Energy Conversion and Management 339 (2025) 119857

Contents lists available at ScienceDirect

Energy Conversion and Management

journal homepage: www.elsevier.com/locate/enconman

Research paper

Data imputation methods for intermittent renewable energy sources: Implications for energy system modeling

Claudio Mantuano [a,c], Olalekan Omoyele [a,b,*], Maximilian Hoffmann [a], Jann Michael Weinand [a], Massimo Panella [d], Detlef Stolten [a,b]

[a] Forschungszentrum Jülich GmbH, Institute of Climate and Energy Systems – Jülich Systems Analysis (ICE-2), 52425 Jülich, Germany
[b] RWTH Aachen University, Chair for Fuel Cells, Faculty of Mechanical Engineering, 52062 Aachen, Germany
[c] Sapienza University of Rome, Department of Computer, Control, and Management Engineering (DIAG), 00185 Rome, Italy
[d] Sapienza University of Rome, Department of Information Engineering, Electronics, and Telecommunications (DIET), 00184 Rome, Italy

ARTICLE INFO

Keywords:
Self-sufficiency
Data imputation
Machine learning
Energy time series
Renewable energy systems
Energy system optimization

ABSTRACT

To incorporate a high share of intermittent renewable sources in energy systems, energy system optimization models rely on weather and climate time series data. However, data for renewable energy sources often contains missing values due to sensor or transmission faults. This study evaluates various data imputation methods for minutely-resolved global horizontal irradiance, direct normal irradiance, and wind speed time series, with missingness ranging from two to ninety percent. Alongside standard statistical tests, a novel validation criterion is introduced by directly evaluating the impact of imputation methods on energy system modeling. While certain imputation methods demonstrate strong point-wise statistical accuracy, they do not necessarily preserve the underlying data distribution. The performance of these methods is strongly influenced by the type of time series and the missingness mechanism, either continuous gaps or randomly missing data points. In energy system optimization, multiple imputation by chained equations, *k*-nearest neighbors, linear interpolation, and simple moving average yield the best results, outperforming more sophisticated deep learning-based methods. Overall, *k*-nearest neighbors consistently outperformed the other approaches across all validation criteria. By comprehensively evaluating the statistical performance of imputation methods and their impact on energy system modeling, this study offers valuable insights for researchers and practitioners addressing missing data in energy system applications.

1. Introduction

The rise in energy demand and climate change threats has increased the need for renewable energy sources. This drives the system-wide integration of intermittent renewable sources into the electricity grid [1–3]. Together with hydro energy, solar and wind energy are the most used renewable sources by total feed-in, relying on solar irradiance and wind speed, respectively. Therefore, accurate and reliable data on global horizontal irradiance (GHI), direct normal irradiance (DNI), and wind speed are crucial for the planning, design, and operation of renewable energy systems [4]. However, missing data can be a common issue in such datasets, which can be caused by various factors such as equipment failure, data transmission errors, and environmental conditions [5,6]. These missing data have a substantial impact on the accuracy and reliability of energy system models, which can result in sub-optimal decision-making and potential financial losses [7–9]. Therefore, methods for data imputation have been extensively employed to fill in the gaps in datasets for wind speed and solar irradiance. These data imputation techniques help predict the missing values based on the information from the available data, thereby improving the accuracy and reliability of energy system modeling outcomes [8].

Several studies have considered data imputation methods in energy time series, as summarized in Fig. 1 and Appendix A. Among these methods, classical imputation techniques based on the mean, mode, median, interpolation, or moving average are widely used and are often employed as benchmark methods, as pointed out by Lin and Tsai [10]. In addition to these, variant methods were proposed, such as the combination of linear interpolation (LI) and linear regression by Sánchez et al. [11]. Multiple imputation by chained equations (MICE) - although not extensively used in the energy field (see Fig. 1) - is a common technique in data imputation [8,12–17]. Machine learning plays a vital role in the development of more advanced methods

* Corresponding author at: Forschungszentrum Jülich GmbH, Institute of Climate and Energy Systems – Jülich Systems Analysis (ICE-2), 52425 Jülich, Germany.
E-mail address: o.omoyele@fz-juelich.de (O. Omoyele).

https://doi.org/10.1016/j.enconman.2025.119857
Received 11 February 2025; Received in revised form 24 April 2025; Accepted 24 April 2025
Available online 22 May 2025

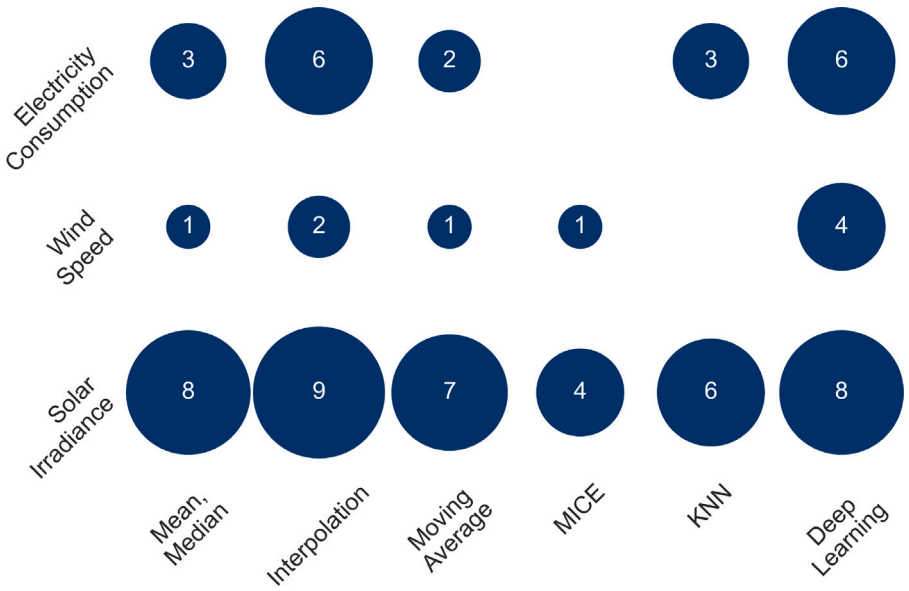

Fig. 1. Frequency of use in previous studies of various imputation techniques on time series of electricity consumption, wind speed, and solar irradiance.

of data imputation, such as *k*-nearest neighbors (KNN), generative adversarial networks (GAN), and many other techniques. KNN is the most widely used machine learning-based imputation method [10], and is often considered a benchmark [8,12,13,18–25]. Other machine learning techniques include linear regression [11,12,15,26,27], support vector machines [11,24,25,28], support vector regressors [21,26,29], and random forests [30]. In recent years, there has been a growing interest in neural network-based methods of data imputation. Kim et al. [31] used artificial neural networks (ANN) and random trees for rainfall data imputation. Garnier et al. [32] employed a feedforward ANN to solar radiation and indoor temperature data. Shukur et al. [33] proposed an autoregressive ANN for wind speed data imputation, while Rahman et al. [34] applied a deep recurrent neural network to electricity consumption data. A promising approach based on GAN was introduced by Goodfellow et al. [35]. GAN were primarily used for image generation but later adapted to imputation tasks with the generative adversarial imputation nets (GAIN) by Yoon et al. [36]. Variants of GAN developed for applications in the energy field include solarGAN by Zhang et al. [13] and the GAN by Khare et al. [23], both specifically designed for solar data imputation. Qu et al. [29] proposed ipGAN for wind speed data imputation. Further recent methods include multivariate time series imputation by Bülte et al. [18] and convolutional denoising autoencoders (CAE) by Liguori et al. [19,37], in addition to ANN combined with encoder–decoder proposed by Centeno et al. [20], extreme gradient boosting by differential evolution of Başakin et al. [38], and convolutional neural networks combined with long short-term memory networks proposed by Hussain et al. [39].

Whether deep learning-based imputation methods can consistently outperform conventional approaches remains an open question, as prior studies have highlighted the limitations of more advanced techniques [16,17]. The outcomes of such comparisons are highly dependent on factors such as the data type, missingness mechanism, missing rates, and evaluation criteria. To date, no study has comprehensively assessed a broad range of imputation techniques for energy time series under different missingness scenarios, nor explicitly evaluated their impact on energy system modeling, an important gap that the present work aims to address.

In this study, we employ both conventional and more advanced imputation methods: mean, median, interpolations, moving averages, MICE, KNN, GAIN, and CAE. These methods are applied to impute high-resolution GHI, DNI, and wind speed time series. The resolution is defined as the time interval between two consecutive measurements. Consequently, high-resolution time series are characterized by smaller time intervals. Missing data are generated under two missingness mechanisms, namely continuous gaps or randomly missing data points, with missing rates ranging from two to ninety percent. The resulting synthetic data are compared to the original ones to evaluate the methods' effectiveness. The accuracy of data obtained from different techniques is evaluated using statistical metrics and, for GHI data, energy system modeling. The statistical tests used to validate the methods include the root mean square error (RMSE) and the Kolmogorov–Smirnov (KS) test. While the statistical approaches compare the original and the imputed data based on the respective time series profile, the energy system modeling quantifies the imputation techniques' effectiveness by evaluating their accuracy when synthetic data are used to optimize a self-sufficient building energy system. The optimization problem is solved using both original and synthetic data, and the resulting outcomes - total system costs and installed capacity of components - are compared to measure the deviation between the values obtained from original data and those from the synthetic data.

This research offers a comprehensive assessment of imputation methods across diverse data types and missingness scenarios, including their impact on energy system optimization. By explicitly quantifying how different imputation techniques affect optimization outcomes and comparing these results with statistical metrics, this study provides a detailed evaluation of approaches for handling missing data in time series for energy system applications.

The remainder of the work is structured as follows: Section 2 outlines our methodology, including the imputation methods, the data used in the experiments, and the selected validation criteria. In Section 3 we discuss the results, while Section 4 presents our conclusions and suggests potential directions for future research.

2. Methodology

In the following sections, the applied imputation methods are described (see Section 2.1), along with the data used in the experiments (see Section 2.2), and the validation methods, including the energy system model (see Section 2.3).

2.1. Applied imputation methods

In this study, we employ both conventional and more advanced imputation methods based on machine learning and deep learning, which are described in the following.

The **mean imputation** replaces any missing data point Y_t with the mean value of available data, as in Eq. (1) [40].

$$Y_t = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{1}$$

where n and Y_i are the total number of observations and the observation at time i, respectively.

The **median imputation** replaces any missing data point Y_t with the median value of available data, as in Eq. (2) [40].

$$Y_t = \begin{cases} Y_{\frac{n+1}{2}} & \text{if } n \text{ is odd} \\ \frac{Y_{\frac{n}{2}} + Y_{\frac{n}{2}+1}}{2} & \text{if } n \text{ is even} \end{cases} \tag{2}$$

where n is the total number of observations.

The **linear interpolation** (LI) assumes a linear relationship between pairs of consecutive available data points, as in Eq. (3) [41,42].

$$Y_t = Y_{t_1} + \frac{Y_{t_2} - Y_{t_1}}{t_2 - t_1}(t - t_1) \tag{3}$$

where t_1 and Y_{t_1} are the first coordinates, t_2 and Y_{t_2} are the second coordinates, t is the missing data point to be interpolated, and Y_t is the interpolated value.

The **cubic interpolation** (CI) employs polynomial curves of degree three to interpolate between pairs of consecutive available data points, as in Eq. (4) [41,42].

$$Y_t = c_1 t^3 + c_2 t^2 + c_3 t + c_4 \tag{4}$$

Given four data points $\{t_0, Y_{t_0}\}$, $\{t_1, Y_{t_1}\}$, $\{t_2, Y_{t_2}\}$, and $\{t_3, Y_{t_3}\}$, the coefficients c_1, c_2, c_3, and c_4 are obtained. Due to its non-linearity, this method can produce outliers - both positive and negative - when there are large gaps in the data.

The **simple moving average** (SMA) computes the average of the available data points in a specified time window, as in Eq. (5) [43,44].

$$Y_t = \frac{1}{k}\sum_{i=t-k}^{t-1} Y_i \tag{5}$$

where k is the size of the time window and Y_i is the observation at time i, with i ranging from $t-k$ to $t-1$. The SMA helps reduce the impact of noise and can emphasize long-term patterns as k increases.

The **exponentially weighted moving average** (EWMA) calculates a moving average by assigning varying weights to observations over time, giving more weight to recent observations and less to older ones [43]. The weights are determined by a parameter β, which ranges between 0 and 1. The values Y_i in Eq. (6) correspond to the observations being summed over time, with i ranging from 1 to $t-1$.

$$Y_t = \sum_{i=1}^{t-1} \beta\,(1-\beta)^{t-i-1}\, Y_i \tag{6}$$

The **autoregressive integrated moving average** (ARIMA) for data imputation [45] consists of three key components: an autoregressive (AR), an integrated (I), and a moving average (MA) term. Together, these components determine the model order (p, d, q), as shown in Eq. (7) [46]. The term I ensures the time series is stationary by replacing its values with differenced values of order d, while the AR and MA terms incorporate the lagged p data points and the lagged q errors, respectively. The ARIMA (p, d, q) model predicts the d^{th}-order differenced Y_t using the α and θ coefficients, which are estimated from the time series data.

$$Y'_t = I + \alpha_1 Y'_{t-1} + \cdots + \alpha_p Y'_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \tag{7}$$

The **multiple imputation by chained equations** (MICE) employs variable correlations to estimate missing data using statistical models such as linear regression [47]. Initially, the missing values are imputed randomly or replaced with the mean value of available observations. Through iterative steps, the missing values are then imputed by leveraging correlations between variables until convergence is reached, resulting in a final dataset where all missing gaps are imputed.

The ***k*-nearest neighbors** (KNN) method is employed to impute missing data points by estimating them through their proximity to available observations, typically measured using the Euclidean distance [48]. After computing the distance between the missing data point and all available observations, the KNN formula in Eq. (8) is used to calculate the missing value Y_t by averaging the k-nearest available observations Y_i, with i ranging from 1 to k.

$$Y_t = \frac{1}{k}\sum_{i=1}^{k} Y_i \tag{8}$$

Goodfellow et al. [35] introduced the generative adversarial networks (GAN), a machine learning model in which two neural networks - the generator and the discriminator - engage in a minimax game. The generator network aims to generate data that mimics the distribution of a given dataset, while the discriminator network evaluates the authenticity of the generated data against the real data. Through iterative training, the generator refines its ability to generate synthetic data that is increasingly indistinguishable from real data. The first adaptation of GAN for the task of missing data imputation was the **generative adversarial imputation nets** (GAIN) by Yoon et al. [36]. Numerous GAIN-based techniques for imputing either wind or solar time series data have emerged in recent years [13,23,29]. Given the diverse nature of the data used in our experiments - both wind speed and solar irradiance - we apply the original GAIN proposed by Yoon et al. [36].

Convolutional denoising autoencoders (CAE) were used by Liguori et al. [19,37] to impute missing data in electricity consumption time series. This model consists of an encoder–decoder pair designed to handle noisy input data. The encoder processes the noisy data through convolutional layers to extract significant pattern information. Afterwards, the decoder reconstructs the encoded data, removing noise and transforming the data back to its original size. Additionally, Liguori et al. [19] proposed an alternative approach that combines CAE with data augmentation (CAE + Aug) to reconstruct missing gaps in settings with limited data availability. Given the sufficiency of available data, we adopt the base CAE model, as data augmentation is not necessary.

2.2. Data

Three different types of time series are considered: GHI, DNI, and wind speed. The data used for imputation come from Milan (latitude: 45.50, longitude: 9.16) for the year 2019, while the data from 2017 and 2018 are used to train the supervised learning models. The solar irradiance data for GHI and DNI are available in the *National Solar Radiation Database (NSRDB)* [49].

2.2.1. Resolution

The resolution (or sampling rate) of a time series refers to the time interval between consecutive measurements and is crucial for capturing patterns in intermittent renewable energy time series data, thereby leading to more reliable results in energy system modeling [50]. However, as summarized in the table in Appendix A and Fig. 2, the most commonly used resolution in previous studies is one hour, followed by one day and 15 min. Indeed, minutely resolution is rarely employed due to the unavailability of data [51]. This study focuses on minutely time series for more accurate sub-hourly modeling of the energy system, deriving high-resolution data from hourly data. Specifically, solar irradiance hourly time series obtained from NSRDB are

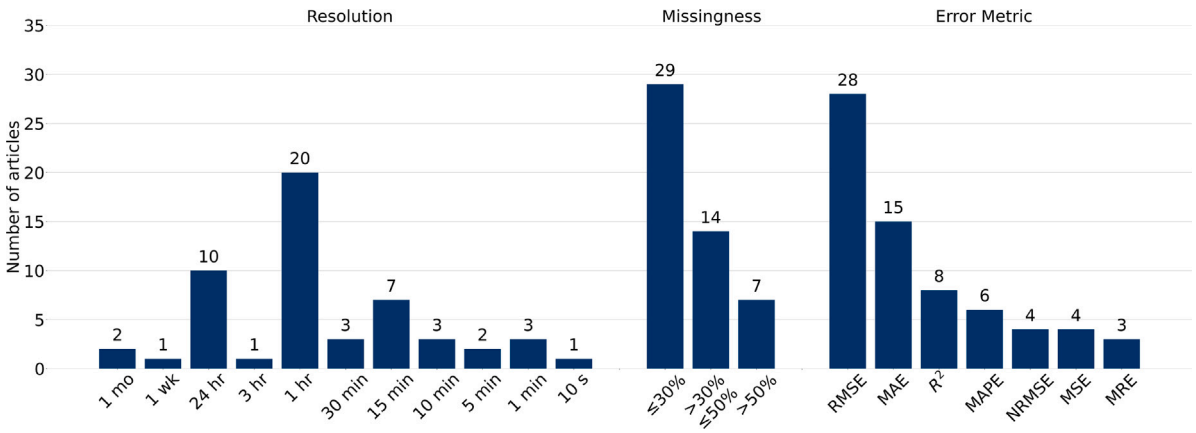

Fig. 2. Frequency of use in previous studies of different resolutions, missing rates, and error metrics.

downscaled to minutely time series using a non-dimensional model developed to generate synthetic data [52,53]. This model applies non-dimensionalization of solar irradiance and time - both stored in a database - to downscale GHI and DNI from hourly to minutely resolution. The hourly data to be downscaled are then parameterized to align with the database using clear-sky irradiance values and their variability [51,54]. The wind speed data are simulated using CorRES [55,56], a time series simulation tool for variable renewable energy. The CorRES tool produces high-resolution renewable energy data from reanalyzed meteorological data and stochastic fluctuations [51].

2.2.2. Missingness

The extent of missing data significantly impacts imputation results because imputation methods rely on the available data to capture patterns and estimate the missing values. While most studies consider missing rates up to 30% (see Fig. 2), this research evaluates a broader range, from 2% to 90%, to assess the performance of imputation methods under varying levels of missingness severity. Two approaches are used to generate missing data, the example of which is illustrated in Fig. 3. The first approach (hereinafter referred to as "continuously missing") randomly generates intervals of 360 to 4320 consecutive missing data points, corresponding to gaps ranging from six hours to three days. This approach mimics a real-world scenario where time is required for the maintenance of measurement equipment or to repair it after a fault. The second approach (hereinafter referred to as "randomly missing") randomly generates single missing data points instead of continuous intervals. The imputation considers both scenarios, enabling a comparative analysis of the statistical validation results. Conversely, the energy system modeling problem only considers the first scenario, as it more closely reflects a real-world case of equipment failure or maintenance.

2.3. Validation

The effectiveness of imputation techniques is evaluated using the RMSE and the KS test as statistical metrics (see Section 2.3.1). Additionally, we introduce a novel validation criterion based on comparing the outcomes of an energy system optimization problem solved using both synthetic and original data (see Section 2.3.2).

2.3.1. Statistical metrics

As reported in Appendix A and Fig. 2, the commonly used validation metrics are RMSE, mean absolute error (MAE), and the coefficient of determination (R^2). Other frequently adopted metrics are the mean absolute percentage error (MAPE), normalized root mean square error (NRMSE), mean square error (MSE), and mean relative error (MRE). Given the comparative nature of our study to previous studies, we employ RMSE to evaluate the point-to-point accuracy of imputation methods. In addition to point-wise performance evaluation, we apply the KS test as a complementary statistical validation criterion. The KS test compares the distributions of the original and imputed data, providing a more comprehensive evaluation of imputation performance [51].

The RMSE, defined in Eq. (9), provides the square root of the mean squared differences between the original data points y_i and the synthetic data points $\hat{y}_i$, with i ranging from 1 to the number of observations, n [57].

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{9}$$

The KS test, defined in Eq. (10), measures the maximum absolute difference D between the synthetic and original distributions [58].

$$D = \max\left|F_n(x) - F(x)\right| \tag{10}$$

where $F_n(x)$ and $F(x)$ denote the empirical cumulative distribution functions of the synthetic and original samples, respectively. The KS test yields the KS statistic, representing the maximum difference between the cumulative distribution functions of the synthetic and original data, and the corresponding p-value, indicating the likelihood that the two samples come from the same population.

2.3.2. Self-sufficient building model

The drive towards net-zero emissions and the declining costs of decentralized off-grid renewable energy systems have intensified the focus on renewable energy deployment, leading to a growing interest in energy autonomy at the residential level [2,60–62]. One proposed approach is the self-sufficient building, which is designed to generate energy from renewable sources to reduce its reliance on energy providers. As a result, there is heightened emphasis on residential energy autonomy, with self-sufficient homes playing a key role in contributing to environmental sustainability [60,63]. In this study, a self-sufficient building model introduced by Kotzur et al. [64], and further explored by Knosala et al. [59] and Omoyele et al. [50] is considered. This model integrates renewable energy sources (solar photovoltaics) and energy storage systems, combined with advanced energy management strategies to optimize resource utilization. The model comprises an electricity grid, a hydrogen grid, and a heat grid, integrated to maximize the utilization of renewable energy and self-sufficiency (see Fig. 4).

Economic aspects of system components. The cost structure of the selected components comprises initial capital expenditures (CAPEX) and fixed operational expenditures (OPEX), covering predictable expenses such as maintenance and scheduled repairs. The techno-economic data of the self-sufficient building, as detailed by Knosala et al. [59] - including both fixed and capacity-specific capital and operational expenditures, along with technology lifetimes - are summarized in the table in Appendix B.

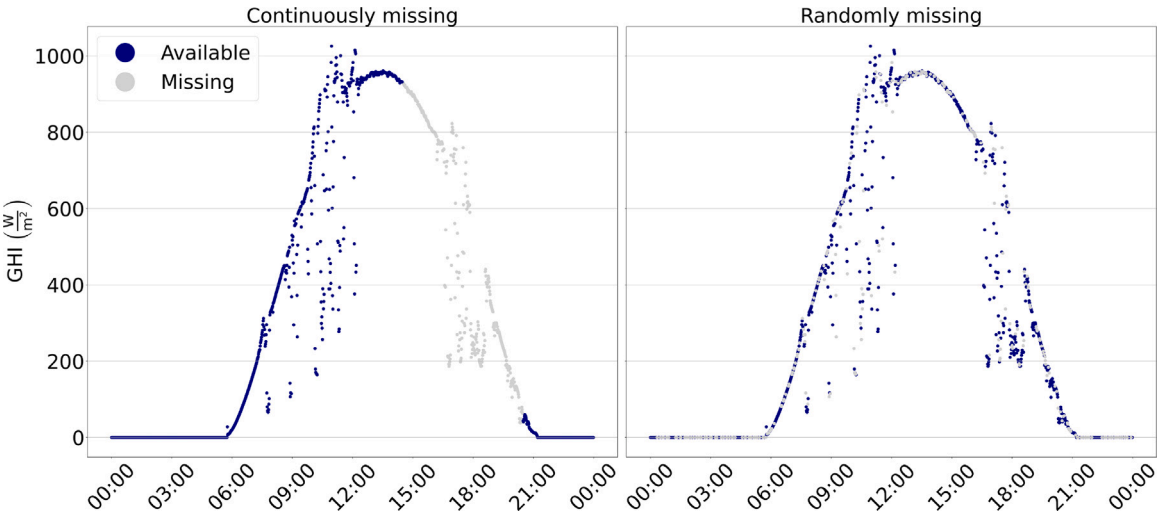

Fig. 3. One-day profile of a GHI minutely time series (Milan, July 7, 2019) with continuously (left) and randomly (right) missing data generation. Available data points are shown in blue, whereas missing data points are shown in gray.

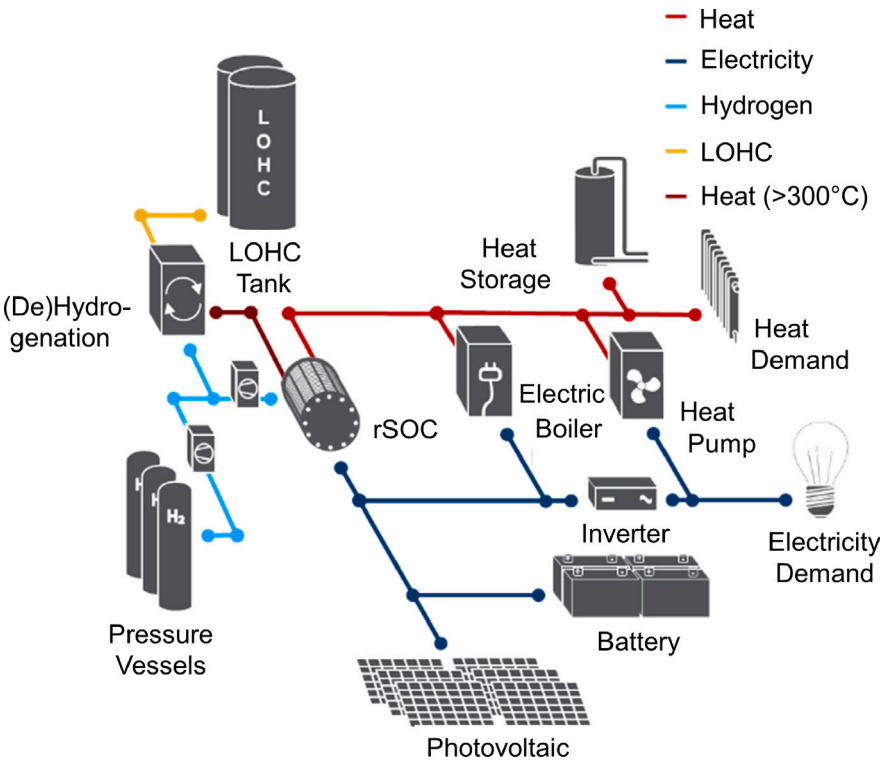

Fig. 4. Self-sufficient building energy system model proposed by Knosala et al. [59]. The abbreviations LOHC and rSOC stand for liquid organic hydrogen carrier and reversible solid oxide cell, respectively.

Energy system optimization problem. The self-sufficient building capacity expansion optimization problem is formulated as a mixed-integer linear programming (MILP) problem. This MILP formulation defines technology selection, component sizing, and operational characteristics. The total annualized costs (TAC) in Eq. (11) are minimized, where the annual economic interest rate, i, is assumed to be 3% over the component lifetime, n.[1]

$$\mathrm{TAC} = \mathrm{CAPEX} \times \left(\frac{i}{1-(1+i)^{-n}} + \mathrm{OPEX}_{\mathrm{rel}} \right) \tag{11}$$

[1] The operational expenditures ($\mathrm{OPEX}_{\mathrm{rel}}$) are considered relative to the capital expenditures (CAPEX).

The complete MILP formulation, as detailed by Refs. [50,65,66], is presented in Appendix C. The energy system optimization problem is modeled using the ETHOS.FINE framework [67,68].

3. Results and discussion

The experiments are conducted on GHI, DNI, and wind speed data under the two different scenarios referred to as continuously missing and randomly missing, as described in Section 2.2.2. The original time series are corrupted by introducing missing data at percentages ranging from 2% to 90%, and are subsequently imputed using the methods described in Section 2.1. The statistical validation results of the 726 experiments are discussed in Section 3.1. Finally, the imputed GHI data

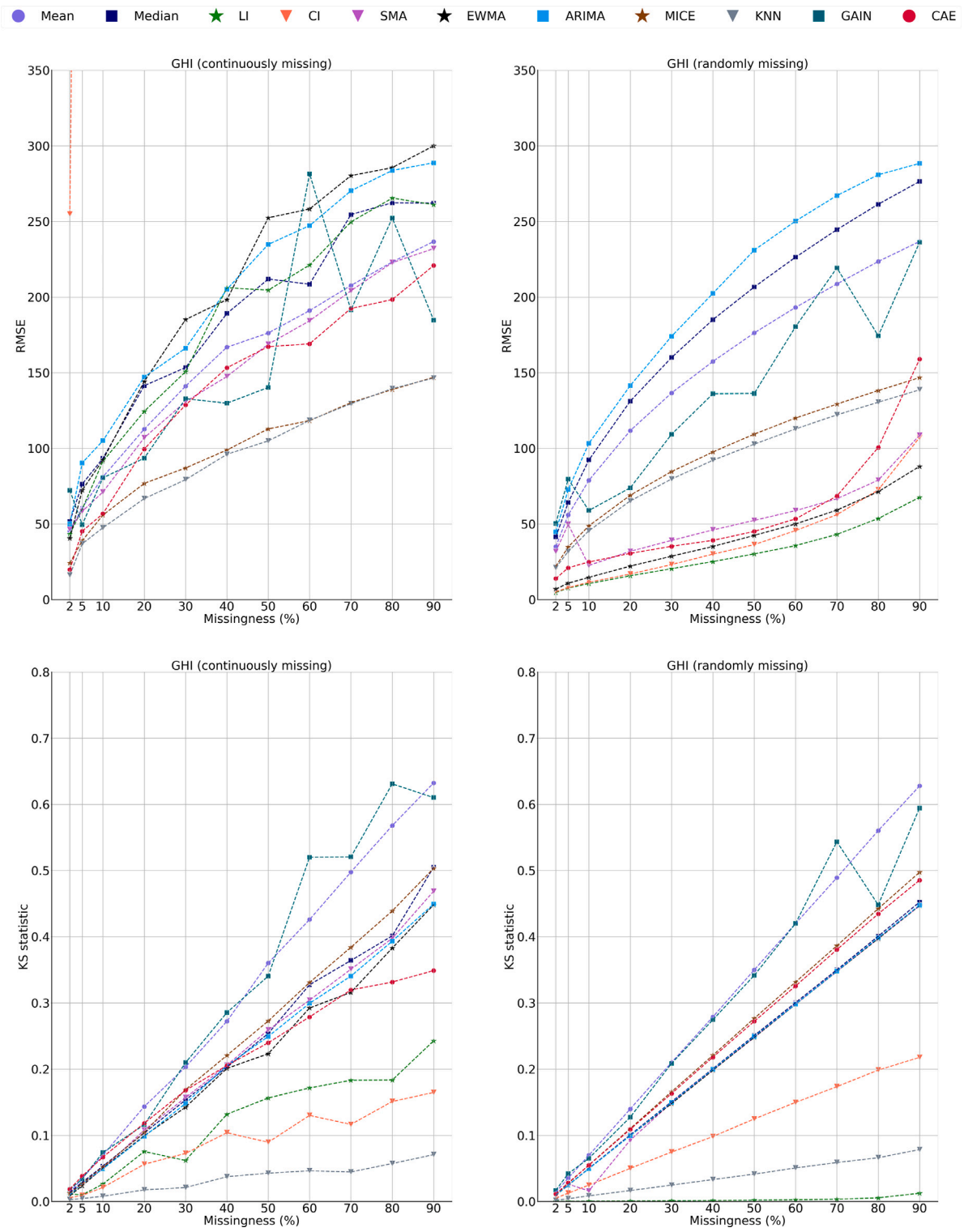

Fig. 5. Statistical validation results of the applied imputation methods for missing rates ranging from 2% to 90% in global horizontal irradiance time series (see also the detailed table of results in Appendix D). Results for the continuously missing scenario are shown on the left, while those for the randomly missing scenario are shown on the right.

from the continuously missing scenario are used in the self-sufficient building capacity expansion optimization problem. To quantify the impact of imputation on energy system modeling, we compute the percentage error between the results - namely TAC and capacities of photovoltaic (PV) modules, inverter, and battery - obtained from the synthetic data and those obtained from the original data, as outlined in Section 3.2.

3.1. Time series imputation

The statistical validation results for the imputation methods applied to GHI, DNI, and wind speed data at different missing rates are shown in Figs. 5–7. The evaluation metrics are computed by comparing the imputed time series to the original ones. Note that the negative values generated, which lack physical interpretation, are replaced with zeros

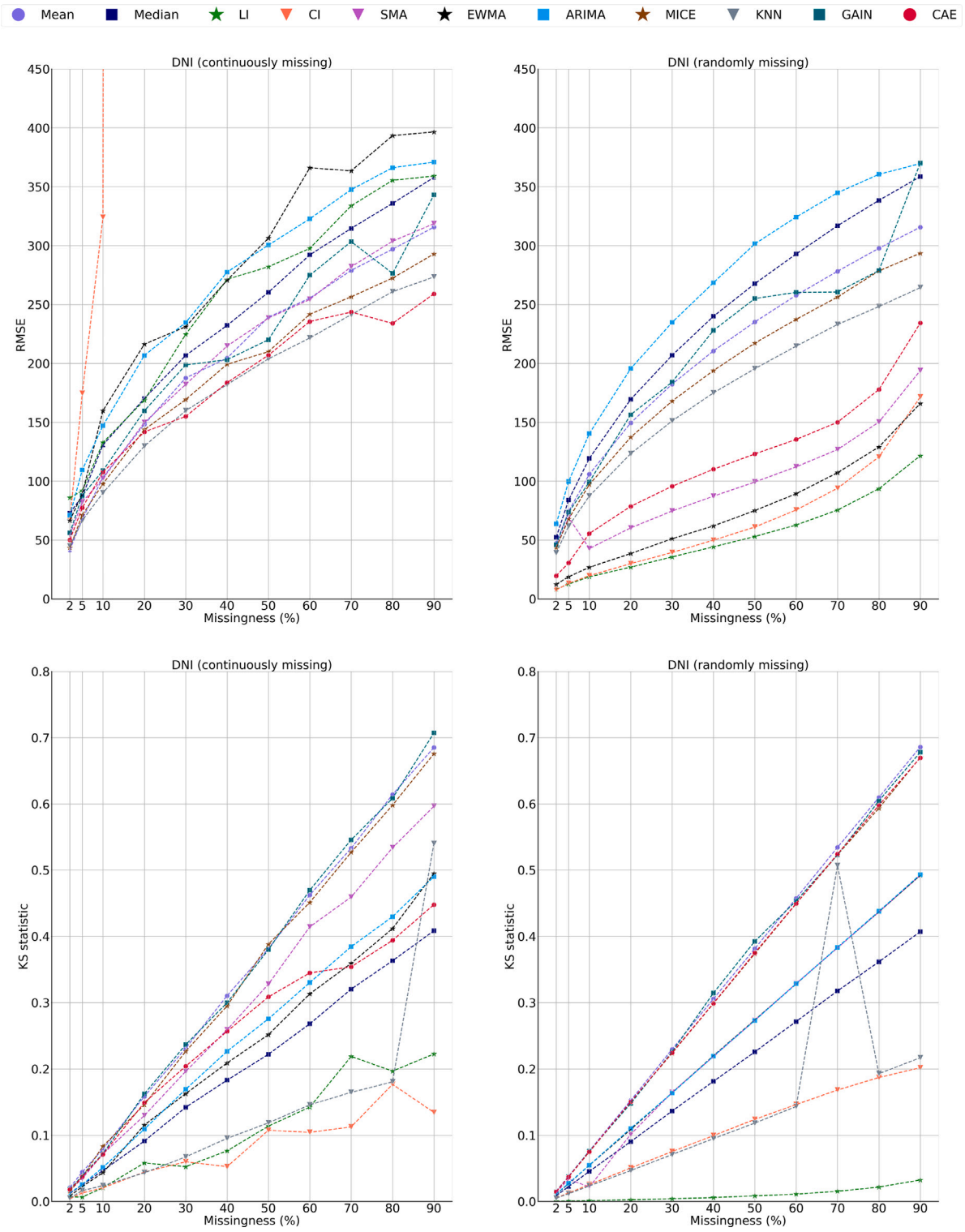

Fig. 6. Statistical validation results of the applied imputation methods for missing rates ranging from 2% to 90% in direct normal irradiance time series (see also the detailed table of results in Appendix D). Results for the continuously missing scenario are shown on the left, while those for the randomly missing scenario are shown on the right.

before validation. A detailed discussion of the RMSE and KS test results is provided below, while the complete numerical results are presented in the tables in Appendix D.

RMSE

Continuously missing data. In GHI imputation (top-left plot of Fig. 5), both KNN and MICE consistently yield the lowest RMSE over the full range of missingness. Notably, KNN demonstrates optimal performance up to 50% of missing rate, after which MICE exhibits a comparable performance. CAE performs well at low missing rates (up to 10%), after which it alternates with GAIN, although the latter demonstrates variability. In DNI imputation (top-left plot of Fig. 6), the mean method yields the lowest RMSE at 2% missingness. As the missingness increases, KNN alternates with CAE in achieving the best

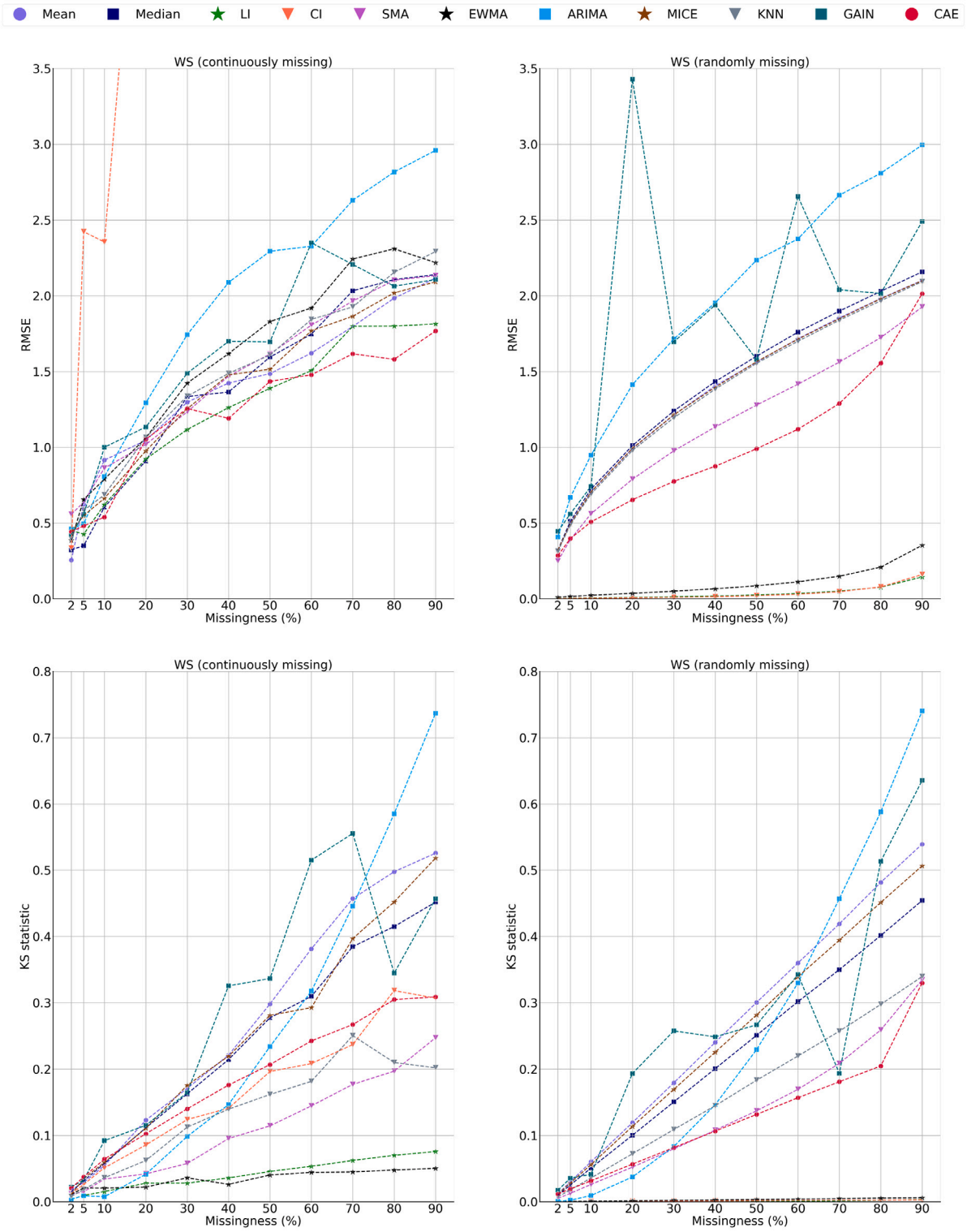

Fig. 7. Statistical validation results of the applied imputation methods for missing rates ranging from 2% to 90% in wind speed time series (see also the detailed table of results in Appendix D). Results for the continuously missing scenario are shown on the left, while those for the randomly missing scenario are shown on the right.

performance, with both methods closely followed by MICE. In wind speed imputation (top-left plot of Fig. 7), mean imputation performs best at 2% missingness, while median imputation outperforms the other methods at 5% and 20%. At other missing rates, LI and CAE alternately achieve the best results. It is noteworthy that, in the continuously missing scenario discussed thus far, CI consistently yields the highest RMSE values, falling outside the plot scale and thus not visible. This poor performance is attributed to the tendency of cubic interpolation to significantly deviate from the original values when interpolating over large continuous gaps, leading to substantial errors.

Randomly missing data. The imputation results in the randomly missing scenario are significantly more regular than those in the continuously missing scenario across all three data types. Specifically, LI, CI, and EWMA consistently outperform the other methods (top-right plots of Figs. 5–7). LI achieves the best RMSE over the entire range of missingness in GHI imputation, while CI performs best at 2% missingness in DNI imputation. In wind speed imputation, CI outperforms the other methods up to 70% missingness, after which LI yields the best RMSE. EWMA follows closely behind LI and CI, outperforming CI at higher missing rates in GHI and DNI. It is observed that both LI and CI are effective in randomly missing scenarios, as they are able to effectively approximate the original time series by interpolating the (single) missing data points. Conversely, when employed to impute large continuous gaps in the data, these methods fail to capture the complex patterns characteristic of high-resolution time series, resulting in larger errors.

KS test

KS statistic. In both GHI and DNI imputation, across continuously and randomly missing data scenarios, the lowest KS statistic values are achieved by LI, CI, and KNN, though their relative performance varies depending on missingness scenario and data type. Notably, in GHI imputation under continuously missing data (bottom-left plot of Fig. 5), KNN yields the lowest KS statistic, followed by CI and LI. In DNI imputation under the same scenario (bottom-left plot of Fig. 6), the KS statistics of LI, CI, and KNN are found to be highly comparable, with CI demonstrating optimal performance from 40% missingness onwards and KNN exhibiting a sharp deviation at 90% missingness. Similar trends are observed for randomly missing data in GHI and DNI imputation (bottom-right plots of Fig. 5 and Fig. 6, respectively), where LI consistently outperforms the other methods, followed by KNN - except for an outlier at 70% missingness in DNI - and CI. In wind speed imputation under continuously missing data (bottom-left plot of Fig. 7), ARIMA performs best up to 10% missingness (matched by LI at 2%), while at higher missing rates, LI and EWMA yield the lowest KS statistics. In the randomly missing scenario (bottom-right plot of Fig. 7), the KS statistics of LI, CI, and EWMA closely reflect the RMSE trends. It is noteworthy that CI yields low KS statistics despite having the highest RMSE values, indicating good distributional similarity in spite of large point-wise errors. Conversely, methods such as MICE and CAE, which achieve low RMSE values, underperform in preserving distributional similarity, as reflected in their KS statistics.

p-value. The significance level for the KS test p-value is set to 0.05. This is the threshold above which the null hypothesis - that the synthetic data distribution does not significantly differ from the original data distribution - cannot be rejected. As outlined in Appendix D, none of the evaluated methods exceeds this threshold in the continuously missing scenario. In contrast, for randomly missing data in GHI, LI and KNN yield p-values above the significance level for missing rates up to 50% and 2%, respectively. In DNI imputation, LI has a p-value above 0.05 up to 20%, while in wind speed imputation, LI, CI, EWMA, and ARIMA show p-values above 0.05 up to 80%, 80%, 40%, and 5%, respectively.

Overall, the best-performing methods based on RMSE in continuously missing scenarios are MICE, KNN, and CAE, whereas LI, CI, and EWMA perform best in randomly missing scenarios. Regarding distributional similarity measured by the KS test, LI, CI, and KNN yield the most favorable results in GHI and DNI imputation, while LI and EWMA perform best in wind speed imputation, alongside CI in the randomly missing case. A salient observation is that the more sophisticated methods (i.e., GAIN and CAE) do not outperform the simpler ones, contrary to expectations. However, similar results have been reported in the literature. Sun et al. [16] compared MICE with GAIN and variational autoencoders, concluding that deep learning-based methods often fail to outperform conventional imputation techniques. In their study, GAIN performed well only under specific missingness mechanisms, while variational autoencoders were prone to mode collapse. They suggest that MICE may be preferable for small- to moderately-sized real-world datasets. Similarly, Wang et al. [17] demonstrated through simulations that MICE outperforms both GAIN and multiple imputation using denoising autoencoders, noting that deep learning-based approaches often generate highly unstable imputations. Furthermore, our results align with those of Liguori et al. [19], who used a similar experimental setup. Their study considered both missingness scenarios, missing rates ranging from 20% to 80%, and three electricity consumption datasets. In the continuously missing scenario, LI, KNN, and CAE alternately achieved the best RMSE, depending on the dataset and missing rate. In contrast, LI consistently outperformed KNN and CAE in imputing randomly missing data across all datasets. However, our study incorporates additional imputation methods and adopts a more comprehensive validation framework. As summarized in Fig. 2 and the table in Appendix A, most previous studies predominantly rely on point-wise error metrics. Nevertheless, the observed discrepancies between RMSE and KS test outcomes highlight the importance of using both point-wise metrics and distribution-based measures to comprehensively assess the statistical performance of imputation methods. This dual approach allows for a more informed selection of imputation techniques tailored to specific use cases characterized by different data types, missingness scenarios, and missing rates.

3.2. Self-sufficient building optimization

The outcomes of the energy system modeling analysis are illustrated in Fig. 8. The results are grouped by missing rates ($\leq$30%, >30% and $\leq$60%, >60% and $\leq$90%) with percentage errors calculated as the average over all missing rates within each group. The baseline at zero represents the results obtained using original data, while the vertical bars represent the percentage deviations resulting from the use of imputed data. As expected, the magnitude of deviations increases with higher missing rates (from top to bottom in Fig. 8). This trend is most pronounced for methods such as median and ARIMA, whereas MICE and KNN show only modest increases, aligning with their RMSE performance and, in the case of KNN, KS statistic. It is noteworthy that methods exhibiting moderate performance in statistical validation, such as LI and SMA, demonstrate deviations similar to those observed in MICE and KNN. These findings highlight two salient points. Firstly, as anticipated, the missing rate can significantly impact the performance of imputation methods in energy system modeling, as demonstrated by the increasing deviations of median and ARIMA. Secondly, for missing rates up to 30% - the most commonly considered range in literature (see Fig. 2) - most methods perform similarly (with the exception of median, EWMA, and ARIMA). Notably, when missing rates are low, mean imputation - a very simple and interpretable method - achieves results that are comparable to those of more complex approaches. However, as the missingness increases, percentage errors rise significantly. Therefore, for low missing rates, simpler methods might be preferable to approximate missing gaps in time series for energy system modeling, particularly in contexts where high-quality training data for more advanced machine learning models are unavailable. Conversely, at higher missing rates, the selection of the most appropriate imputation method becomes crucial to prevent substantial under- or overestimation of component capacities and TAC. Among all the evaluated methods, KNN demonstrates the most consistent performance across the three validation metrics: best RMSE and KS statistic for GHI, and low deviations in energy system modeling over the full range of missingness.

While the self-sufficient building model used in this study provides a robust framework to evaluate the impact of imputation on energy system modeling, its specificity may limit the generalizability of the findings. Therefore, Appendix E includes the validation results for a grid-connected variant of the same system, assuming a constant electricity price of 0.40 €/kWh. In this case - characterized by a reduced

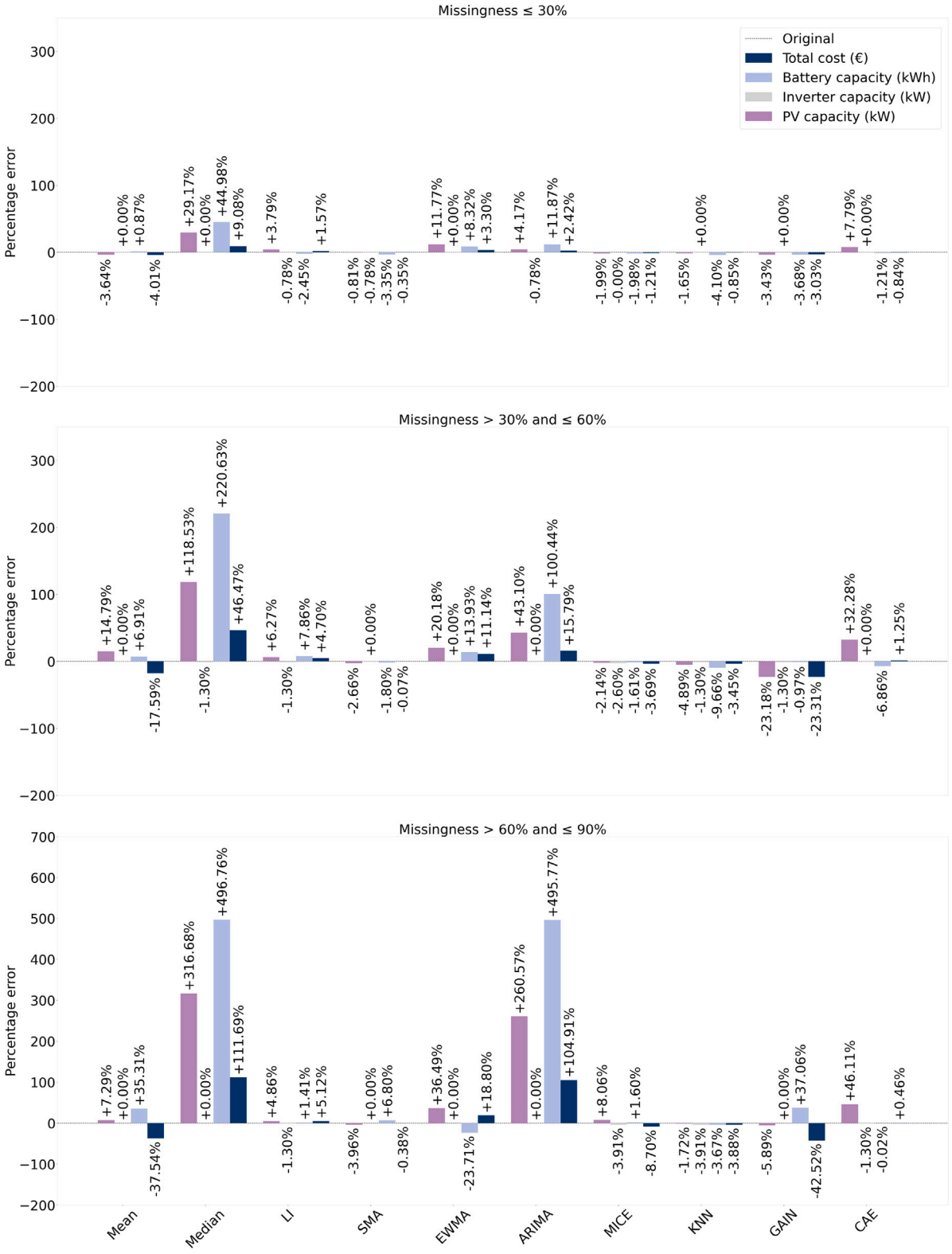

Fig. 8. Percentage error in total annualized costs and component capacities in the self-sufficient building (off-grid system). Zero indicates the baseline corresponding to the results obtained using the original data, while the vertical bars represent the deviations resulting from imputed data. Results are grouped by missing rates (≤30%, >30% and ≤60%, >60% and ≤90%), and the error values shown are group averages.

contribution of renewable energy sources due to the availability of electricity from the grid - KNN confirms its strong performance, particularly under high missingness. To extend these insights and their applicability, future work could explore similar analyses using diverse energy system configurations and geographical locations.

4. Conclusions

This study examined the problem of data gaps in time series used for energy system modeling by applying and evaluating several imputation techniques: mean, median, linear interpolation, cubic interpolation, simple moving average, exponentially weighted moving average, autoregressive integrated moving average, multiple imputation by chained equations, *k*-nearest neighbors, generative adversarial imputation nets, and convolutional denoising autoencoders. Two types of missingness were defined, namely continuous gaps (ranging from six hours to three days) and randomly missing single data points, with missing rates ranging from 2% to 90%. The analysis was conducted on global horizontal irradiance, direct normal irradiance, and wind speed time series. The imputation methods were evaluated using three validation criteria: the root mean square error, the Kolmogorov–Smirnov test, and the impact on a self-sufficient building capacity expansion optimization problem.

The results highlighted substantial differences in the statistical performance of imputation methods across scenarios defined by missingness type and data type. Moreover, discrepancies between the outcomes of the root mean square error and Kolmogorov–Smirnov test underscored the importance of using both point-wise and distributional error metrics. In continuously missing scenarios, the best-performing methods according to root mean square error were multiple imputation by chained equations, *k*-nearest neighbors, and convolutional denoising autoencoders, while linear and cubic interpolation and exponentially weighted moving average exhibited optimal performance in randomly missing scenarios. In terms of distributional similarity, linear interpolation, cubic interpolation, and *k*-nearest neighbors performed best for global horizontal irradiance and direct normal irradiance, while linear interpolation and exponentially weighted moving average showed strong performance for wind speed, alongside cubic interpolation in the randomly missing case. Energy system modeling confirmed the strong performance of multiple imputation by chained equations and *k*-nearest neighbors, and showed that simpler methods such as linear interpolation and simple moving average can achieve results comparable to those of more advanced techniques. Overall, *k*-nearest neighbors emerged as the most consistently effective approach across the three validation criteria.

This work provides an explicit quantification of the impact of missing data imputation on energy system modeling, alongside a comprehensive evaluation of statistical performance across diverse imputation methods and time series. However, the specificity of the case study may limit the generalizability of the findings. Future research could expand this analysis by considering energy systems in different geographical locations and incorporating diverse renewable energy sources. Intermediate missingness scenarios, including both continuously and randomly missing data within the same time series, could also be considered to investigate potential variations in the performance of the different techniques. Additionally, the evaluation of emerging state-of-the-art methods, such as transformers or generative AI-based imputation, could provide valuable insights into their effectiveness in filling data gaps in time series for energy system modeling. Evaluating the impact of these techniques on the optimal sizing and operation of system components could foster the integration of renewable energy sources in future energy systems, contributing to the attainment of climate and environmental goals.

CRediT authorship contribution statement

Claudio Mantuano: Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Olalekan Omoyele:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Maximilian Hoffmann:** Writing – review & editing, Visualization, Supervision, Formal analysis, Conceptualization. **Jann Michael Weinand:** Writing – review & editing, Visualization, Supervision, Formal analysis, Conceptualization. **Massimo Panella:** Writing – review & editing, Conceptualization. **Detlef Stolten:** Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Helmholtz Association, Germany under the program “Energy System Design”.

Appendix A. Literature review

Reviewed studies categorized by year of publication, authors, application domain, proposed imputation method, comparative methods, solution approach, missing rates, data resolution, and error metrics (see Tables A.1 and A.2).

Table A.1
List of abbreviations used in Table A.2.

List of abbreviations			
AD	Adaptive boosting	MASE	Mean absolute scaled error
AE	Absolute error/Autoencoder	MBE	Mean bias error
AE-CD	Autoencoder - coordinate descent	MFA	Mixture factor analysis
ANN	Artificial neural network	MGEL-ELM	Meta-learning extreme learning machine optimized with golden eagle optimization and logistic map
ANNEM	Artificial neural network estimation method		
AR-ANN	Autoregressive artificial neural network	MICE	Multiple imputation by chained equations
ARIMA	Autoregressive integrated moving average	MIDA	Multiple imputation using denoising autoencoders
ARMA	Autoregressive moving average	MIDWM	Modified inverse distance weighting method
Aug	Augmentation	ML	Machine learning imputation algorithm
AVG	Average algorithm	MLP	Multi-layer perceptron
AvgNRMSE	Average normalized root mean square error	MLP-AVG	Ensemble model adopting the average of subnetworks
BN	Bayesian network	MLPNN	Multi-layer perceptron neural network
BPNN	Back-propagation neural network	MLP-S	Single multi-layer perceptron
BRITS	Bidirectional recurrent imputation for time series	MLR	Multiple linear regression
CAE	Convolutional denoising autoencoder neural network	MPCA	Multilinear principal component analysis

(*continued on next page*)

Table A.1 (*continued*).

List of abbreviations			
CBRi	Case-based reasoning imputation	MRD	Maximum rank distance
CCWM	Coefficient of correlation weighting method	MRE	Mean relative error
CD	Correlation dimension	MSE	Mean square error
CI	Cubic interpolation	MVTSI	Multivariate time series imputation
CNN	Convolutional neural network	NI	Nearest interpolation
CRM	Coefficient of residual mass	NMSE	Normalized mean square error
C-SpI	Cubic spline interpolation	N – N	Nearest neighbor interpolation
CVMAE	Coefficient of variation of the mean absolute error	NN	Neural network
CVME	Coefficient of variation of the mean error	NNWM	Nearest neighbor distance weighting method
DAE	Denoising autoencoder	NR	Normal ratio
DBN	Deep belief network	NRMSE	Normalized root mean square error
DM	Decision matrix	NRWC	Normal ratio weighted with correlations
DT-NN	Decision tree neural network	NSE	Nash–Sutcliffe efficiency
e	Percentage error	OWA	Optimally weighted average
eLAI	Extended learning-based adaptive imputation method	PCHIP	Shape-preserving piecewise cubic hermite interpolating polynomial interpolation
EM	Expectation maximization		
EM-MCMC	Expectation–maximization Monte Carlo Markov chain	PPCA	Probabilistic principal component analysis
FA	Factor analysis	R^2	Coefficient of determination
FCNN	Fully connected neural network	rMAE	Relative mean absolute error
FFSGAM	Fixed functional set genetic algorithm method	RMSD	Root mean square deviation
FIR	Fuzzy inductive reasoning	RMSE	Root mean square error
GAIN	Generative adversarial imputation nets	RNN	Recurrent neural network
GAN	Generative adversarial networks	RNNWM	Revised nearest neighbor weighting method
GBM	Gradient boosting machine	rRMSE	Relative root mean square error
GRU	Gated recurrent unit	RT	Regression tree
HA	Historical average	SAA	Simple arithmetic average
H-S-A	Hargreaves, Samani and Annandale method	SAE-CD	Sparse autoencoder - coordinate descent
IDWM	Inverse distance weighting method	SAME	Same datetime interval averaged algorithm
IEWM	Inverse exponential weighting method	SASR	Statistically adjusted solar radiation
KEM	Kriging estimation method	SC	Seasonal component
KF	Kalman filter	SENet	Softmax ensemble network
KGE	Kling-Gupta efficiency	SI	Spline interpolation
KNN	*k*-nearest neighbors	SMA	Simple moving average
LAI	Learning-based adaptive imputation method	sMAPE	Symmetric mean absolute percentage error
LANN	Local average of nearest neighbors	SSA	Singular spectrum analysis
LI	Linear interpolation	StI	Stineman interpolation
LOCF	Last observation carried forward	SVM	Support vector machine
LR	Linear regression	TBA	Temperature-based approach
LSTM	Long-short term memory	TMMVAE	Temporal multi-modal variational autoencoder
LSTM-BIT	Deep learning and transfer learning-based method	VGI	Vector-autoregressive gaussian interpolation
MA	Moving average	WMAPE	Weighted mean absolute percentage error
MAE	Mean absolute error	XGBoost	Extreme gradient boosting
MAKIMA	Modified Akima interpolation	XGBoost-DE	Extreme gradient boosting by differential evolution
MAPE	Mean absolute percentage error	ZERO	Zero-replace algorithm
MARS	Multivariate adaptive regression splines		

Table A.2
Literature review of studies on missing data imputation in time series for energy applications.

Year	Authors	Application	Proposed method	Comparative methods	Approach	Missingness	Resolution	Err. metrics
2023	Bülte et al. [18]	Energy data	MVTSI	LSTM, LOCF, KNN	ML, Statistical	1%–23%	1 h	MSE, MAE
2023	Boriratrit et al. [69]	Solar irradiance	SAME	AVG, ZERO, ML	ML, Statistical	20%	1 h	RMSE, *p*-value
2023	Liguori et al. [19]	Electricity consumption	CAE + Aug	CAE, RF, KNN, LI, Mean	ML, Statistical	20%, 40%, 60%, 80%	15 min	MAE, RMSE, NRMSE
2023	Centeno et al. [20]	Solar power	ANN + Encoder-Decoder	Random Sample, Mean, Mode, Median, EM, KNN, solarGAN	ML, Statistical	10%–90%	15 min	WMAPE, RMSE, R^2
2023	Başakin et al. [38]	Solar irradiance	XGBoost-DE	LI, SI, MARS, RF	ML, Statistical	5%, 10%, 20%, 30%	24 h	RMSE, R^2 MAE, NSE, KGE
2022	Phan et al. [12]	Solar power	RF-MICE	ZERO, Mean, Median, Mode, LR, interpolation, MA, KNN, MICE, SC + Mean, SC + Median, SC + Mode, SC + Mean + LR, SC + Median + LR, SC + Mode + LR	ML, Statistical	37%	1 h	RMSE
2022	Mohamad et al. [70]	Solar irradiance	–	N-N, LI, C-SpI, PCHIP, MAKIMA, StI, Bezier curve, SMA	ML, Statistical	10%–50%	5 min	MAE, RMSE, MIE, MBE
2022	Hussain et al. [39]	Electricity consumption	CNN-LSTM	CNN, LSTM	ML	N/A	24 h	RMSE

(*continued on next page*)

Table A.2 (*continued*).

Year	Authors	Application	Proposed method	Comparative methods	Approach	Missingness	Resolution	Err. metrics
2021	Denhard et al. [43]	Solar irradiance	–	KF, LI, SI, StI, SMA, Linear weighted MA, Exponential weighted MA, LOCF, NOCF, Random sample	Statistical	N/A	1 min, 30 min	RMSE, MAE, MBE
2021	Zhang et al. [13]	Solar power	solarGAN	Mean, LOCF, MF, KNN, MICE, GAIN, GAN-Z	ML, Statistical	10%–90%	1 h	MSE
2021	Yelchuri et al. [71]	Solar irradiance	–	KF, ARIMA, LI, SI, StI, SMA, Linear weighted MA, Exponential weighted MA	Statistical	N/A	15 min	MAE
2021	Shen et al. [72]	Solar power	TMMVAE	Mean, GAIN, TMVAE, TVAE-Num, MMVAE	ML, Statistical	10%–90%	30 min	AvgNRMSE
2021	Flores et al. [73]	Solar irradiance	CBRi2	CBRi, LANN, ARIMA	ML, Statistical	10%, 20%, 30%	24 h	RMSE, MAE, MAPE, R^2
2021	Jeong et al. [5]	Electricity consumption	MFA	Mean, LI, MA, BRITS, PPCA, FA, MPCA	ML, Statistical	10%–50%	15 min	RMSE, CV(RES)
2021	Liu et al. [14]	Wind turbines (SCADA)	SAE-CD	Mean, MIDA, GAIN, AE-CD, MICE	ML, Statistical	N/A	10 s	NRMSE
2021	Ho et al. [74]	Solar irradiance	ANN	–	ML	N/A	5 min	RMSE
2021	Wang et al. [21]	Electricity consumption	–	KNN, SVR, MLP, LI, ARIMA	ML, Statistical	N/A	1 min	MAPE
2020	Qu et al. [29]	Wind speed	ipGAN	ARMA, BPNN, SVR, DBN, CNN, DAE	ML, Statistical	N/A	10 min	MAE, MSE, NRMSE
2020	Lindig et al. [75]	Solar irradiance	–	Isotropic, Klucher, Hay-Davies, Reindl, King, Perez, RF, Extra trees, Gradient boosting, Histogram-based gradient boosting	ML, Statistical	20%	15 min	RMSE
2020	Park et al. [22]	Fault detection (PV fleet)	–	AR, Simple regression, Multiple regression, KNN	ML, Statistical	N/A	1 h	NRMSE
2020	Khare et al. [23]	Solar irradiance	GAN	Mean, KNN	ML, Statistical	14%, 16%, 18%	24 h	MSE, RMSE, R^2
2020	Ma et al. [24]	Electricity consumption	LSTM-BIT	Mean, LI, KNN, SVM, RF, FCNN, RNN, LSTM	ML, Statistical	10%–90%	15 min	RMSE, R^2
2020	Zhao et al. [76]	Electricity consumption	Intelligent electricity data imputation method	Transaction imputation method, Arithmetic average method	Statistical	N/A	24 h	Standard error
2020	Khan et al. [11]	Electricity consumption	Machine learning-based hybrid ensemble model	–	ML	20%	1 h	MAE, R^2
2020	Chen et al. [25]	Solar irradiance	DT-NN interpolation	NN, DT-NN, ARMA, SVM, Weighted KNN	ML, Statistical	34%	1 h	MAE, RMSE
2020	Li et al. [77]	Electricity consumption	BPNN	–	ML	N/A	1 h	Relative error
2020	Jung et al. [26]	Electricity consumption	SENet + MLP	LR, AD, SVR, GBM, XGBoost, RF, MLP-S, CNN, RNN, MLP-AVG, SENet + CNN, SENet + RNN	ML	10%–30%	1 h	MAPE, RMSE
2019	Kim et al. [8]	Solar power	–	LI, Mode, KNN, MICE	ML, Statistical	10%, 15%, 20%	1 h	RMSE, MRE, RMSD, MRD
2018	Demirhan et al. [40]	Solar irradiance	–	Interpolation, KF, Persistence, Weighted MA, Mean, Mode, Median, Random sample, Seasonal decomposition, Seasonal splitting	Statistical	5%, 10%, 25%, 50%	1 min, 1 h, 24 h, 1 wk	MASE, rMAE, rRMSE
2018	Sánchez et al. [78]	Wind speed	LI + LR	CI, LI, LR, SVM, MLP, LI + SVM, LI + MLP	ML, Statistical	1.5%	10 min	MAE
2018	Rahman et al. [34]	Electricity consumption	Deep RNN	MLP	ML	3%	1 h	RMSE, Pearson coefficient
2017	Layanun et al. [28]	Solar irradiance	SVM + Mean	Interpolation, MA, Mean	ML, Statistical	14%	1 h, 3 h	MAE, RMSE
2017	Kim et al. [79]	Electricity consumption	LAI, eLAI	LI, OWA, PPCA	ML, Statistical	1%–30%	30 min	MAPE, RMSE
2017	Jurado et al. [80]	Electricity consumption	Flexible FIR	–	ML	9%–81%	1 h	NMSE, sMAPE

(*continued on next page*)

Table A.2 (*continued*).

Year	Authors	Application	Proposed method	Comparative methods	Approach	Missingness	Resolution	Err. metrics
2016	Peppanen et al. [81]	Electricity consumption	OWA	HA, LI	Statistical	N/A	15 min	MAPE
2015	Zainudin et al. [41]	Solar irradiance	–	LI, CI, NI, SI, Bezier/Said-Ball (Piecewise interpolation)	Statistical	10%–50%	1 h	RMSE, R^2
2015	Shukur et al. [33]	Wind speed	AR-ANN	LI, N-N, State space	ML, Statistical	10%, 20%, 30%	24 h	RMSE
2014	Turrado et al. [15]	Solar irradiance	MICE	IDW, MLR	Statistical	0%–1%	10 min	RMSE, MAE
2014	Saaban et al. [82]	Solar irradiance	–	LI, CI, NI, SI, Bezier/Said-Ball (Piecewise interpolation)	Statistical	10%–50%	1 h	RMSE, R^2
2014	Kasam et al. [83]	Temperature	VGI	–	Statistical	N/A	1 h	Relative error
2014	Ogunsola et al. [84]	Solar irradiance	–	SSA, SASR, TBA	Statistical	N/A	1 h	CVMAE, CVRMSE, CVME, R^2
2012	Garnier et al. [32]	Solar irradiance, Indoor temperature	Feedforward ANN	–	ML	2%	1 h	MRE, Weighted MRE
2012	Yozgatligil et al. [85]	Precipitation, Temperature	–	SAA, NR, NRWC, MLPNN, EM-MCMC	ML, Statistical	10%, 20%, 50%	1 mo	CVRMSE, CD
2011	Daut et al. [27]	Solar irradiance, Temperature	Hargreaves + LR	Hargreaves, LR	Statistical	58%	1 mo	RMSE, CRM, NSE, *e*
2010	Kim et al. [31]	Precipitation	ANN + RT	ANN, RT	ML	0%–2%	24 h	RMSE, Pearson coefficient
2009	Teegavarapu et al. [86]	Precipitation	FFSGAM	IDWM, CCWM	ML, Statistical	N/A	24 h	RMSE, MAE, AE, Correlation coefficient
2005	Teegavarapu et al. [87]	Precipitation	–	IDWM, MIDWM, CCWM, IEWM, NNWM, RNNWM, ANNEM, KEM	ML, Statistical	N/A	24 h	MAE, MRE, RMSE, R^2
2005	Jin et al. [88]	Temperature	Stochastic binning	–	Statistical	N/A	1 h	MAPE

Appendix B. Techno-economic data of the self-sufficient building

	CAPEX				OPEX		Lifetime
Components	**Fixed**		**Capacity-specific**		**Fixed + Capacity-specific**		**Years**
Photovoltaic ground	—	—	4000.00	€/kW$_p$	1.00	% Inv./a	20
Photovoltaic rooftop	—	—	769.00	€/kW$_p$	1.00	% Inv./a	20
Inverter	—	—	75.00	€/kW$_p$	—	—	20
Battery	—	—	301.00	€/kWh$_p$	—	—	15
Reversible solid oxide cell	5000.00	€	2400.00	€/kW$_{el}$	1.00	% Inv./a	15
Heat pump	4230.00	€	504.90	€/kW$_{th}$	1.50	% Inv./a	20
Thermal storage	—	—	90.00	€/kWh$_{th}$	0.01	% Inv./a	25
E-heater & E-boiler	—	—	60.00	€/kW$_{th}$	2.00	% Inv./a	30
Tank	—	—	0.79	€/kWh$_{H2}$	—	—	25
Dibenzyltoluene	—	—	1.25	€/kWh$_{H2}$	—	—	25
Hydrogen vessels	—	—	15.00	€/kWh$_{H2}$	—	—	25
Hydrogenizer	2123.30	€	761.10	€/kW$_{H2}$	1.00	% Inv./a	20
Dehydrogenizer	1140.00	€	408.60	€/kW$_{H2}$	1.00	% Inv./a	20
Low pressure compressor	—	—	1716.71	€/kW$_p$	1.00	% Inv./a	25
High pressure compressor	560.00	€	1329.80	€/kW$_p$	1.00	% Inv./a	25
Heat-exchangers 1 and 2	—	—	1.00	€/kW$_{th}$	1.00	% Inv./a	—
Expanders 1 and 2	—	—	1.00	€/kW$_{th}$	1.00	% Inv./a	25

Appendix C. Self-sufficient building capacity expansion optimization problem

The formulation of the self-sufficient building capacity expansion optimization problem is presented in (C.1 - C.10) using the notation described in the following table, as detailed in Refs. [50,65,66].

Symbol	Description
Sets	
T	Time steps
M	Components
G	Commodities
M^{source}	Subset of components representing sources
M^{sink}	Subset of components representing sinks
M^{store}	Subset of components representing storage units
M^{conv}	Subset of components representing conversion units
M^{g}	Components associated with a commodity in g
Parameters	
C_c^{CAPEX}	Capital expenditures of component c
C_c^{OPEX}	Operational expenditures of component c
η_c^{ch}	Efficiency (charging) of storage unit c
η_c^{dis}	Efficiency (discharging) of storage unit c
γ_c	Conversion factor (from one commodity to another) of conversion unit c
Variables	
x_c^{cap}	Installed capacity of component c
$x_{c,t}^{SOC}$	State of charge of storage unit c at time t
$x_{c,t}^{op}$	Operation rate of component c at time t
$x_{c,t}^{\text{op,ch}}$	Operation rate (charging) of storage unit c at time t
$x_{c,t}^{\text{op,dis}}$	Operation rate (discharging) of storage unit c at time t
$f_{c,t}$	Flow of commodity c at time t

$$\min \quad \left(\sum_{c\in M}\left(C_c^{\text{CAPEX}} + \sum_{t\in T} C_c^{\text{OPEX}} x_{c,t}^{\text{op}}\right)\right) \tag{C.1}$$

s.t. $\forall g \in G,\ t \in T$:

$$\sum_{c\in M^g} f_{c,t} = 0 \tag{C.2}$$

$$f_{c,t} = x_{c,t}^{\text{op}} \qquad \forall c \in M^{\text{source}} \cap M^{\text{g}} \tag{C.3}$$

$$f_{c,t} = -x_{c,t}^{\text{op}} \qquad \forall c \in M^{\text{sink}} \cap M^{\text{g}} \tag{C.4}$$

$$f_{c,t} = \gamma_c x_{c,t}^{\text{op}} \qquad \forall c \in M^{\text{conv}} \cap M^{\text{g}} \tag{C.5}$$

$$f_{c,t} = x_{c,t}^{\text{op,dis}} - x_{c,t}^{\text{op,ch}} \qquad \forall c \in M^{\text{store}} \cap M^{\text{g}} \tag{C.6}$$

s.t. $\forall t \in T$:

$$x_{c,t}^{\text{op}} \geq 0 \qquad \forall c \in M^{\text{source, sink, conv, store}} \tag{C.7}$$

$$x_{c,t}^{\text{op}} \leq x_c^{\text{cap}} \qquad \forall c \in M^{\text{source, sink, conv}} \tag{C.8}$$

$$x_{c,t+1}^{\text{SOC}} = x_{c,t}^{\text{SOC}} + \eta_c^{\text{ch}} x_{c,t}^{\text{op, ch}} - \frac{x_{c,t}^{\text{op, dis}}}{\eta_c^{\text{dis}}} \qquad \forall c \in M^{\text{store}} \tag{C.9}$$

$$0 \leq x_{c,t}^{\text{SOC}} \leq x_c^{\text{cap}} \qquad \forall c \in M^{\text{store}} \tag{C.10}$$

Appendix D. Imputation results

The RMSE values for all imputation methods, missing rates, and experimental settings are presented. The best results, relative to each experimental setting, are highlighted in green, the worst results in red, and intermediate results in white.

RMSE

	Mean	Median	LI	CI	SMA	EWMA	ARIMA	MICE	KNN	GAIN	CAE
GHI - CONTINUOUSLY MISSING											
2%	45,145	51,693	43,318	254,799	46,302	40,501	50,018	24,117	16,221	72,278	19,880
5%	60,684	76,359	59,818	1001,802	58,342	72,344	90,346	38,926	36,889	49,542	45,252
10%	81,101	93,380	91,425	2074,577	71,100	92,318	105,027	55,877	47,476	80,491	56,755
20%	112,680	141,383	124,444	10184,113	106,981	144,175	147,144	76,808	66,776	93,484	99,509
30%	141,125	153,397	150,701	3792,035	130,911	185,167	166,204	87,024	79,252	132,816	128,662
40%	166,874	189,145	206,162	13666,534	147,460	198,313	205,123	98,906	96,048	129,819	153,283
50%	176,247	211,920	204,580	13256,809	168,851	252,395	234,852	112,851	104,895	140,240	167,230
60%	191,049	208,493	221,163	10891,713	184,311	258,239	247,222	118,300	118,526	281,501	169,095
70%	207,683	254,532	249,673	16801,088	204,145	280,287	270,253	130,250	129,517	191,504	192,483
80%	223,211	262,357	265,458	17879,520	222,778	285,597	283,717	138,849	139,485	252,277	198,337
90%	236,708	262,143	261,138	21603,467	232,068	299,958	288,767	146,937	146,437	184,771	220,869
GHI - RANDOMLY MISSING											
2%	34,976	41,423	4,663	5,165	31,828	6,962	44,731	22,000	21,075	50,309	13,890
5%	55,948	64,133	7,659	7,886	49,694	10,830	72,712	34,769	31,847	79,682	21,020
10%	78,842	92,300	10,577	11,433	22,580	14,700	103,289	48,748	45,534	58,908	24,852
20%	111,710	131,174	15,746	16,905	31,895	22,118	141,448	68,959	65,200	73,866	30,537
30%	136,598	160,040	20,447	23,157	39,121	28,683	174,074	84,643	79,701	109,262	35,244
40%	157,447	185,068	25,138	30,029	46,057	35,102	202,440	97,578	92,081	136,081	39,150
50%	176,292	206,641	30,172	36,309	52,376	42,446	231,025	109,380	102,739	136,272	45,050
60%	193,169	226,367	35,677	45,627	58,813	49,972	250,261	120,031	112,836	180,411	53,357
70%	208,671	244,597	43,101	56,146	66,679	59,192	267,129	129,258	122,137	219,326	68,452
80%	223,515	261,385	53,611	72,526	78,959	71,334	280,858	138,194	130,494	174,365	100,572
90%	236,775	276,465	67,585	107,222	108,636	88,063	288,375	146,736	138,679	236,203	158,934
DNI - CONTINUOUSLY MISSING											
2%	41,479	72,914	86,001	66,542	54,275	66,555	70,943	43,467	44,874	56,071	50,111
5%	67,733	87,158	91,884	174,607	82,243	89,205	109,574	71,178	66,633	87,768	77,390
10%	103,260	130,716	132,718	323,945	102,147	159,863	147,098	97,756	90,048	109,017	107,653
20%	148,215	170,168	168,716	29067,282	149,749	216,280	206,728	143,715	129,691	159,646	141,783
30%	187,617	206,736	224,596	6028,606	182,063	231,104	234,648	169,266	159,815	198,692	154,921
40%	204,450	232,210	271,372	9922,564	214,675	270,551	277,470	199,181	181,966	203,101	183,560
50%	239,258	260,346	282,004	15276,187	238,463	306,366	300,435	209,942	203,823	220,056	206,948
60%	255,389	292,144	297,652	22446,680	254,281	366,075	322,691	241,718	221,532	274,864	235,551
70%	278,911	314,585	333,778	21132,360	282,062	363,472	347,534	256,532	241,382	303,357	243,554
80%	296,953	335,833	355,430	28371,223	303,710	393,381	366,093	272,407	260,812	276,391	233,951
90%	315,649	357,923	359,056	34462,974	318,553	396,497	370,894	292,892	273,363	343,038	258,987
DNI - RANDOMLY MISSING											
2%	47,083	52,245	8,240	7,717	43,942	12,314	63,654	43,557	38,957	46,019	19,569
5%	74,421	83,953	12,628	13,218	69,185	18,852	99,787	67,850	61,782	73,570	30,650
10%	105,857	119,259	18,776	19,697	42,770	26,790	140,350	97,080	87,311	99,250	55,529
20%	149,448	169,422	26,987	29,957	60,308	38,487	195,520	137,243	123,292	156,439	78,590
30%	182,350	206,778	35,625	39,394	74,622	51,067	234,717	167,943	151,096	184,037	95,631
40%	210,558	239,939	44,179	49,825	87,213	61,974	268,436	193,873	174,888	227,961	110,059
50%	235,121	267,620	52,979	61,002	99,238	75,011	301,448	217,164	195,461	255,020	123,057
60%	257,786	292,872	62,760	75,521	112,134	89,309	324,232	237,367	214,628	260,309	135,409
70%	278,205	316,860	75,459	93,963	126,837	107,143	344,862	256,333	233,039	260,628	149,995
80%	297,724	338,377	93,539	120,530	150,152	129,004	360,508	278,458	248,234	278,777	177,764
90%	315,593	358,601	121,560	171,439	193,884	165,940	369,660	293,502	264,304	370,135	234,260
WS - CONTINUOUSLY MISSING											
2%	0,255	0,323	0,452	0,334	0,557	0,386	0,461	0,441	0,395	0,424	0,443
5%	0,625	0,351	0,427	2,423	0,637	0,655	0,495	0,549	0,584	0,558	0,481
10%	0,915	0,605	0,621	2,355	0,863	0,789	0,807	0,663	0,686	1,000	0,539
20%	1,046	0,910	0,923	5,557	1,018	1,062	1,293	0,976	1,065	1,133	1,053
30%	1,297	1,335	1,117	8,474	1,234	1,423	1,743	1,254	1,337	1,488	1,256
40%	1,424	1,365	1,262	8,833	1,469	1,618	2,088	1,478	1,487	1,699	1,191
50%	1,486	1,595	1,389	10,028	1,616	1,830	2,294	1,515	1,609	1,695	1,435
60%	1,621	1,747	1,506	12,280	1,806	1,919	2,327	1,769	1,844	2,350	1,478
70%	1,797	2,033	1,799	10,584	1,965	2,242	2,630	1,865	1,926	2,206	1,617
80%	1,984	2,107	1,799	9,864	2,102	2,310	2,817	2,019	2,155	2,063	1,580
90%	2,116	2,139	1,814	10,419	2,133	2,219	2,959	2,091	2,292	2,107	1,768
WS - RANDOMLY MISSING											
2%	0,315	0,321	0,002	0,001	0,249	0,010	0,408	0,321	0,313	0,445	0,286
5%	0,497	0,511	0,003	0,002	0,391	0,016	0,668	0,495	0,488	0,560	0,397
10%	0,705	0,724	0,005	0,003	0,561	0,023	0,948	0,705	0,692	0,742	0,507
20%	0,991	1,011	0,009	0,006	0,789	0,036	1,413	0,993	0,979	3,429	0,654
30%	1,213	1,239	0,013	0,009	0,976	0,050	1,715	1,214	1,195	1,695	0,774
40%	1,406	1,433	0,018	0,014	1,134	0,066	1,954	1,396	1,384	1,939	0,875
50%	1,562	1,601	0,025	0,020	1,279	0,086	2,235	1,565	1,555	1,579	0,990
60%	1,717	1,760	0,035	0,030	1,416	0,112	2,375	1,714	1,700	2,656	1,119
70%	1,854	1,899	0,050	0,046	1,562	0,149	2,664	1,849	1,839	2,039	1,289
80%	1,982	2,029	0,076	0,078	1,724	0,209	2,809	1,980	1,967	2,015	1,555
90%	2,102	2,158	0,144	0,160	1,925	0,353	2,995	2,099	2,093	2,491	2,013

The KS test statistic values for all imputation methods, missing rates, and experimental settings are presented. The best results, relative to each experimental setting, are highlighted in green, the worst results in red, and intermediate results in white.

KOLMOGOROV-SMIRNOV TEST STATISTIC

	Mean	Median	LI	CI	SMA	EWMA	ARIMA	MICE	KNN	GAIN	CAE
GHI - CONTINUOUSLY MISSING											
2%	0,014	0,014	0,011	0,006	0,012	0,009	0,010	0,012	0,003	0,018	0,019
5%	0,035	0,029	0,010	0,009	0,026	0,024	0,027	0,027	0,005	0,034	0,038
10%	0,070	0,053	0,026	0,021	0,049	0,050	0,049	0,052	0,008	0,074	0,067
20%	0,143	0,103	0,075	0,056	0,108	0,099	0,098	0,104	0,018	0,116	0,118
30%	0,203	0,153	0,062	0,073	0,157	0,142	0,148	0,169	0,021	0,210	0,168
40%	0,272	0,203	0,132	0,104	0,206	0,201	0,205	0,221	0,037	0,285	0,205
50%	0,360	0,254	0,156	0,089	0,259	0,223	0,249	0,273	0,043	0,340	0,240
60%	0,426	0,327	0,172	0,130	0,304	0,292	0,299	0,331	0,046	0,520	0,279
70%	0,498	0,364	0,183	0,116	0,351	0,316	0,340	0,384	0,045	0,520	0,319
80%	0,568	0,401	0,183	0,151	0,398	0,383	0,393	0,439	0,057	0,631	0,331
90%	0,632	0,505	0,243	0,165	0,469	0,448	0,449	0,503	0,071	0,610	0,349
GHI - RANDOMLY MISSING											
2%	0,014	0,010	0,000	0,005	0,010	0,010	0,010	0,011	0,002	0,017	0,011
5%	0,035	0,025	0,000	0,013	0,026	0,025	0,025	0,027	0,004	0,042	0,028
10%	0,070	0,050	0,000	0,025	0,016	0,050	0,050	0,055	0,008	0,065	0,055
20%	0,140	0,101	0,001	0,050	0,093	0,099	0,099	0,110	0,017	0,127	0,109
30%	0,210	0,151	0,001	0,075	0,149	0,148	0,150	0,166	0,025	0,209	0,163
40%	0,279	0,200	0,002	0,098	0,199	0,198	0,200	0,221	0,033	0,274	0,218
50%	0,350	0,251	0,002	0,125	0,248	0,248	0,249	0,277	0,041	0,341	0,272
60%	0,420	0,300	0,003	0,150	0,298	0,298	0,298	0,331	0,051	0,420	0,325
70%	0,489	0,350	0,004	0,173	0,348	0,348	0,348	0,386	0,059	0,543	0,380
80%	0,560	0,401	0,005	0,199	0,397	0,397	0,398	0,442	0,066	0,448	0,434
90%	0,628	0,452	0,012	0,218	0,447	0,447	0,448	0,497	0,078	0,595	0,485
DNI - CONTINUOUSLY MISSING											
2%	0,021	0,012	0,007	0,004	0,017	0,008	0,013	0,016	0,005	0,019	0,018
5%	0,044	0,025	0,007	0,013	0,035	0,023	0,026	0,037	0,016	0,039	0,037
10%	0,079	0,047	0,021	0,021	0,072	0,043	0,051	0,083	0,024	0,072	0,071
20%	0,158	0,091	0,058	0,044	0,130	0,115	0,109	0,146	0,043	0,162	0,149
30%	0,231	0,142	0,052	0,060	0,196	0,163	0,169	0,226	0,067	0,236	0,204
40%	0,310	0,183	0,076	0,052	0,259	0,209	0,226	0,295	0,095	0,300	0,257
50%	0,381	0,222	0,114	0,107	0,328	0,252	0,276	0,388	0,119	0,380	0,309
60%	0,462	0,268	0,142	0,104	0,414	0,313	0,330	0,451	0,146	0,470	0,345
70%	0,533	0,320	0,219	0,113	0,459	0,359	0,384	0,527	0,165	0,546	0,354
80%	0,614	0,363	0,197	0,177	0,534	0,412	0,430	0,598	0,180	0,608	0,394
90%	0,685	0,408	0,223	0,134	0,596	0,495	0,490	0,676	0,540	0,707	0,448
DNI - RANDOMLY MISSING											
2%	0,015	0,009	0,000	0,005	0,013	0,011	0,011	0,015	0,005	0,015	0,015
5%	0,038	0,022	0,001	0,013	0,034	0,027	0,027	0,037	0,012	0,037	0,037
10%	0,076	0,045	0,001	0,025	0,022	0,055	0,055	0,075	0,023	0,076	0,075
20%	0,152	0,090	0,003	0,050	0,101	0,109	0,110	0,150	0,047	0,148	0,150
30%	0,230	0,136	0,004	0,075	0,164	0,164	0,164	0,224	0,071	0,226	0,224
40%	0,306	0,181	0,006	0,099	0,219	0,220	0,219	0,299	0,095	0,314	0,299
50%	0,381	0,226	0,008	0,124	0,273	0,274	0,273	0,374	0,118	0,392	0,375
60%	0,457	0,271	0,011	0,146	0,329	0,328	0,328	0,449	0,143	0,453	0,450
70%	0,534	0,318	0,015	0,168	0,383	0,383	0,383	0,523	0,507	0,523	0,524
80%	0,610	0,362	0,022	0,187	0,437	0,437	0,438	0,594	0,193	0,605	0,598
90%	0,686	0,407	0,032	0,202	0,492	0,492	0,493	0,670	0,217	0,678	0,670
WS - CONTINUOUSLY MISSING											
2%	0,018	0,015	0,004	0,012	0,012	0,011	0,004	0,011	0,008	0,022	0,021
5%	0,029	0,032	0,010	0,026	0,015	0,021	0,009	0,035	0,018	0,033	0,037
10%	0,054	0,058	0,015	0,051	0,034	0,020	0,008	0,059	0,036	0,092	0,064
20%	0,123	0,111	0,028	0,085	0,042	0,022	0,041	0,111	0,062	0,115	0,102
30%	0,171	0,163	0,028	0,124	0,058	0,036	0,098	0,175	0,112	0,164	0,140
40%	0,221	0,214	0,036	0,141	0,095	0,026	0,147	0,219	0,140	0,326	0,176
50%	0,298	0,277	0,046	0,196	0,114	0,040	0,234	0,281	0,162	0,337	0,207
60%	0,381	0,310	0,053	0,208	0,145	0,044	0,318	0,293	0,182	0,515	0,242
70%	0,457	0,385	0,062	0,237	0,177	0,045	0,446	0,397	0,250	0,556	0,267
80%	0,498	0,415	0,070	0,318	0,197	0,048	0,585	0,452	0,210	0,345	0,305
90%	0,526	0,452	0,076	0,307	0,247	0,050	0,737	0,518	0,202	0,457	0,309
WS - RANDOMLY MISSING											
2%	0,012	0,010	0,000	0,000	0,005	0,000	0,000	0,011	0,007	0,017	0,012
5%	0,030	0,025	0,000	0,000	0,013	0,000	0,002	0,028	0,018	0,035	0,019
10%	0,060	0,051	0,000	0,000	0,026	0,001	0,009	0,056	0,036	0,041	0,032
20%	0,119	0,100	0,000	0,001	0,052	0,001	0,037	0,113	0,072	0,193	0,057
30%	0,179	0,150	0,000	0,001	0,080	0,002	0,083	0,169	0,109	0,258	0,082
40%	0,240	0,201	0,000	0,001	0,108	0,003	0,146	0,225	0,145	0,249	0,106
50%	0,300	0,251	0,001	0,002	0,137	0,003	0,229	0,282	0,183	0,267	0,132
60%	0,360	0,302	0,001	0,002	0,169	0,004	0,330	0,339	0,220	0,342	0,157
70%	0,419	0,350	0,001	0,002	0,209	0,005	0,457	0,394	0,257	0,193	0,181
80%	0,481	0,401	0,002	0,002	0,259	0,005	0,588	0,451	0,297	0,513	0,205
90%	0,539	0,454	0,003	0,003	0,335	0,006	0,740	0,506	0,339	0,636	0,330

The KS test p-values for all imputation methods, missing rates, and experimental settings are presented. The best results, relative to each experimental setting, are highlighted in green.

KOLMOGOROV-SMIRNOV TEST P-VALUE

	Mean	Median	LI	CI	SMA	EWMA	ARIMA	MICE	KNN	GAIN	CAE
GHI - CONTINUOUSLY MISSING											
2%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,031	0,000	0,000
5%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
10%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
20%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
30%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
40%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
50%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
60%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
70%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
80%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
90%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
GHI - RANDOMLY MISSING											
2%	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,383	0,000	0,000
5%	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
10%	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
20%	0,000	0,000	0,998	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
30%	0,000	0,000	0,897	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
40%	0,000	0,000	0,387	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
50%	0,000	0,000	0,240	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
60%	0,000	0,000	0,020	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
70%	0,000	0,000	0,004	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
80%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
90%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
DNI - CONTINUOUSLY MISSING											
2%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
5%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
10%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
20%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
30%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
40%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
50%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
60%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
70%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
80%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
90%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
DNI - RANDOMLY MISSING											
2%	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
5%	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
10%	0,000	0,000	0,772	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
20%	0,000	0,000	0,056	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
30%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
40%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
50%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
60%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
70%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
80%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
90%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
WS - CONTINUOUSLY MISSING											
2%	0,000	0,000	0,003	0,000	0,000	0,000	0,001	0,000	0,000	0,000	0,000
5%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
10%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
20%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
30%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
40%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
50%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
60%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
70%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
80%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
90%	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
WS - RANDOMLY MISSING											
2%	0,000	0,000	1,000	1,000	0,000	1,000	1,000	0,000	0,000	0,000	0,000
5%	0,000	0,000	1,000	1,000	0,000	1,000	0,177	0,000	0,000	0,000	0,000
10%	0,000	0,000	1,000	1,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000
20%	0,000	0,000	1,000	1,000	0,000	0,725	0,000	0,000	0,000	0,000	0,000
30%	0,000	0,000	1,000	0,956	0,000	0,230	0,000	0,000	0,000	0,000	0,000
40%	0,000	0,000	1,000	0,754	0,000	0,067	0,000	0,000	0,000	0,000	0,000
50%	0,000	0,000	0,999	0,444	0,000	0,008	0,000	0,000	0,000	0,000	0,000
60%	0,000	0,000	0,958	0,219	0,000	0,001	0,000	0,000	0,000	0,000	0,000
70%	0,000	0,000	0,694	0,099	0,000	0,000	0,000	0,000	0,000	0,000	0,000
80%	0,000	0,000	0,236	0,106	0,000	0,000	0,000	0,000	0,000	0,000	0,000
90%	0,000	0,000	0,009	0,040	0,000	0,000	0,000	0,000	0,000	0,000	0,000

Appendix E. Energy system modeling results for the grid-connected self-sufficient building

See Fig. 9.

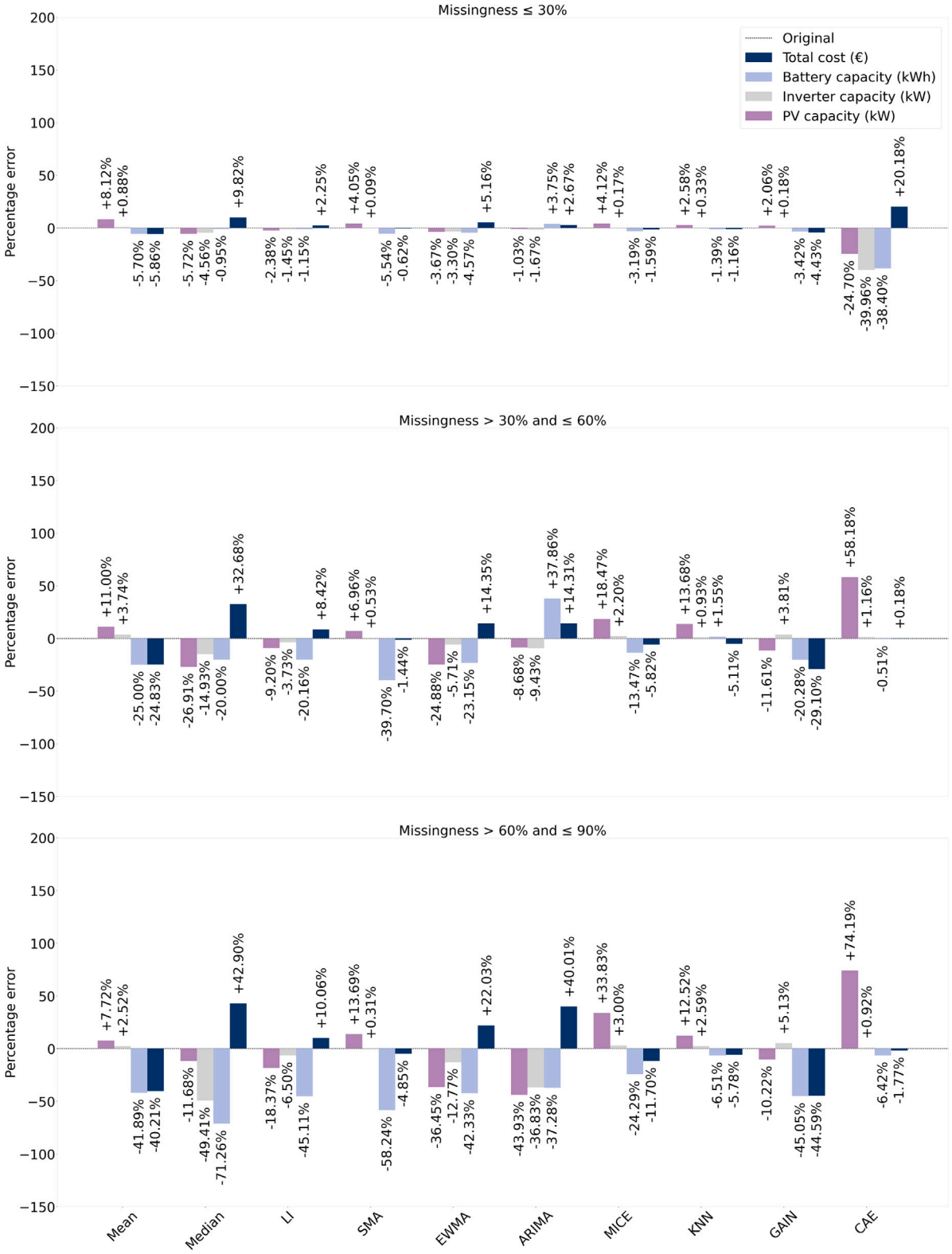

Fig. 9. Percentage error in total annualized costs and component capacities in the self-sufficient building (grid-connected system). Zero indicates the baseline corresponding to the results obtained using the original data, while the vertical bars represent the deviations resulting from imputed data. Results are grouped by missing rates (≤30%, >30% and ≤60%, >60% and ≤90%), and the error values shown are group averages.

Data availability

Data will be made available on request.

References

[1] Babatunde Olubayo Moses, Munda Josiah L, Hamam YJER. Power system flexibility: A review. Energy Rep 2020;6:101–6. http://dx.doi.org/10.1016/j.egyr.2019.11.048.

[2] Risch Stanley, Weinand Jann Michael, Schulze Kai, Vartak Sammit, Kleinebrahm Max, Pflugradt Noah, et al. Scaling energy system optimizations: Techno-economic assessment of energy autonomy in 11 000 German municipalities. Energy Convers Manage 2024;309:118422. http://dx.doi.org/10.1016/j.enconman.2024.118422.

[3] Quiñones Jhon J, Pineda Luis R, Ostanek Jason, Castillo Luciano. Towards smart energy management for community microgrids: Leveraging deep learning in probabilistic forecasting of renewable energy sources. Energy Convers Manage 2023;293:117440. http://dx.doi.org/10.1016/j.enconman.2023.117440.

[4] Alkhayat Ghadah, Mehmood Rashid. A review and taxonomy of wind and solar energy forecasting methods based on deep learning. Energy AI 2021;4:100060. http://dx.doi.org/10.1016/j.egyai.2021.100060.

[5] Jeong Dongyeon, Park Chiwoo, Ko Young Myoung. Missing data imputation using mixture factor analysis for building electric load data. Appl Energy 2021;304:117655. http://dx.doi.org/10.1016/j.apenergy.2021.117655.

[6] Fan Hang, Zhang Xuemin, Mei Shengwei. Wind power time series missing data imputation based on generative adversarial network. In: 2021 IEEE 4th international electrical and energy conference. IEEE; 2021, p. 1–6. http://dx.doi.org/10.1109/CIEEC50170.2021.9510923.

[7] Altan Aytaç, Karasu Seçkin, Zio Enrico. A new hybrid model for wind speed forecasting combining long short-term memory neural network, decomposition methods and grey wolf optimizer. Appl Soft Comput 2021;100:106996. http://dx.doi.org/10.1016/j.asoc.2020.106996.

[8] Kim Taeyoung, Ko Woong, Kim Jinho. Analysis and impact evaluation of missing data imputation in day-ahead PV generation forecasting. Appl Sci 2019;9(1):204. http://dx.doi.org/10.3390/app9010204.

[9] Mayer Martin János. Effects of the meteorological data resolution and aggregation on the optimal design of photovoltaic power plants. Energy Convers Manage 2021;241:114313. http://dx.doi.org/10.1016/j.enconman.2021.114313.

[10] Lin Wei-Chao, Tsai Chih-Fong. Missing value imputation: a review and analysis of the literature (2006–2017). Artif Intell Rev 2020;53:1487–509. http://dx.doi.org/10.1007/s10462-019-09709-4.

[11] Khan Prince Waqas, Byun Yung-Cheol, Lee Sang-Joon, Park Namje. Machine learning based hybrid system for imputation and efficient energy demand forecasting. Energies 2020;13(11):2681. http://dx.doi.org/10.3390/en13112681.

[12] Phan Quoc-Thang, Wu Yuan-Kang, Phan Quoc-Dung, Lo Hsin-Yen. A study on missing data imputation methods for improving hourly solar dataset. In: 2022 8th international conference on applied system innovation. IEEE; 2022, p. 21–4. http://dx.doi.org/10.1109/ICASI55125.2022.9774453.

[13] Zhang Wenjie, Luo Yonghong, Zhang Ying, Srinivasan Dipti. SolarGAN: Multivariate solar data imputation using generative adversarial network. IEEE Trans Sustain Energy 2020;12(1):743–6. http://dx.doi.org/10.1109/TSTE.2020.3004751.

[14] Liu Xin, Zhang Zijun. A two-stage deep autoencoder-based missing data imputation method for wind farm SCADA data. IEEE Sensors J 2021;21(9):10933–45. http://dx.doi.org/10.1109/JSEN.2021.3061109.

[15] Turrado Concepción Crespo, López María del Carmen Meizoso, Lasheras Fernando Sánchez, Gómez Benigno Antonio Rodríguez, Rollé José Luis Calvo, de Cos Juez Francisco Javier. Missing data imputation of solar radiation data under different atmospheric conditions. Sensors 2014;14(11):20382–99. http://dx.doi.org/10.3390/s141120382.

[16] Sun Yige, Li Jing, Xu Yifan, Zhang Tingting, Wang Xiaofeng. Deep learning versus conventional methods for missing data imputation: A review and comparative study. Expert Syst Appl 2023;227:120201. http://dx.doi.org/10.1016/j.eswa.2023.120201.

[17] Wang Zhenhua, Akande Olanrewaju, Poulos Jason, Li Fan. Are deep learning models superior for missing data imputation in large surveys? Evidence from an empirical comparison. 2021, http://dx.doi.org/10.48550/arXiv.2103.09316, arXiv preprint arXiv:2103.09316.

[18] Bülte Christopher, Kleinebrahm Max, Yilmaz Hasan Ümitcan, Gómez-Romero Juan. Multivariate time series imputation for energy data using neural networks. Energy AI 2023;13:100239. http://dx.doi.org/10.1016/j.egyai.2023.100239.

[19] Liguori Antonio, Markovic Romana, Ferrando Martina, Frisch Jérôme, Causone Francesco, van Treeck Christoph. Augmenting energy time-series for data-efficient imputation of missing values. Appl Energy 2023;334:120701. http://dx.doi.org/10.1016/j.apenergy.2023.120701.

[20] de Paz-Centeno Iván, García-Ordás María Teresa, García-Olalla Óscar, Alaiz-Moretón Héctor. Imputation of missing measurements in PV production data within constrained environments. Expert Syst Appl 2023;217:119510. http://dx.doi.org/10.1016/j.eswa.2023.119510.

[21] Wang Ming-Chang, Tsai Chih-Fong, Lin Wei-Chao. Towards missing electric power data imputation for energy management systems. Expert Syst Appl 2021;174:114743. http://dx.doi.org/10.1016/j.eswa.2021.114743.

[22] Park You-Jin, Fan Shu-Kai S, Hsu Chia-Yu. A review on fault detection and process diagnostics in industrial processes. Processes 2020;8(9):1123. http://dx.doi.org/10.3390/pr8091123.

[23] Khare Priyanshi, Wadhvani Rajesh, Shukla Sanyam. Missing data imputation for solar radiation using generative adversarial networks. In: Proceedings of international conference on computational intelligence: ICCI 2020. Springer; 2022, p. 1–14. http://dx.doi.org/10.1007/978-981-16-3802-2_1.

[24] Ma Jun, Cheng Jack CP, Jiang Feifeng, Chen Weiwei, Wang Mingzhu, Zhai Chong. A bi-directional missing data imputation scheme based on LSTM and transfer learning for building energy data. Energy Build 2020;216:109941. http://dx.doi.org/10.1016/j.enbuild.2020.109941.

[25] Chen Xinyi, Liu Yinan, Shen Yu, Zhang Kanjian, Wei Haikun. A data interpolation method for missing irradiance data of photovoltaic power station. In: 2020 Chinese automation congress. IEEE; 2020, p. 4735–40. http://dx.doi.org/10.1109/CAC51589.2020.9326730.

[26] Jung Seungwon, Moon Jihoon, Park Sungwoo, Rho Seungmin, Baik Sung Wook, Hwang Eenjun. Bagging ensemble of multilayer perceptrons for missing electricity consumption data imputation. Sensors 2020;20(6):1772. http://dx.doi.org/10.3390/s20061772.

[27] Daut I, Irwanto M, Irwan YM, Gomesh N, Ahmad NS. Combination of Hargreaves method and linear regression as a new method to estimate solar radiation in Perlis, Northern Malaysia. Sol Energy 2011;85(11):2871–80. http://dx.doi.org/10.1016/j.solener.2011.08.026.

[28] Layanun Vichaya, Suksamosorn Supachai, Songsiri Jitkomut. Missing-data imputation for solar irradiance forecasting in Thailand. In: 2017 56th annual conference of the society of instrument and control engineers of Japan. IEEE; 2017, p. 1234–9. http://dx.doi.org/10.23919/SICE.2017.8105472.

[29] Qu Fuming, Liu Jinhai, Ma Yanjuan, Zang Dong, Fu Mingrui. A novel wind turbine data imputation method with multiple optimizations based on GANs. Mech Syst Signal Process 2020;139:106610. http://dx.doi.org/10.1016/j.ymssp.2019.106610.

[30] Zhang Chu, Wang Yuhan, Fu Yongyan, Qiao Xiujie, Nazir Muhammad Shahzad, Peng Tian. A novel DWTimesNet-based short-term multi-step wind power forecasting model using feature selection and auto-tuning methods. Energy Convers Manage 2024;301:118045. http://dx.doi.org/10.1016/j.enconman.2023.118045.

[31] Kim Jung-Woo, Pachepsky Yakov A. Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. J Hydrol 2010;394(3–4):305–14. http://dx.doi.org/10.1016/j.jhydrol.2010.09.005.

[32] Garnier Antoine, Eynard Julien, Caussanel Matthieu, Grieu Stéphane. Missing data estimation for energy resources management in tertiary buildings. In: CCCA12. IEEE; 2012, p. 1–6. http://dx.doi.org/10.1109/CCCA.2012.6417902.

[33] Shukur Osamah Basheer, Lee Muhammad Hisyam. Imputation of missing values in daily wind speed data using hybrid AR-ANN method. Mod Appl Sci 2015;9(11):1. http://dx.doi.org/10.5539/mas.v9n11p1.

[34] Rahman Aowabin, Srikumar Vivek, Smith Amanda D. Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks. Appl Energy 2018;212:372–85. http://dx.doi.org/10.1016/j.apenergy.2017.12.051.

[35] Goodfellow Ian, Pouget-Abadie Jean, Mirza Mehdi, Xu Bing, Warde-Farley David, Ozair Sherjil, et al. Generative adversarial nets. Adv Neural Inf Process Syst 2014;27. https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

[36] Yoon Jinsung, Jordon James, Schaar Mihaela. GAIN: Missing data imputation using generative adversarial nets. In: International conference on machine learning. PMLR; 2018, p. 5689–98, https://proceedings.mlr.press/v80/yoon18a.html.

[37] Liguori Antonio, Markovic Romana, Dam Thi Thu Ha, Frisch Jérôme, van Treeck Christoph, Causone Francesco. Indoor environment data time-series reconstruction using autoencoder neural networks. Build Environ 2021;191:107623. http://dx.doi.org/10.1016/j.buildenv.2021.107623.

[38] Başakın Eyyup Ensar, Ekmekcioğlu Ömer, Özger Mehmet. Developing a novel approach for missing data imputation of solar radiation: A hybrid differential evolution algorithm based extreme gradient boosting model. Energy Convers Manage 2023;280:116780. http://dx.doi.org/10.1016/j.enconman.2023.116780.

[39] Hussain Syed Nazir, Abd Aziz Azlan, Hossen Md Jakir, Ab Aziz Nor Azlina, Murthy G Ramana, Mustakim Fajaruddin Bin. A novel framework based on CNN-LSTM neural network for prediction of missing values in electricity consumption time-series datasets. J Inf Process Syst 2022;18(1):115–29, http://gnanaganga.inflibnet.ac.in:8080/jspui/handle/123456789/15039.

[40] Demirhan Haydar, Renwick Zoe. Missing value imputation for short to mid-term horizontal solar irradiance data. Appl Energy 2018;225:998–1012. http://dx.doi.org/10.1016/j.apenergy.2018.05.054.

[41] Zainudin Mohd Lutfi, Saaban Azizan, Bakar Mohd Nazari Abu. Estimation of missing values in solar radiation data using piecewise interpolation methods: Case study at Penang city. In: AIP conference proceedings, vol. 1691, no. 1, AIP Publishing; 2015, http://dx.doi.org/10.1063/1.4937079.

[42] Bayen Alexandre M, Siauw Timmy. Chapter 14 - Interpolation. In: An introduction to MATLAB® programming and numerical methods for engineers. Boston: Academic Press; 2015, p. 211–23. http://dx.doi.org/10.1016/B978-0-12-420228-3.00014-2.

[43] Denhard Alexis, Bandyopadhyay Soutir, Habte Aron, Sengupta Manajit. Evaluation of time-series gap-filling methods for solar irradiance applications. Tech. rep., Golden, CO, United States: National Renewable Energy Lab. (NREL); 2021, http://dx.doi.org/10.2172/1826664.

[44] Ellis Craig A, Parbery Simon A. Is smarter better? A comparison of adaptive, and simple moving average trading strategies. Res Int Bus Financ 2005;19(3):399–411. http://dx.doi.org/10.1016/j.ribaf.2004.12.009.

[45] Velasco-Gallego Christian, Lazakis Iraklis. Real-time data-driven missing data imputation for short-term sensor data of marine systems. A comparative study. Ocean Eng 2020;218:108261. http://dx.doi.org/10.1016/j.oceaneng.2020.108261.

[46] Kotu Vijay, Deshpande Bala. Time series forecasting. Data Sci 2019;2:395–445. http://dx.doi.org/10.1016/B978-0-12-801460-8.00010-0.

[47] Moons Karel GM, Donders Rogier ART, Stijnen Theo, Harrell Jr Frank E. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol 2006;59(10):1092–101. http://dx.doi.org/10.1016/j.jclinepi.2006.01.009.

[48] Troyanskaya Olga, Cantor Michael, Sherlock Gavin, Brown Pat, Hastie Trevor, Tibshirani Robert, et al. Missing value estimation methods for DNA microarrays. Bioinformatics 2001;17(6):520–5. http://dx.doi.org/10.1093/bioinformatics/17.6.520.

[49] Sengupta Manajit, Xie Yu, Lopez Anthony, Habte Aron, Maclaurin Galen, Shelby James. The national solar radiation data base (NSRDB). Renew Sustain Energy Rev 2018;89:51–60. http://dx.doi.org/10.1016/j.rser.2018.03.003.

[50] Omoyele Olalekan, Matrone Silvana, Hoffmann Maximilian, Ogliari Emanuele, Weinand Jann Michael, Leva Sonia, et al. Impact of temporal resolution on the design and reliability of residential energy systems. Energy Build 2024;319:114411. http://dx.doi.org/10.1016/j.enbuild.2024.114411.

[51] Omoyele Olalekan, Hoffmann Maximilian, Koivisto Matti, Larraneta Miguel, Weinand Jann Michael, Linßen Jochen, et al. Increasing the resolution of solar and wind time series for energy system modeling: A review. Renew Sustain Energy Rev 2024;189:113792. http://dx.doi.org/10.1016/j.rser.2023.113792.

[52] Larrañeta Miguel, Cantón-Marín Carlos, Silva-Pérez Manuel Antonio, Lillo-Bravo Isidoro. Use of the ND tool: An open tool for the synthetic generation of 1-min solar data from hourly means with geographic flexibility. In: AIP conference proceedings, vol. 2445, no. 1, AIP Publishing; 2022, http://dx.doi.org/10.1063/5.0085901.

[53] Larrañeta Miguel, Fernandez-Peruchena C, Silva-Pérez Manuel Antonio, Lillo-Bravo Isidoro. Methodology to synthetically downscale DNI time series from 1-h to 1-min temporal resolution with geographic flexibility. Sol Energy 2018;162:573–84. http://dx.doi.org/10.1016/j.solener.2018.01.064.

[54] Omoyele Olalekan, Hoffmann Maximilian, Weinand Jann Michael, Larraneta Miguel, Linßen Jochen, Stolten Detlef. A high-resolution downscaling approach for solar irradiance using statistical parameter matching. 2025, http://dx.doi.org/10.2139/ssrn.5222834, Available at SSRN 5222834.

[55] DTU Wind Energy. Correlations in renewable energy sources (CorRES), URL: https://corres.windenergy.dtu.dk/.

[56] Koivisto Matti, Jónsdóttir Guðrún Margrét, Sørensen Poul, Plakas Konstantinos, Cutululis Nicolaos. Combination of meteorological reanalysis data and stochastic simulation for modelling wind generation variability. Renew Energy 2020;159:991–9. http://dx.doi.org/10.1016/j.renene.2020.06.033.

[57] Singla Pardeep, Duhan Manoj, Saroha Sumit. Different normalization techniques as data preprocessing for one step ahead forecasting of solar global horizontal irradiance. In: Artificial intelligence for renewable energy systems. Elsevier; 2022, p. 209–30. http://dx.doi.org/10.1016/B978-0-323-90396-7.00004-3.

[58] Espinar Bella, Ramírez Lourdes, Drews Anja, Beyer Hans Georg, Zarzalejo Luis F, Polo Jesús, et al. Analysis of different comparison parameters applied to solar radiation data from satellite and german radiometric stations. Sol Energy 2009;83(1):118–25. http://dx.doi.org/10.1016/j.solener.2008.07.009.

[59] Knosala Kevin, Kotzur Leander, Röben Fritz TC, Stenzel Peter, Blum Ludger, Robinius Martin, et al. Hybrid hydrogen home storage for decentralized energy autonomy. Int J Hydrog Energy 2021;46(42):21748–63. http://dx.doi.org/10.1016/j.ijhydene.2021.04.036.

[60] Kleinebrahm Max, Weinand Jann Michael, Naber Elias, McKenna Russell, Ardone Armin, Fichtner Wolf. Two million European single-family homes could abandon the grid by 2050. Joule 2023;7(11):2485–510. http://dx.doi.org/10.1016/j.joule.2023.09.012.

[61] Weinand Jann Michael, Ried Sabrina, Kleinebrahm Max, McKenna Russell, Fichtner Wolf. Identification of potential off-grid municipalities with 100% renewable energy supply for future design of power grids. IEEE Trans Power Syst 2022;37(4):3321–30. http://dx.doi.org/10.1109/TPWRS.2020.3033747.

[62] Weinand Jann Michael, Hoffmann Maximilian, Göpfert Jan, Terlouw Tom, Schönau Julian, Kuckertz Patrick, et al. Global LCOEs of decentralized off-grid renewable energy systems. Renew Sustain Energy Rev 2023;183:113478. http://dx.doi.org/10.1016/j.rser.2023.113478.

[63] Gstöhl Ursin, Pfenninger Stefan. Energy self-sufficient households with photovoltaics and electric vehicles are feasible in temperate climate. PloS One 2020;15(3):e0227368. http://dx.doi.org/10.1371/journal.pone.0227368.

[64] Kotzur Leander, Markewitz Peter, Robinius Martin, Stolten Detlef. Kostenoptimale Versorgungssysteme für ein vollautarkes Einfamilienhaus. Int Energiewirtschaftstagung 2017;10:1–14.

[65] Omoyele Olalekan, Hoffmann Maximilian, Weinand Jann Michael, Stolten Detlef. Accelerating computational efficiency in sub-hourly renewable energy systems modeling. 2024, http://dx.doi.org/10.2139/ssrn.5004752, Available at SSRN 5004752.

[66] Hoffmann Maximilian, Schyska Bruno U, Bartels Julian, Pelser Tristan, Behrens Johannes, Wetzel Manuel, et al. A review of mixed-integer linear formulations for framework-based energy system models. Adv Appl Energy 2024;100190. http://dx.doi.org/10.1016/j.adapen.2024.100190.

[67] Welder Lara, Ryberg D Severin, Kotzur Leander, Grube Thomas, Robinius Martin, Stolten Detlef. Spatio-temporal optimization of a future energy system for power-to-hydrogen applications in Germany. Energy 2018;158:1130–49. http://dx.doi.org/10.1016/j.energy.2018.05.059.

[68] Klütz Theresa, Knosala Kevin, Behrens Johannes, Maier Rachel, Hoffmann Maximilian, Pflugradt Noah, et al. ETHOS.FINE: A framework for integrated energy system assessment. J Open Source Softw 2025;10(105):6274. http://dx.doi.org/10.21105/joss.06274.

[69] Boriratrit Sarunyoo, Fuangfoo Pradit, Srithapon Chitchai, Chatthaworn Rongrit. Adaptive meta-learning extreme learning machine with golden eagle optimization and logistic map for forecasting the incomplete data of solar irradiance. Energy AI 2023;13:100243. http://dx.doi.org/10.1016/j.egyai.2023.100243.

[70] Mohamad Noor Bariah, Lai An-Chow, Lim Boon-Han. A case study in the tropical region to evaluate univariate imputation methods for solar irradiance data with different weather types. Sustain Energy Technol Assess. 2022;50:101764. http://dx.doi.org/10.1016/j.seta.2021.101764.

[71] Yelchuri Srinath, Rangaraj AG, Xie Yu, Habte Aron, Joshi Mohit Chandra, Boopathi K, et al. A short-term solar forecasting platform using a physics-based smart persistence model and data imputation method. Tech. rep., Golden, CO, United States: National Renewable Energy Lab. (NREL); 2021, http://dx.doi.org/10.2172/1837967.

[72] Shen Meng, Zhang Huaizheng, Cao Yixin, Yang Fan, Wen Yonggang. Missing data imputation for solar yield prediction using temporal multi-modal variational auto-encoder. In: Proceedings of the 29th ACM international conference on multimedia. 2021, p. 2558–66. http://dx.doi.org/10.1145/3474085.3475430.

[73] Flores Anibal, Paxi-Apaza Walter, Clares-Perca Juan. CBRi2: Imputation of solar radiation time series with case based reasoning. In: 2021 IEEE XXVIII international conference on electronics, electrical engineering and computing. IEEE; 2021, p. 1–4. http://dx.doi.org/10.1109/INTERCON52678.2021.9532750.

[74] Ho Kah-Ching, Lim Boon-Han, Lai An-Chow. Recovery of the solar irradiance data using artificial neural network. In: IOP conference series: Earth and environmental science, vol. 721, no. 1, IOP Publishing; 2021, 012006, https://iopscience.iop.org/article/10.1088/1755-1315/721/1/012006.

[75] Lindig Sascha, Louwen Atse, Moser David, Topic Marko. Outdoor PV system monitoring—input data quality, data imputation and filtering approaches. Energies 2020;13(19):5099. http://dx.doi.org/10.3390/en13195099.

[76] Zhao Yuliang, Ge Dehui, Huang Shufan, Zhou Hui, Peng Chuning, Wang Qi, et al. An intelligent imputation method for electricity data. In: 2020 international conference on intelligent computing, automation and systems. IEEE; 2020, p. 12–8. http://dx.doi.org/10.1109/ICICAS51530.2020.00011.

[77] Li Hui, Chen Xin, Shan Mingzhu, Duan Peiyong. Missing data filling methods of air-conditioning power consumption for public buildings. In: 2020 39th Chinese control conference. IEEE; 2020, p. 3183–7. http://dx.doi.org/10.23919/CCC50068.2020.9188857.

[78] Sánchez-Gómez Claudia, Velázquez Ramiro. Analysis of wind missing data for wind farms in Isthmus of Tehuantepec. OPENAIRE 2019. http://dx.doi.org/10.1109/ROPEC.2018.8661457.

[79] Kim Minkyung, Park Sangdon, Lee Joohyung, Joo Yongjae, Choi Jun Kyun. Learning-based adaptive imputation method with kNN algorithm for missing power data. Energies 2017;10(10):1668. http://dx.doi.org/10.3390/en10101668.

[80] Jurado Sergio, Nebot Àngela, Mugica Fransisco, Mihaylov Mihail. Fuzzy inductive reasoning forecasting strategies able to cope with missing data: A smart grid application. Appl Soft Comput 2017;51:225–38. http://dx.doi.org/10.1016/j.asoc.2016.11.040.

[81] Peppanen Jouni, Zhang Xiaochen, Grijalva Santiago, Reno Matthew J. Handling bad or missing smart meter data through advanced data imputation. In: 2016 IEEE power & energy society innovative smart grid technologies conference. IEEE; 2016, p. 1–5. http://dx.doi.org/10.1109/ISGT.2016.7781213.

[82] Saaban Azizan, Zainudin Lutfi, Bakar Mohd Nazari Abu. On piecewise interpolation techniques for estimating solar radiation missing values in Kedah. In: AIP conference proceedings, vol. 1635, no. 1, American Institute of Physics; 2014, p. 217–21. http://dx.doi.org/10.1063/1.4903586.

[83] Kasam Alisha A, Lee Benjamin D, Paredis Christiaan JJ. Statistical methods for interpolating missing meteorological data for use in building simulation. In: Building simulation, vol. 7, Springer; 2014, p. 455–65. http://dx.doi.org/10.1007/s12273-014-0174-7.

[84] Ogunsola Oluwaseyi T, Song Li. Restoration of long-term missing gaps in solar radiation. Energy Build 2014;82:580–91. http://dx.doi.org/10.1016/j.enbuild.2014.07.088.

[85] Yozgatligil Ceylan, Aslan Sipan, Iyigun Cem, Batmaz Inci. Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. Theor Appl Climatol 2013;112:143–67. http://dx.doi.org/10.1007/s00704-012-0723-x.

[86] Teegavarapu Ramesh SV, Tufail Mohammad, Ormsbee Lindell. Optimal functional forms for estimation of missing precipitation data. J Hydrol 2009;374(1–2):106–15. http://dx.doi.org/10.1016/j.jhydrol.2009.06.014.

[87] Teegavarapu Ramesh SV, Chandramouli Viswanathan. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. J Hydrol 2005;312(1–4):191–206. http://dx.doi.org/10.1016/j.jhydrol.2005.02.015.

[88] Jin Zhou, Yezheng Wu, Gang Yan. A stochastic method to generate bin weather data in Nanjing, China. Energy Convers Manage 2006;47(13–14):1843–50. http://dx.doi.org/10.1016/j.enconman.2005.10.006.