiation budgets and machine learning: A review," *Remote Sensing*, vol. 16, p. 2883, 2024.

[11] Joint Research Center, "Photovoltaic Geographical Information System," https://re.jrc.ec.europa.eu/pvg_tools/en/#MR.

[12] P. Lazzeroni, F. Moretti, and F. Stirano, "Economic potential of PV for Italian residential end-users," *Energy*, vol. 200, p. 117508, 2020.

[13] O. Gounari, A. Martinez, N. Taylor, and N. Alexandris, "PVGIS 5.3 Dataset Update," European Commission, Joint Research Centre, Ispra, Italy, Tech. Rep. JRC139355, 2024.

[14] Joint Research Center, "Photovoltaic Geographical Information System - API non-interactive service," https://joint-research-centre.ec.europa.eu/photovoltaic-geographical-information-system-pvgis/getting-started-pvgis/api-non-interactive-service_en.

[15] International Organization for Standardization, "Hygrothermal performance of buildings - calculation and presentation of climatic data - part 4: Hourly data for assessing the annual energy use for heating and cooling," ISO Standard 15927-4:2005, 2005.

[16] A. Eschenbach, G. Yepes, C. Tenllado, J. I. Gomez-Perez, L. Piñuel, L. Zarzalejo, and S. Wilbert, "Spatio-temporal resolution of irradiance samples in machine learning approaches for irradiance forecasting," *IEEE Access*, vol. 8, pp. 51 518–51 531, 2020.

[17] C. Crespo-Turrado, M. Meizoso-López, F. Lasheras, B. A. Rodríguez-Gómez, J. Calvo-Rolle, and F. J. D. C. Juez, "Missing data imputation of solar radiation data under different atmospheric conditions," *Sensors (Basel, Switzerland)*, vol. 14, pp. 20 382–20 399, 2014.

[18] A. Livera, M. Theristis, E. Koumpli, S. Theocharides, G. Makrides, J. Sutterlueti, J. Stein, and G. Georghiou, "Data processing and quality verification for improved photovoltaic performance and reliability analytics," *Progress in Photovoltaics: Research and Applications*, vol. 29, pp. 143–158, 2020.

[19] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes: The Art of Scientific Computing*, 3rd ed. Cambridge University Press, 2007.

[20] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[21] D. Shepard, "A two-dimensional interpolation function for irregularly-spaced data," *Proceedings of the 1968 ACM National Conference*, pp. 517–524, 1968.