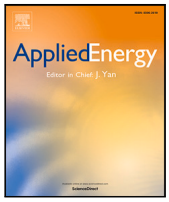
Applied Energy 356 (2024) 122434

Contents lists available at ScienceDirect

Applied Energy

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



Can end-to-end data-driven models outperform traditional semi-physical models in separating 1-min irradiance?

Yinghao Chu [a], Dazhi Yang [b,*], Hanxin Yu [a], Xin Zhao [c], Mengying Li [d,**]

[a] Department of Systems Engineering, City University of Hong Kong, Hong Kong Special Administrative Region of China
[b] School of Electrical Engineering and Automation, Harbin Institute of Technology, Harbin, Heilongjiang, China
[c] Flare & Co Information Co., Ltd., China
[d] Department of Mechanical Engineering & Research Institute for Smart Energy, The Hong Kong Polytechnic University, Hong Kong Special Administrative Region of China

ARTICLE INFO

Keywords:
Solar radiation modeling
Separation modeling
Diffuse radiation
Benchmarking data
Data-driven models

ABSTRACT

As a crucial component of the model chain, which facilitates irradiance-to-power conversion during solar resource assessment and forecasting, separation modeling continues to draw attention in both academia and industry. However, when evaluating even the best separation model today, one can quickly recognize its limited accuracy compared to other energy meteorology models such as transposition models. The task of separating global horizontal irradiance into diffuse and beam components does not seem soluble by any derivative effort aimed at tweaking the existing semi-physical models. As a result, an appealing alternative is to consider end-to-end data-driven models, which have demonstrated predictive capability in scenarios where the volume of data is substantial and the interaction among variables is complex. This work discusses the separation of 1-min irradiance from a data-driven perspective. In this preliminary study, a total of 10 representative data-driven separation models are developed and compared to the state-of-the-art semi-physical models, using a comprehensive 1-min irradiance database that spans five years and covers numerous climate types. The average error of the data-driven models is found to be 15.2% to 22.6% lower than that of the semi-physical models for training locations and 7.9% to 17.6% lower for completely unseen locations. Data-driven models also have significantly lower standard deviations (up to 87.2% even for completely unseen locations), highlighting their robustness. In addition, this work provides a guideline for choosing between data-driven and semi-physical models based on data availability, application needs, computational resources, interpretability, and model adaptability. Furthermore, the study underscores the challenges in accurately predicting the diffuse fraction using available input features and indicates that the incorporation of additional weather-related variables and domain knowledge could enhance the performance of data-driven separation models.

1. Introduction

Separation models, or decomposition models, divide global horizontal irradiance (GHI, G_h) into two additive components: beam horizontal irradiance (BHI, B_h) and diffuse horizontal irradiance (DHI, D_h). These models estimate the diffuse fraction k using G_h and other geological and meteorological parameters, and then, with the estimated k, calculate D_h and B_h values via:

$$\hat{D}_h = \hat{k}\, G_h, \tag{1}$$

$$\hat{B}_h = G_h - \hat{D}_h. \tag{2}$$

Since BHI is related to the more frequently used beam normal irradiance (BNI, B_n) through the cosine of the solar zenith angle, obtaining B_n is trivial if D_h or B_h is known, i.e., $\hat{B}_n = \hat{B}_h / \cos(Z)$.

Separation modeling plays a pivotal role in solar resource assessment [1,2] and forecasting [3,4], which are the two most critical domains of solar energy meteorology. For instance, photovoltaic (PV) or concentrated solar power (CSP) power forecast submission is mandated by many grid operators [5,6]. However, solar forecasting research primarily focuses on the forecasting of solar irradiance. This underscores the significance of irradiance-to-power conversion methodologies using a physical model chain, which is a cascade use of several energy meteorology models, where the output of a preceding model serves as the

Abbreviations: BNI, beam normal irradiance; DHI, diffuse horizontal irradiance; GHI, global horizontal irradiance; QC, quality control

* Corresponding author.
** Corresponding author.
E-mail addresses: yangdazhi.nus@gmail.com (D. Yang), mengying.li@polyu.edu.hk (M. Li).

https://doi.org/10.1016/j.apenergy.2023.122434
Received 26 July 2023; Received in revised form 14 November 2023; Accepted 27 November 2023
Available online 8 December 2023

input to the next, until the AC power is obtained [7]. Given the fact that many numerical weather prediction (NWP) models produce just GHI forecasts, separation models become indispensable for acquiring DHI and BNI forecasts, thereby enabling PV and CSP power forecasts [8–10]. Consequently, separation modeling has garnered substantial attention in energy meteorology research and practical applications for several decades, culminating in an array of over 150 distinct model options.

Despite the large number of models, they can be collected into two categories. Physics-based separation models, such as Hollands1 [11] or Hollands2 [12],[1] are exceptionally rare due to the difficulty in creating an effective surrogate representation of the complex radiative-transfer process under all-sky conditions. In contrast, a vast majority of models are semi-physical (or equivalently, semi-empirical), aiming to establish suitable mathematical relationships between diffuse and global irradiance components. Semi-physical models offer flexibility, ease of construction, computational efficiency, and adaptability to local irradiance regimes, which jointly explain to a large extent their popularity.

Surveying the literature, it is evident that several comprehensive assessments have been conducted regarding the performance of separation models. In a pioneering study, Gueymard and Ruiz-Arias [13] validated 140 models using 1-min data from 54 research-grade radiometry stations, consisting of a total of 25 million quality-controlled data points. The study identified Engerer2 [14] as the best-performing model, largely attributed to its innovative consideration of the cloud-enhancement effect which significantly enhanced its predictive prowess. Since then, Engerer2 has become a benchmark that subsequent models aspire to surpass. Later, a comprehensive follow-up review was conducted by Yang [1]. That review compared 10 major models proposed after 2016, using an even larger dataset from 126 world-wide research-grade radiometry stations, encompassing over 80 million quality-controlled 1-min data points. While several later-introduced models managed to outperform Engerer2, Yang4 [15] emerged as the best-performing model. Yang4 employs a novel temporal-cascade modeling strategy that effectively captures the low-frequency variations in the diffuse fraction. Given the extensive dataset used by Yang [1], Yang4 can be confidently regarded as the most accurate 1-min separation model to date.[2]

In addition to the incorporation of the cloud-enhancement effect and low-frequency variation components, much of the success of Engerer2 and Yang4 can be attributed to their functional form that is the logistic function. More precisely, since separation modeling aims to predict the diffuse fraction (k, the ratio of DHI and GHI, ranging from 0 to 1) using the clearness index (k_t, the ratio of GHI and extraterrestrial GHI, usually ranging from 0 to 1) along with other auxiliary variables, the logistic function aptly mirrors the shape of the k_t–k relationship. Mathematically,

$$k = \frac{1}{1 + e^{\beta_1 k_t + \beta_0}}, \tag{3}$$

where β_0 and β_1 are the model coefficients that need to be identified empirically, i.e., via fitting. Given the univariate nature of Eq. (3), which can only model injective relationship (i.e., one-to-one mapping), both Engerer2 and Yang4 sought to include additional predictors into the exponent, such that the non-injective mapping between k_t and k can be better apprehended. Visually, Fig. 1 shows the k_t–k scatter using some sample data collected at Carpentras (44.083°N, 5.059°E), France, alongside the predictions made using the logistic function, Engerer2, and Yang4. The rule-of-thumb in interpreting plots of this sort is that the prediction scatter should sufficiently cover the background scatter, the more precisely covered the better—in this regard, the advantage of Yang4 is immediately evident.

In addition to Engerer2 and Yang4, an expansive array of separation models has been proposed, and many of them are also based on the logistic function. These models incorporate diverse novel adaptations, such as regime switching [17] or piecewise modeling [18]. Regrettably, as elucidated by the comprehensive reviews of Gueymard and Ruiz-Arias [13], Yang [1], none of these semi-physical models was able to revolutionize separation modeling in terms of accuracy. For instance, the normalized root mean square error (nRMSE) for transposition models, another significant category within the realm of energy meteorology models, can often descend to as low as a few percent [19]. Conversely, the RMSE for separation models rarely falls below 10%, with a range of 20%–40% being commonplace. Given these observations, it is posited that further modifications of existing semi-physical models are unlikely to engender considerably greater success than that already achieved by Engerer2 and Yang4. Therefore, it becomes necessary to explore alternative modeling strategies that diverge fundamentally from the principles we have grown accustomed to. Machine learning may well provide such an opportunity.

Applying machine learning (or data-driven) methods to solve engineering problems has long been regarded by the scientific community as a trivial task. What separates the good applications from those not-so-good ones is often how well the domain knowledge can be built into the data-driven model or whether the construct of the problem warrants the use of machine learning. The former issue does not arouse concern here, since half a century of experience on separation modeling should provide a decent amount of domain knowledge in terms of the provision of input features. (Although many deep-learning models do not require feature engineering, it is not in conflict with having meaningful features before learning.) On the other hand, regarding the conceptual construct, the intuitive linkage between existing logistic-function-based modeling and neural networks becomes evident when acknowledging that the logistic function is analogous to a single-neuron neural network employing a sigmoid activation function.

The logistic function resembles a single artificial neuron with a sigmoid activation function, which can be mathematically defined as:

$$k = g(z) = \frac{1}{1 + e^{-z}}, \tag{4}$$

where z is the weighted sum of meteorological inputs, that is,

$$z = h(\boldsymbol{x}) = \sum_{i=1}^{n} w_i x_i, \tag{5}$$

where w_i are the weights, x_i are the features, such as the clearness index k_t, solar zenith angle, cloud-enhancement quantifier, or the low-frequency variability index. Therefore, it is intuitively to write Eq. (4) as a composite function:

$$k = g \circ h(\boldsymbol{x}) = f(\boldsymbol{x}), \tag{6}$$

and develop data-driven models to directly map meteorological input features x to k using end-to-end learning approaches.

In short, the novelty and contribution of this work encompass the following three aspects. First, it explores the possibility of employing data-driven models, including advanced deep-learning models, to develop separation models via end-to-end learning. In comparison with conventional semi-physical models, data-driven models present notable advantages, such as the automatic learning of intricate relationships between input variables and generating predictions, or obviating the need for manual feature engineering. Second, a comprehensive evaluation of the performance of the 10 representative data-driven models is performed and compared against that of 10 state-of-the-art semi-physical models from the literature, on a comprehensive 1-min irradiance database covering many climate types. Since separation

[1] It is customary in the separation modeling literature to denote a model by the inventor's last name in Small Caps; when multiple models are proposed by the same person, a number is added.

[2] In parallel to this work, another work on separation modeling was published by [16], who proposed Yang5, which is not discussed in this paper for brevity.

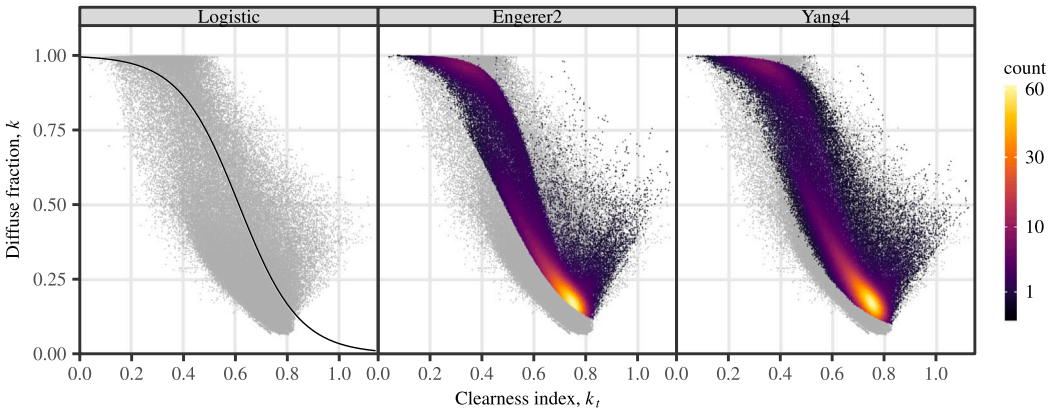


Fig. 1. One-minute diffuse fraction prediction using the logistic function, Engerer2, and Yang4 models, using data from Carpentras (44.083°N, 5.059°E), France, over a period of one year. Measurements are shown as the gray background, and predictions are shown as colored scatter.

models are often criticized for their low transferability, for models fitted at one location behave poorly elsewhere, the performance evaluation is conducted not just at training locations but also at unseen locations. Third, this work delves into the challenges intrinsic to further enhance the performance of data-driven approaches in separation modeling. In that, it contributes to a broader understanding of the capabilities and limitations of data-driven models within the domain, fostering confidence in their applicability to real-world challenges.

The results and implications of this work are profound. This study underscores the significance of adopting contemporary data-driven techniques within the constantly evolving realm of separation modeling. By showcasing the potential of data-driven methodologies and elucidating their strengths and weaknesses, this research advocates a selection guideline between data-driven and semi-empirical models. In terms of predictive accuracy, the average error of the data-driven models is found to be 15.2% to 22.6% lower than that of the semi-physical models on testing datasets from training locations and 7.9% to 17.6% lower on datasets from completely unseen locations. Data-driven models also exhibit significantly lower standard deviations (up to 87.2% even for completely unseen locations), highlighting their robustness and adaptability. Last but not least, through an in-depth investigation of the existing challenges, this study aids in pinpointing areas where additional research and development are requisite for fully exploiting the potential of data-driven models in separation modeling. To further enhance data-driven models, recommendations, such as incorporating domain knowledge, using additional exogenous variables or features, or ensemble modeling, are made. This inquiry not only furthers the refinement of data-driven separation modeling but also contributes to the establishment of best practices and guidelines for their implementation in the field.

The following pages are divided into five sections. The data used in this work is introduced in Section 2. Section 3 presents the methodology used in this work. The experiment and result are discussed in Section 4, followed by the conclusions in Section 5.

2. Data

Evaluating separation models, as well as other radiation models, requires access to high-quality, research-grade, ground-based radiometry data, which has hitherto been limited in availability. The Baseline Surface Radiation Network (BSRN) is the largest and most prominent network of radiometry stations, comprising about 60 active stations [20]. In addition to BSRN, there are smaller networks managed by independent organizations, such as the Bureau of Meteorology (Australia), the National Renewable Energy Laboratory (United States), or the Southern African Universities Radiometric Network (South Africa and several neighboring countries). Recognizing the need for comprehensive data, members from the International Energy Agency (IEA) PVPS Task 16 Activity 1.4 have collected and compiled several years of data from a total of 126 sites to support various solar energy meteorology research endeavors [21]. All of these stations employ thermopile pyranometers for G_h and D_h measurements and tracker-mounted thermopile pyrheliometers for B_n measurements, adhering to the recommended radiometry standards and undergoing regular calibration.

Due to propriety reasons and the constraints of available computational resources, this study employs 1-min data from 12 selected sites (as shown in Fig. 2), spanning five years from 2015 to 2020. Among these sites, seven are utilized for both model training and validation, while the remaining five sites in a different continent are reserved exclusively for model assessment in order to evaluate the performance of the models under unseen data instances. This approach ensures a comprehensive and rigorous evaluation of the models, taking into account their adaptability and generalization capabilities, which are critical aspects of academic research and real-world applications.

To ensure the utilization of the highest-quality data for comparison purposes, a stringent quality control (QC) process, including both automatic tests and manual screening, was implemented and carried out by the IEA members, as described by Forstinger et al. [21]. The entire QC process can be divided into four distinct quantitative stages, with each stage comprising several filters or tests. Specifically, if a data point fails to pass any of the tests, it is flagged and subsequently excluded from further analysis. Conversely, if a data point successfully passes all tests, or when the tests cannot be performed due to inapplicable conditions, it is considered "usable" for the analysis. This rigorous approach to data quality management ensures that the data entering the models is reliable and contributes to the validity of any conclusion made. More details of the QC process can also be found in [1]. In addition, two more filters are employed in this work: (1) discarding data instances when $Z > 85°$, and (2) discarding data instances when $G_h, B_n, D_h < 0$. This is due to the significant errors in both the radiometry data and various separation models under low-sun conditions [13].

3. Methodology

3.1. Problem formulation

The hypothesis of this work is that data-driven models, utilizing meteorological inputs, can be employed to separate global horizontal irradiance into its diffuse and beam components. In the training phase, let x_{ij} denote an input feature, with $i = 1, \ldots, n$ indexing the training samples and $j = 1, \ldots, m$ indexing the features, such as the solar zenith angle or clearness index, see Table 1 for a list. Thus, for the ith sample,

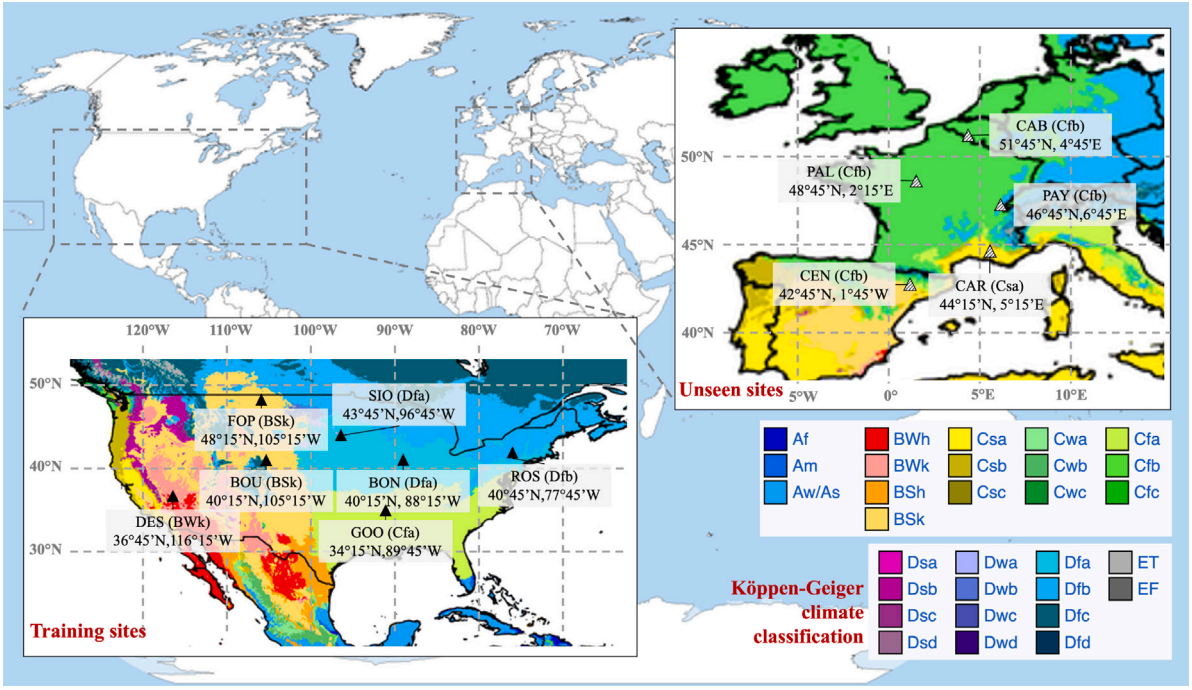


Fig. 2. The geographical distribution of 12 selected sites (triangle symbols) within the most up-to-date Köppen–Geiger climate classification system [22].

one can write the vector $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{im})$. When all n samples are framed into a matrix, one has

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ x_{21} & \cdots & x_{2m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix}, \tag{7}$$

which is the feature matrix that is to be used for training. The target vector is a length-n vector holding the corresponding diffuse fractions, that is,

$$\boldsymbol{k} = \left(k_1, \ldots, k_n\right)^{\top}. \tag{8}$$

With the above notation, the objective of this work is to develop end-to-end separation models, denoted as f, such that the diffuse fraction at a future index $n+1$ can be predicted by the corresponding input at that index, that is,

$$\hat{k}_{n+1} = f(\boldsymbol{x}_{n+1} | \boldsymbol{X}; \boldsymbol{\theta}), \tag{9}$$

where $\boldsymbol{\theta}$ denote a vector of model parameters. Once $\hat{k}_{n+1}$ is estimated, the prediction of D_h and B_n follows. Mathematically, $\hat{D}_{h,n+1} = \hat{k}_{n+1} G_{h,n+1}$ and $\hat{B}_{n,n+1} = \left(G_{h,n+1} - \hat{D}_{h,n+1}\right)/\cos Z_{n+1}$.

In constructing various versions of f (i.e., different data-driven models), four assumptions/notes are made. First, it is assumed that the models are representative in terms of their predictive mechanisms and the data is representative in terms of its spatio-temporal coverage. Second, the parameter estimation of data learning models, such as neural network models, may involve elements of randomness; this work assumes that the influence of this randomness on the final prediction is minor. That said, it is noted that most data-driven models in this work, such as k-nearest neighbors (kNN) and extreme gradient boosting (XGBoost), are expected to converge to the same parameter set and produce identical predictions on the same test dataset. Third, the data measurement uncertainty is assumed to be negligible, in that, the ground-based measurements are used as the truth, against which model predictions are verified. Fourth, it is noted that the nRMSE is sufficient for model evaluation, as it can be decomposed into a series of error statistics, which jointly gauge the quality of predictions—this is what the Murphy–Winkler verification framework suggests. On this point, other metrics such as the mean absolute error (MAE) are redundant, and using both RMSE and MAE violates the statistical theory of consistency [23,24].

3.2. Benchmark separation models

Here, we select 10 state-of-the-art separation models of different types—see Yang [1] for their formulation. These 10 models are to be used as the benchmark models to evaluate the performance of the data-driven models of concern. Input parameters for all considered separation models are summarized in Table 1. As can be seen, different models take different input parameters, whereas k_t is employed as an essential input by all models. In the following paragraph, a brief justification for the choice of these 10 models is offered.

Starting from Engerer2 [14], this model is selected because it has won the separation modeling "contest" conducted by Gueymard and Ruiz-Arias [13] in 2016, and has since been used as a benchmark for almost all subsequent proposals of separation models. Whereas the model parameters of Engerer2 were fitted using Australia data, Bright and Engerer [25] refitted the parameters using a more comprehensive dataset, resulting in Engerer4, which is a "world" version of Engerer2. Starke1 and Starke2 [18] typify the piecewise modeling technique in separation modeling, in that, the inventors introduced a piecewise model that distinguishes between conditions with and without cloud enhancement. Since both Starke1 and Starke2 are based wholly on the BRL model [26], the BRL model is also used as a benchmark. Moving on to Abreu, it is a univariate model proposed very recently. Although the performance of Abreu lags most of the other benchmarks due to its univariate nature, it is considered here because it has possibly the best performance amongst the univariate separation models. Whereas Paulescu [27] represents a regression-based modeling philosophy, Every1 and Every2 [17] represent the climate-based modeling philosophy. In the case of the form, the separation model is viewed as a regression problem, and the authors adopted a linear model with indicator functions for that purpose. In the case of the latter, Every2 employs the BRL model as the basic function form, but fits one set of coefficients for each major climate type of concern. Last but not least, Yang4 [15] has won the "contest" conducted by Yang [1] and is the best separation model to date, as such it has to be included.

Table 1
Various input parameters as required by different separation models.

Parameter	Calculation method	Interpretation	Engerer2 Engerer4	Starke1 Starke2 BRL	Abreu	Paulescu	Every1 Every2	Yang4	Data-driven
E_0	Computed via solar positioning	Extraterrestrial GHI [W/m^2]							△
G_h	Obtained by local measurements	GHI [W/m^2]							△
G_{csky}	McClear clear-sky model	Clear-sky GHI [W/m^2]		△					△
Z	Computed via solar positioning	Solar zenith angle [°]	△					△	△
α	$90° - Z$	Solar elevation angle [°]		△			△		
AST	Computed via solar positioning	Apparent solar time	△	△			△	△	
k_t	G_h/E_0	Clearness index	△	△	△	△	△	△	△
$k_{t,\text{daily}}$	Averaging k_t over a day	Low-frequency k_t signal, a form of variability index		△		△	△		
ψ	Three-point moving average of k_t	Low-frequency k_t signal, a form of variability index		△			△		
Δk_{tc}	$k_{tc} - k_t$	Difference between clearness index of the clear-sky GHI ($k_{tc} = G_{\text{csky}}/E_0$) and clearness index	△					△	△
k_e	$\max\left(0, 1 - G_{\text{csky}}/G_h\right)$	Portion of the diffuse fraction that is attributable to cloud enhancement events	△					△	△
k_{csi}	G_h/G_{csky}	Starke's quantifier for cloud enhancement							△
$k_{\text{hourly}}^{\text{Engerer2}}$	Applying Engerer2 on hourly G_h	Hourly diffuse fraction estimated from Engerer2						△	
cc_{clim}	Acquired from database	Cloud frequency climatology							△
aod_{clim}	Acquired from database	Aerosol optical depth climatology							△
abd_{clim}	Acquired from database	Albedo climatology							△

3.3. End-to-end data-driven models

In this work, data-driven separation models are proposed and developed, to predict the diffuse fraction k using a range of popular machine-learning techniques. These techniques enable capturing the complex, nonlinear relationships between input features and the target variable, thereby potentially enhancing the accuracy of separation. Here, a total of 10 representative methods are used to build end-to-end data-driven models that can learn features from existing data and then adapt to different locations with or without new training data. By leveraging the strengths and mitigating the weaknesses of each method, the present setup aims to identify a comprehensive and robust data-driven separation model. The choice of these methods is motivated by their ability to handle different aspects of the problem, such as dealing with noisy data, capturing non-linear relationships, and providing interpretability. Details about the selected 10 methods are presented in Appendix.

Based on the thorough review of state-of-the-art semi-physical separation models in the literature, a total of 11 essential input features are selected for developing the data-driven models. The selected input features are summarized in Table 1. Among the 11 input features, eight are commonly used in semi-physical separation models, while the remaining three features are climatology variables, namely cloud frequency climatology (cc_{clim}), aerosol optical depth climatology (aod_{clim}) and albedo climatology (abd_{clim}). The climatology variables are obtained based on their importance to surface radiation. Cloud is the most important factor affecting the transmission of irradiance down to the earth's surface, therefore, cloud climatology data employed in this work comes from Wilson and Jetz [28]. Right next to clouds, the next-important atmospheric constituent that influences the surface radiation is aerosol, as such the aerosol optical depth (AOD) product merged from several major AOD databases by Yang and Gueymard [29] is herein selected. Last but not least, surface albedo, which controls the backscattering process, is selected as the third climatology variable, and its source is the ERA5 reanalysis Hersbach et al. [30]. All three climatology databases cover the entire world, and the pixel values collocated with the ground-based stations used in this work are elicited from the corresponding climatology maps. Most input features for the data-driven separation model can be derived or obtained from NWP products. Only the surface albedo, which has low inter-day and inter-season variability, is obtained from climatology data. Therefore, all data can be easily obtained or estimated, which could ensure the applicability of the data-driven models. Note that each feature undergoes standardization via the Z-score normalization method before being inputted into the data-driven models.

3.4. Evaluation of model performance

The performance evaluation of the data-driven and benchmark models involves a two-stage process utilizing data from 12 sites. The first stage is model development and validation using data from seven North American sites (BON, BOU, DES, FOP, GOO, ROS, SIO). Here, 70% of the data is allocated for training and 30% for testing. The second stage is for model transferability assessment, using data from five European sites (CAB, CAR, CEN, PAL, PAY), which are solely for model testing at unseen sites. This strategy ensures the models are robust against meteorological variations across different regions. For evaluation at both stages, the estimated diffuse fraction is then converted to corresponding diffuse and beam normal irradiance values for calculating the error metrics between the model estimated and measured irradiance. The primary performance metric is nRMSE, complemented by prediction skill metrics for a comprehensive comparison between data-driven and semi-physical models. Additionally, the Murphy–Winkler factorizations

are employed to provide an extensive examination of the prediction and observation distributions, offering a holistic view of prediction quality.

The nRMSE is used as the main metric to assess and compare the performance of D_h and B_n estimated by all models. The nRMSE is calculated as:

$$\mathrm{nRMSE} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{I}_i - I_i\right)^2}}{\frac{1}{n}\sum_{i=1}^{n} I_i}, \tag{10}$$

where n is the number of observation instances, $\hat{I}$ and I represents the estimated and observed values (either D_h or B_n), respectively.

The use of normalized metrics provides several advantages in this context. First, normalized metrics facilitate a fair comparison between models by accounting for variations in the magnitude of the target variable across different datasets or locations. By scaling the errors with respect to a baseline, such as the mean value of the target variable, the error metrics are dimensionless and become comparable across different scenarios. In addition, normalized metrics allow for a more intuitive interpretation of the results, as they provide a relative measure of the performance of the model. By expressing the errors as a percentage or fraction of a baseline value, the degree of deviation from the actual target values can be easily gauged.

Besides calculating the errors for individual models, it is also of interest to know the overall performance comparison between data-driven and semi-physical models. On this point, a method that can compare the two classes of models in an aggregated fashion is needed. The concept of prediction skill is considered for that matter. This concept encompasses three aspects: mean, standard deviation, and the best nRMSE. The prediction skills of those three kinds are defined mathematically as follows:

$$s_{\mathrm{mean}} = 1 - \frac{\sum_i \mathrm{nRMSE}_{\mathrm{d}}^i}{\sum_i \mathrm{nRMSE}_{\mathrm{b}}^i}, \tag{11}$$

$$s_{\mathrm{deviation}} = 1 - \sqrt{\frac{\sum_i \left(\mathrm{nRMSE}_{\mathrm{d}}^i - \sum_i \mathrm{nRMSE}_{\mathrm{d}}^i/N\right)^2}{\sum_i \left(\mathrm{nRMSE}_{\mathrm{b}}^i - \sum_i \mathrm{nRMSE}_{\mathrm{b}}^i/N\right)^2}}, \tag{12}$$

$$s_{\mathrm{best}} = 1 - \frac{\min\left(\mathrm{nRMSE}_{\mathrm{d}}^i\right)}{\min\left(\mathrm{nRMSE}_{\mathrm{b}}^i\right)}, \tag{13}$$

where N is 10, representing the number of models, $\mathrm{nRMSE}_{\mathrm{d}}^i$ and $\mathrm{nRMSE}_{\mathrm{b}}^i$ represent the nRMSE of the ith data-driven model and the ith benchmark model, respectively.

The last component of the verification exercise pertains to the Murphy–Winkler factorizations [31]. In evaluating point predictions, one often focuses just on the accuracy, which is but one aspect of prediction quality. Other aspects of quality include association, calibration, refinement, and discrimination, among others. In this regard, Murphy and Winkler [31] proposes to examine the joint distribution of prediction and observation, which contains all information relevant to prediction. Furthermore, since the joint distribution can be decomposed into marginal and conditional distributions, those are also scrutinized. The Murphy–Winkler verification framework was first introduced to solar engineering by Yang and Perez [32], and was subsequently advocated by a group of 33 energy meteorologists [24]. Ever since, the framework has gained much popularity, and has been applied to a wide range of energy meteorology problems [e.g., 29,33–35]. In this work, the formal introduction of the Murphy–Winkler factorizations is not reiterated, and the reader is referred to the aforementioned works for details. However, we should give the necessary information as the verification exercise progresses.

4. Results and discussion

4.1. Model assessment at locations with training data

The evaluation of the model performance at the seven locations used for training is presented in Table 2—the evaluation is performed on the remaining 30% data at these locations. The results indicate that the mean and best performance of data-driven models for estimating both D_h and B_n is markedly superior to those of the benchmark models. In particular, the relative improvement in the average nRMSE of data-driven models compared to benchmark models ranges from 14.9% to 20.6% for D_h, and from 17.3% to 22.6% for B_n across the seven sites. When comparing the best-performing models from the two groups (Yang4 and XGBoost), the relative improvement of XGBoost in terms of nRMSE ranges from −1.4% to 8.9% for D_h and from 2.8% to 16.1% for B_n across the seven sites. The majority of data-driven models consistently outperform the average performance of benchmark models. This superior performance can be attributed to the adaptability and capability of data-driven models to learn complex, non-linear relationships. In contrast, benchmark models often rely on pre-defined assumptions or relationships, which may constrain their capacity to capture the intricacies of the data, ultimately resulting in higher variation among models. The standard deviation among data-driven models is notably lower than that of the benchmark models. This observation suggests that data-driven models are effective in learning the inherent data patterns, leading to similar performance levels. More specifically, these data-driven models may identify congruent relationships between input variables and the target variable, indicating that they exploit distinct aspects of the data while converging on a comparable underlying structure. This consistency in data patterns across different models allows each model to achieve a similar level of accuracy. In conclusion, when training data is available, data-driven models exhibit significant advantages over the semi-physical benchmarks in estimating both D_h and B_n in the context of irradiance separation.

4.2. Model assessment at unseen locations

The performance of the models at unseen locations is presented in Table 3. For these locations, the advantage of data-driven models over benchmark models persists, albeit with a noticeable reduction. Specifically, data-driven models generally outperform the majority of benchmark models in terms of average nRMSE for estimating both D_h and B_n. The relative improvement of average nRMSE of data-driven models over benchmark models ranges from 7.9% to 17.6% for D_h and from 12.4% to 17.5% for B_n across the five unseen sites. In a similar vein to locations with training data, the standard deviation of data-driven models is significantly lower than that of the benchmark models. However, for the best-performing models, the relative improvement of nRMSE of data-driven models over benchmark models ranges from −12.2% to 0.3% for D_h and −8.8% to −0.1% for B_n among the five sites. These observations suggest that data-driven models may exhibit a marginal degree of overfitting, leading to increased error when estimating for sites without training data. This outcome can be attributed to the inherent complexity of the models and their propensity to capture noise in the training data, which may not generalize well to previously unseen locations.

4.3. Murphy–Winkler verification of the top performers

In this study, a wide range of models have been evaluated. The following part concentrates on a detailed comparison between the two top performers, one of the data-driven models and the other of benchmark models, namely, LightGBM and Yang4. The comparison is performed at the five unseen locations, in terms of the diffuse fraction k. It is important to note that the differences among the top data-driven models, such as ANN, kNN, XGBoost, or LightGBM, are relatively minor. Further analysis confirms the similarity in performance among these top-performing data-driven models.

By selecting this pair of models, we showcase the best overall performance for both data-driven and benchmark categories. To enable a comprehensive examination of their predictive capabilities, a series of illustrative figures are presented following the recommendation

Table 2
Comparisons of nRMSE (%) for estimating D_h (left) and B_n (right) at locations with training data. $\mathbb{E}$ denotes the expectation, which applies to both D_h or B_n and has a unit of W/m^2. The mean, standard deviation, and best (i.e, smallest) nRMSE values among the 10 benchmark and the 10 data-driven models are also presented. In the last three rows, the three skill scores (%) introduced in Section 3.4 are displayed.

	nRMSE of D_h predictions							nRMSE of B_n predictions						
	BON	BOU	DES	FOP	GOO	ROS	SIO	BON	BOU	DES	FOP	GOO	ROS	SIO
$\mathbb{E}$ (W/m^2)	161	138	112	133	161	169	146	561	636	739	568	558	519	560
Engerer2	42.3	39.0	34.6	37.1	40.1	43.3	38.3	16.4	16.4	15.0	17.1	14.6	16.4	16.2
Engerer4	47.2	43.2	38.7	40.2	46.1	49.1	42.3	18.4	18.2	15.5	18.1	16.8	18.6	17.5
Starke1	39.7	37.8	30.9	35.6	36.8	40.0	36.4	15.9	15.9	12.9	16.8	13.4	15.3	15.9
Starke2	46.7	43.6	36.1	40.2	39.8	46.5	42.2	18.9	18.7	15.4	19.3	14.6	17.9	18.8
BRL	52.2	53.5	41.8	46.2	49.4	55.7	47.9	20.8	23.2	17.5	21.5	18.0	21.2	20.5
Abreu	47.9	48.1	41.1	42.4	45.9	49.7	44.2	19.6	20.1	17.0	20.0	18.8	19.7	19.2
Paulescu	41.3	40.8	33.9	38.0	37.4	41.8	38.6	17.2	17.9	14.5	18.2	15.0	16.8	17.2
Every1	60.5	60.8	48.0	51.2	57.3	64.3	53.5	23.4	25.7	19.8	23.5	20.5	23.9	22.6
Every2	60.5	53.5	41.8	51.7	57.2	62.8	47.9	23.3	23.2	17.5	23.9	20.3	23.9	20.5
Yang4	37.3	35.6	29.2	33.3	34.3	38.0	33.5	14.9	15.4	13.3	16.2	12.8	14.7	14.8
Mean	47.6	45.6	37.6	41.6	44.4	49.1	42.5	18.9	19.5	15.8	19.5	16.5	18.8	18.3
Deviation	8.1	8.1	5.7	6.3	8.2	9.2	6.1	2.9	3.5	2.1	2.7	2.8	3.3	2.4
Best	37.3	35.6	29.2	33.3	34.3	38.0	33.5	14.9	15.4	12.9	16.2	12.8	14.7	14.8
MLR	43.8	41.3	35.4	40.5	42.5	44.7	42.2	17.2	17.4	14.7	18.3	15.5	17.2	17.8
QPR	38.7	36.3	30.9	34.3	37.2	40.0	36.0	15.1	15.1	12.6	15.4	13.4	15.3	14.9
DT	38.9	36.2	30.4	33.8	37.7	40.4	35.8	15.3	15.0	12.4	15.3	13.7	15.5	14.9
RF	38.6	35.9	30.2	33.6	37.3	40.1	35.5	15.1	14.9	12.3	15.2	13.5	15.3	14.8
GB	40.1	36.5	30.7	34.8	38.6	41.5	36.4	15.5	15.1	12.5	15.6	14.1	15.8	15.1
Light GBM	38.1	35.4	30.0	33.4	36.6	39.3	35.3	14.9	14.7	12.2	15.0	13.2	15.1	14.6
XGBoost	35.6	32.8	27.7	30.4	34.8	36.6	32.5	13.9	13.7	11.4	13.6	12.5	13.9	13.4
SVR	39.9	37.7	33.2	37.1	39.2	40.6	38.5	15.7	15.8	13.7	16.6	14.4	15.6	16.1
kNN	38.2	35.2	29.7	33.0	36.9	39.4	35.2	14.9	14.7	12.1	14.9	13.3	15.1	14.5
ANN	37.0	34.5	29.1	31.9	35.8	38.2	34.0	14.5	14.3	11.8	14.3	12.9	14.6	14.0
Mean	38.9	36.2	30.7	34.3	37.7	40.1	36.1	15.2	15.1	12.6	15.4	13.6	15.3	15.0
Deviation	2.2	2.2	2.2	2.8	2.1	2.1	2.6	0.9	1.0	1.0	1.3	0.8	0.8	1.2
Best	35.6	32.8	27.7	30.4	34.8	36.6	32.5	13.9	13.7	11.4	13.6	12.5	13.9	13.4
s_{mean}	18.2	20.6	18.3	17.6	15.2	18.4	14.9	19.4	22.6	20.6	20.8	17.3	18.6	18.3
$s_{\text{deviation}}$	73.1	72.9	62.5	55.6	74.0	76.8	56.8	70.0	71.8	54.7	52.8	69.8	74.5	50.9
s_{best}	4.5	8.0	5.1	8.9	−1.4	3.9	3.1	7.2	11.1	11.8	16.1	2.8	5.4	9.6

of Murphy and Winkler [31] and Yang et al. [24], including marginal, joint, and conditional distribution plots of observed and estimated diffuse fractions. The subsequent discussion offers valuable insights into the strengths and weaknesses of these elite models, illuminating their relative advantages and the factors contributing to their superior performance.

4.3.1. Verifying the marginal distributions

Fig. 3 presents the marginal distributions of predicted and observed diffuse fractions for both top performers. The distribution patterns of the observed and predicted diffuse fractions for both models across training and testing sets exhibit a similar bimodal distribution. This distribution arises from two-state weather conditions, with clear-sky periods (minimal cloud cover) corresponding to lower diffuse fractions and cloudy periods (significant cloud cover) corresponding to higher diffuse fractions, resulting in one peak near 0.2 and another close to 1, respectively.

Upon closer examination of Fig. 3, the two peaks of the distributions for both LightGBM and Yang4 predictions have shifted from those of the observations. The peak near 0.2 shifts rightward, and the peak near 1 shifts leftward, a tendency especially pronounced during cloudy periods where the diffuse fraction is close to 1. This phenomenon can be attributed to the conservative prediction tendency of data-driven models that aim to minimize statistical metrics like the mean square error (MSE). By avoiding extreme predictions, these models reduce the likelihood of significant errors, albeit sometimes at the expense of prediction accuracy. When comparing the two models, the distribution of Yang4's predicted k is better aligned with the observations than the LightGBM model, particularly during the cloudy period when k is close to 1.

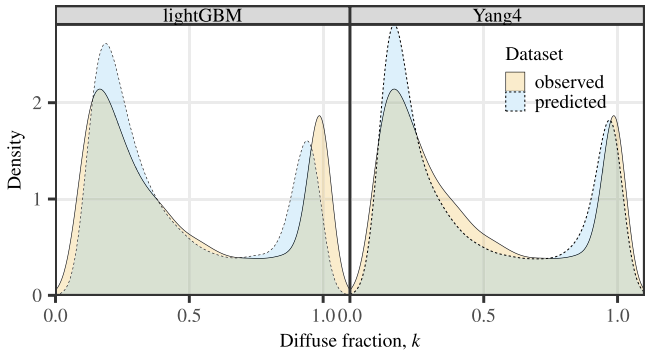


Fig. 3. Marginal distributions of predicted and observed k using LightGBM and Yang4 when evaluated at the five locations without training data.

4.3.2. Verifying the joint distributions

To examine the distribution of predictions from both models in greater detail, the joint distributions of the observed and predicted diffuse fraction, for both LightGBM and Yang4, are plotted in Fig. 4. On the top and right margin of the sub-figures, the marginal distributions are again shown in the form of histograms for information. Both models tend to slightly over-predict k when k is smaller than 0.5 and under-predict k when k is larger than 0.5. The joint distribution of the Yang4 case demonstrates better alignment along the diagonal than that of the LightGBM model, particularly for large k conditions, which in turn explains its smaller statistical errors. However, a common challenge for all models, regardless of the availability of training data, is the prediction under partially cloudy conditions when measured diffuse fractions are

Table 3
Comparisons of nRMSE (%) for estimating D_h (left) and B_n (right) at locations without training data. $\mathbb{E}$ denotes the expectation, which applies to both D_h or B_n and has a unit of W/m^2. The mean, standard deviation, and best (i.e., smallest) nRMSE values among the 10 benchmark and the 10 data-driven models are also presented. In the last three rows, the three skill scores (%) introduced in Section 3.4 are displayed.

	nRMSE of D_h predictions					nRMSE of B_n predictions				
	CAB	CAR	CEN	PAL	PAY	CAB	CAR	CEN	PAL	PAY
$\mathbb{E}$ (W/m^2)	156	125	147	177	139	308	610	553	404	631
Engerer2	34.4	33.9	37.3	39.7	36.5	18.7	19.0	19.7	21.2	19.7
Engerer4	39.5	38.9	42.6	44.4	40.9	22.2	21.7	22.6	24.2	22.3
Starke1	28.6	30.6	34.6	34.1	37.6	16.4	17.0	18.3	18.8	20.2
Starke2	33.8	31.8	39.1	40.1	45.1	19.0	17.5	20.5	22.2	24.7
BRL	43.3	39.9	48.3	50.3	48.4	24.2	22.7	25.5	27.2	26.2
Abreu	38.0	40.5	44.5	44.2	42.1	22.1	25.6	25.1	24.8	24.1
Paulescu	30.2	33.4	37.3	35.0	40.0	18.2	20.4	20.7	20.1	21.9
Every1	49.1	41.8	53.1	57.2	55.4	26.9	24.4	27.9	30.3	29.4
Every2	49.4	42.2	52.2	56.9	53.5	27.6	25.5	28.0	30.7	29.2
Yang4	28.9	28.0	31.8	33.8	34.7	16.2	16.5	17.3	18.5	18.9
Mean	37.5	36.1	42.1	43.6	43.4	21.1	21.0	22.6	23.8	23.7
Deviation	7.8	5.1	7.3	8.8	7.1	4.1	3.5	3.9	4.5	3.8
Best	28.9	28.0	31.8	33.8	34.7	16.2	16.5	17.3	18.5	18.9
MLR	36.1	36.7	41.3	41.3	38.3	20.4	19.5	21.3	22.5	20.6
QPR	34.3	32.5	37.2	38.8	36.6	18.9	17.5	19.0	20.9	19.8
DT	32.4	32.7	35.4	37.9	35.1	18.3	18.0	18.8	20.7	19.2
RF	32.2	32.3	35.1	37.6	34.7	18.1	17.8	18.6	20.6	19.0
GB	33.0	33.1	35.8	37.9	34.6	18.4	18.6	19.2	20.5	19.0
Light GBM	31.6	32.4	34.6	37.6	34.6	17.7	17.4	18.3	20.6	19.0
XGBoost	32.9	32.1	36.2	38.6	36.9	18.6	17.3	19.0	21.0	20.1
SVR	32.4	35.7	38.3	37.9	35.6	18.4	18.6	19.7	21.2	19.8
kNN	32.4	31.5	35.5	37.9	36.2	18.3	17.0	18.7	20.6	19.8
ANN	32.2	33.5	36.0	37.5	34.9	17.6	18.3	18.9	20.0	18.9
Mean	33.0	33.3	36.5	38.3	35.8	18.5	18.0	19.1	20.9	19.5
Deviation	1.3	1.7	2.0	1.1	1.2	0.8	0.7	0.8	0.6	0.6
Best	31.6	31.5	34.6	37.5	34.6	17.6	17.0	18.3	20.0	18.9
s_{mean}	12.2	7.9	13.1	12.1	17.6	12.6	14.4	15.2	12.4	17.5
$s_{\text{deviation}}$	83.0	67.3	73.0	87.2	82.5	81.5	78.5	78.3	85.7	84.4
s_{best}	−10.4	−12.2	−8.7	−11.1	0.3	−8.8	−3.1	−5.8	−8.3	−0.1

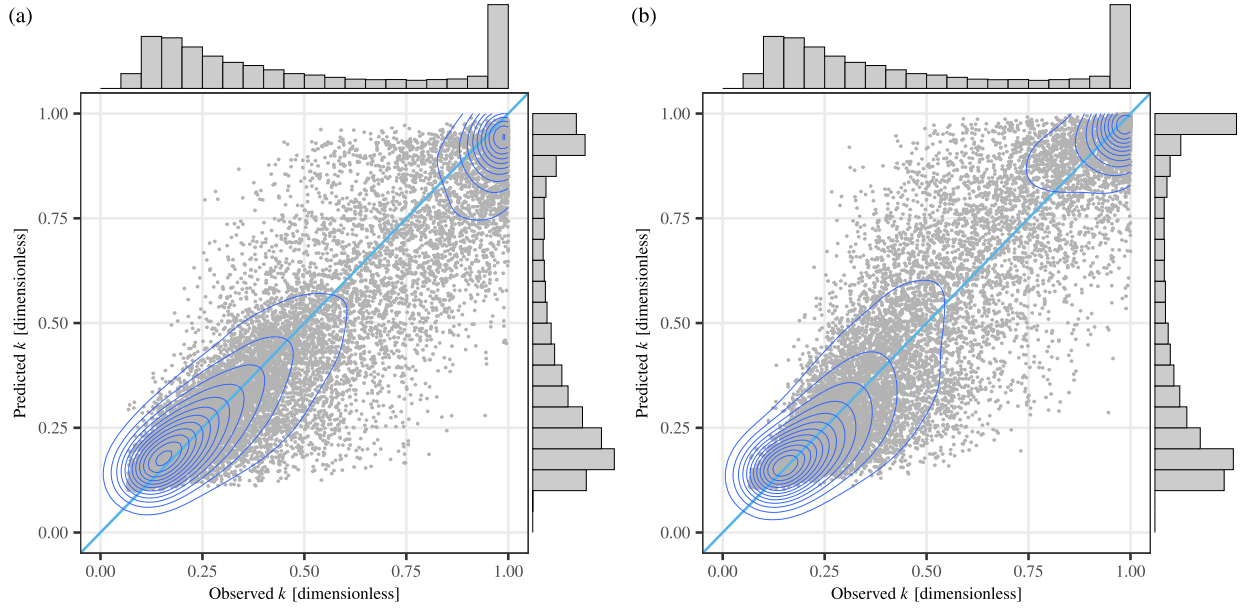


Fig. 4. Joint and marginal distributions of observed and predicted k using (a) LightGBM and (b) Yang4 when evaluated at the five locations without training data. The contour lines show the 2D kernel densities.

around 0.5. Here, the models exhibit a highly dispersed distribution away from the diagonal, reflecting the high variability of the cloud field during such conditions. This demonstrates that accurately separating solar irradiance under partially cloudy conditions remains a significant challenge, irrespective of the model used or the availability of training data.

4.3.3. Verifying the conditional distributions

In addition to the joint and marginal distributions, Fig. 5 presents the conditional distributions to investigate the conditional dependence between prediction and observation. Fig. 5(a) and (c) correspond to the observation conditional on prediction, and Fig. 5(b) and (d) correspond to the prediction conditional on observation. It can be seen from Fig. 5(b) and (d) that both models tend to overestimate at the lower

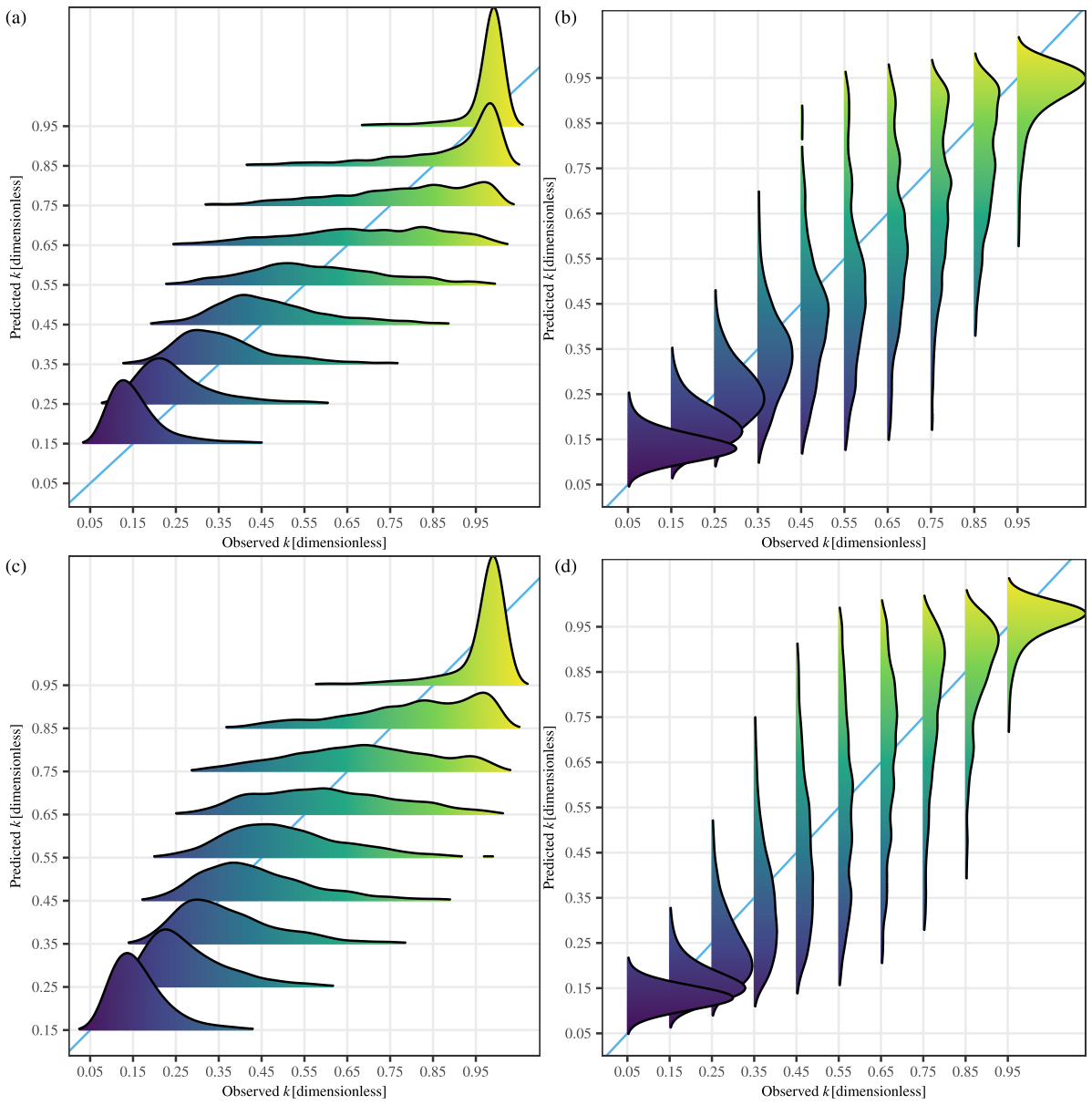


Fig. 5. Conditional distributions of observed and predicted k when evaluated at the five location without training data. $f(y|x)$ are shown in (a) and (c) for LightGBM and Yang4, respectively. $f(x|y)$ are shown in (b) and (d) for the two models, respectively.

k range and underestimate at the higher k range. Fig. 5(a) and (c), on the other hand, reveal that local distribution maxima of LightGBM estimations align better with the diagonal line than Yang4 when $k < 0.15$ or $k > 0.95$. However, Yang4 demonstrates better alignment when $0.15 < k < 0.35$ or $0.75 < k < 0.95$. A common observation across both models is the relatively flat probability density around k values of 0.5, indicating larger prediction errors. This pattern again highlights the challenges of estimating diffuse fractions under partially cloudy conditions, a difficulty that persists regardless of the availability of training data.

This analysis that compares the top-performing data-driven model and the benchmark model offers valuable insights into their respective strengths and weaknesses. The detailed exploration of their performance under different conditions can inform future research and model development in the field of solar irradiance separation. By understanding the specific areas where these models excel or fall short, we can target improvements to enhance their effectiveness, with the ultimate goal of furthering the practical utility of these models in the renewable energy sector.

4.4. Discussion of results

The analysis of performance for data-driven models reveals an interesting insight related to D_h and B_n predictions: There is a strong correlation between the performance of D_h and B_n predictions, particularly for locations with training data. This suggests that a model proficient at predicting D_h for a specific site is also likely to excel at predicting B_n at the same site. This correlation can be attributed to the closure relationship between D_h, B_n, and G_h, i.e., $G_h = D_h + B_n \cos Z$. As the predictand k is the ratio of D_h to G_h, and the inter-dependence of the three irradiance components implies that the performance of a separation model in predicting one variable directly affects its performance in predicting the other.

Next, among all data-driven models, MLR generally exhibits high nRMSE values, likely due to its assumption of linearity. The relationship between the meteorological inputs and the output diffuse fraction is inherently nonlinear, causing MLR to struggle in accurately modeling this association and resulting in increased prediction errors. Models capable of establishing nonlinear mathematical mappings, such as

LightGBM, kNN, or ANN, exhibit similar behavior in terms of nRMSE. It is important to note that the performance differences between the top models and their closest competitors are just marginal, emphasizing their comparable nature. This comparability among data-driven models indicates their proficiency in capturing underlying data structures and patterns, leading to consistent predictive accuracy.

Moreover, as observed in Fig. 3, both the observation and prediction sets exhibit similar distributions with bimodality. Upon examining the predictions made by the models, it is evident that all models tend to draw the two peaks closer together, with the peak near $k = 0.2$ shifting rightward and the peak near $k = 1$ shifting leftward. As discussed earlier, this observation can be attributed to the fact that machine-learning models often predict conservatively in order to minimize statistical metrics such as MSE. By avoiding extreme predictions, these models can reduce the risk of incurring large errors, even if this strategy leads to less accurate predictions in some instances. In a statistical sense, MSE-optimized predictions models are necessarily under-dispersed, as mathematically proven by Mayer and Yang [23]. Another reason for the described misalignment is likely due to the misidentification of the sky conditions. One intuitive approach to addressing the observed issue involves dividing the data into clear and cloudy subsets, e.g., based on the ranges of the clear-sky index, and then developing separate models tailored to each weather condition. This piecewise modeling strategy has been employed by Starke et al. [18], but not here.

To assess the above hypothesis, four scatter plots as shown in Fig. 6 are examined. In these plots, the diffuse fraction is on the y-axis and various input features, namely GHI (G_h), solar zenith angle (Z), clearness index (k_t), and clear-sky index (k_{csi}), are on the x-axis. Orange dots represent training data, and blue dots represent testing data. Both the training and testing data display similar distributions. For the GHI plot, when G_h ranges from 0 to 600 W/m^2, one can observe that the k values mainly concentrate at regions close to 0 or close to 1, echoing the bimodal distribution seen in the previous analysis (Fig. 3). However, from a data-driven modeling perspective, it is challenging to predict k solely based on G_h, as G_h alone does not provide sufficient information about the sky condition, which plays a significant role in determining k. The solar zenith angle plot shows a similar distribution, indicating that the bimodal distribution of k is less likely to be caused by diurnal effects. Relying solely on features, such as G_h or solar zenith angle, to differentiate between the two peaks in the distribution would also prove difficult, as it does not capture the nuances of sky-condition-related effects on k.

The clearness index and clear-sky index plots are more informative as indicators of weather conditions. However, the mathematical relationship between these two factors and k is not as straightforward or simple as one might anticipate. Intuitively, lower k_t and k_{csi} values indicate more cloudy conditions, and vice versa. As expected, k is higher (close to 1) under more cloudy conditions, as the sun is obscured by clouds, and lower under clear sky conditions. Therefore, data-driven models may learn patterns from these two useful inputs to achieve higher accuracies. However, as observed in these subplots, there are still numerous instances that could be considered outliers. For example, for k_t ranging from 0.4 to 0.6, k ranges from 0.2 to 1, and for k_{csi} values close to 1, k ranges from 0.2 to 0.8. These inconsistencies in the data can adversely affect the modeling process, causing data-driven models to make conservative predictions, as observed in the previous analysis.

4.5. Remarks and recommendation for future research

Based on the aforementioned results and discussions, one can conclude that the choice between data-driven models and the available benchmark models largely depends on the availability of training data and the specific requirements of the application. Below are some guidelines for selecting an appropriate separation model:

- Availability of training data: If training data is available for the location of interest, data-driven models, such as lightGBM, XGBoost, kNN, or ANN, are recommended as they consistently outperform the semi-physical benchmark models in terms of prediction accuracy and consistency. Potential scenarios include but are not limited to maintenance of PV plants during which temporary and movable sensors measuring all irradiance components are placed on site for a period of time, or only historical D_h or B_n are measured during the initial resource assessment period.
- Absence of training data: When training data is unavailable for a specific location, the choice between data-driven and semi-physical benchmarks becomes apparent. While data-driven models may still outperform benchmark models in some cases, such better performance is usually not guaranteed. In these situations, one should consider additional factors, such as domain knowledge, model interpretability, robustness, transfer learning, or ensemble modeling, as mentioned earlier.
- Model complexity and computational resources: Data-driven models tend to be more computationally intensive as compared to semi-physical benchmarks. Depending on the available computational resources and the required response time for predictions, it would be preferred to opt for a simpler semi-physical model or a more complex data-driven model.
- Model interpretability and decision-making: If gaining insights into the underlying processes and relationships is crucial for decision-making, a more interpretable semi-physical model or kNN model might be preferred. However, if prediction accuracy is of primary concern, a data-driven model may be more suitable.
- Flexibility and adaptability: Data-driven models can adapt to new data when coupled with automated machine-learning framework [36,37], allowing them to learn and improve over time. If the goal is to have a model that can adjust to changing conditions, a data-driven model may be a better choice.

The above analysis also offers valuable insights into the relationships between key input features and diffuse fraction distribution, underscoring the challenges in predicting k using available input features. To improve model performance and pave the way for future research, several key takeaways are drawn and corresponding recommendations are made:

- Incorporating domain knowledge: Incorporating more domain knowledge and understanding the underlying physical mechanisms that cause the varying or bimodal distributed k under the same levels of G_h or k_t is essential. Additionally, guided by domain knowledge, we could also incorporate more relevant original or extracted features [38]. This knowledge can guide feature engineering and model development, helping to create more accurate and robust predictions.
- Additional weather-related variables and feature engineering: Exploring the inclusion of other weather-related variables, such as temperature or humidity, might provide a more comprehensive representation of the underlying data patterns. This can aid in capturing the complex, nonlinear relationships between meteorological inputs and diffuse fraction, ultimately improving model performance. Moreover, combining input features or creating new ones based on domain knowledge and data-driven insights may reveal more informative representations of the data. This could enhance the predictive power of data-driven models by allowing them to identify and exploit previously unrecognized patterns.
- Model selection and ensemble techniques: This work builds and investigates several data-driven models as individuals. However, investigating alternative modeling techniques or using ensemble methods that combine the strengths of different models could lead to improved prediction accuracy. These approaches may help mitigate overfitting and thus enhance the generalizability of data-driven models to unseen locations or conditions.

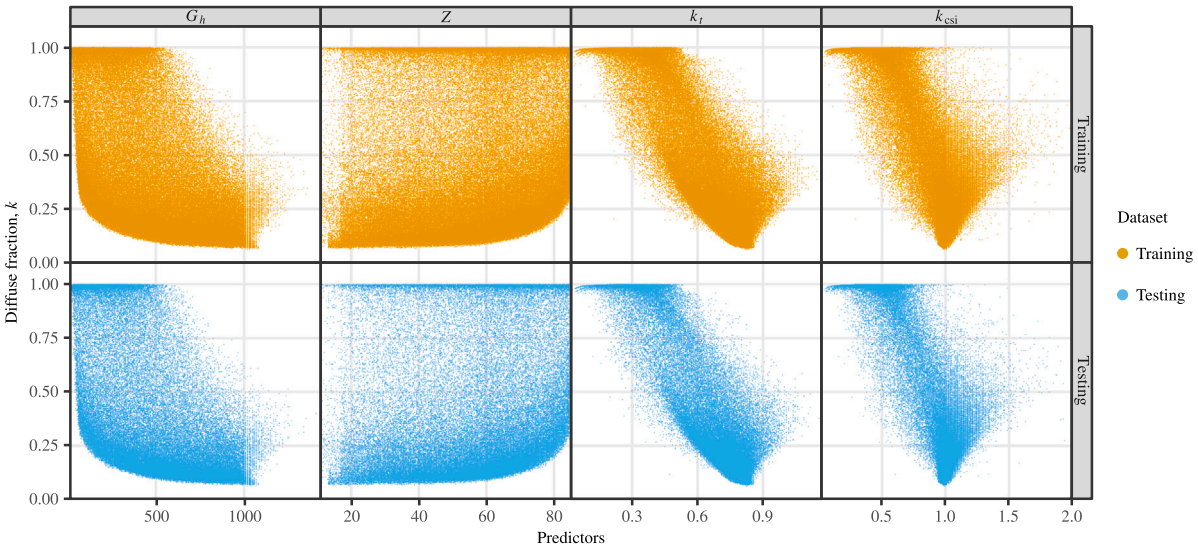


Fig. 6. Scatter plot of diffuse fraction (y-axis) versus selected key input features (x-axis). Selected key input features are GHI (G_h), solar zenith angle (Z), clearness index (k_t), and clear-sky index (k_{csi}). Blue and red colors represent data from the training and testing sets, respectively.

5. Conclusion

This study has explored the performance of 10 representative data-driven models in predicting diffuse fraction, and subsequently, diffuse horizontal irradiance and beam normal irradiance. A total of 10 state-of-the-art semi-physical models from the most recent literature are used as benchmarks. Employing the most informative error metric for solar applications, that is the normalized root mean square error, a comprehensive evaluation of model performance across 12 sites spanning two continents is conducted. Whereas seven stations in one continent are used for training and on-site evaluation, the remaining five in another continent are set aside for off-site (i.e., unseen) evaluation.

Data-driven models, when compared to semi-physical models, have demonstrated an error reduction of 15.2% to 22.6% on datasets from training locations and 7.9% to 17.6% on datasets from unseen locations. However, the degree of performance enhancement of data-driven models tends to decrease for sites without training data, particularly for the top-performing models. Moreover, data-driven models exhibit similar predictive behaviors, suggesting they might be learning analogous data patterns, and consequently achieve significantly lower standard deviations. Challenges in accurately predicting the diffuse fraction are highlighted in scatter plots of predictions, which illustrate the association between key input features and the diffuse fraction distribution. To enhance the performance of data-driven models, future work could consider incorporating additional weather-related variables or features and integrating relevant domain knowledge into the modeling process. In summary, this study offers valuable insights into the performance of data-driven models for diffuse fraction prediction, the strategies for selecting these models, and the challenges encountered during their development. These findings pave the way for future research in this field.

CRediT authorship contribution statement

Yinghao Chu: Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Dazhi Yang:** Writing – review & editing, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Hanxin Yu:** Writing – review & editing, Validation, Methodology. **Xin Zhao:** Writing – review & editing, Resources, Conceptualization. **Mengying Li:** Writing – review & editing, Visualization, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

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.

Data availability

Data will be made available on request.

Acknowledgments

Yinghao Chu is substantially supported by a grant from City University of Hong Kong (Project No. 9610625). Dazhi Yang is supported by the National Natural Science Foundation of China (Project No. 42375192), and China Meteorological Administration Climate Change Special Program (CMA-CCSP; Project No. QBZ202315). Mengying Li is substantially supported by a grant from the Research Grants Council of Hong Kong (Project No. 25213022). Both Yinghao Chu and Mengying Li are substantially supported by a collaborative grant from the Research Grants Council of Hong Kong (Project No. C6003-22Y).

Appendix. Selected data-driven methods

The data-driven algorithms used in this work are summarized here.

A.1. Multivariate regression

Multivariate regression (MLR) method is one of the most fundamental and straightforward statistical-learning techniques that captures the relationship between input variables $\boldsymbol{x}$ and output variable y using linear predictor functions. The unknown model parameters are estimated using the least squares method, minimizing the squared difference between the observed and predicted values of the output variable [39]. The mathematical expression of MLR is [40]:

$$y_i = \boldsymbol{\omega} \cdot \boldsymbol{x}_i + \omega_0, \tag{A.1}$$

where $\boldsymbol{\omega}$ represents the weight vector, $\boldsymbol{x}_i$ denotes the input vector, and ω_0 is an intercept term. Model coefficients $\boldsymbol{\omega}$ and ω_0 are to be determined by minimizing a loss function. While widely used and easy to interpret, MLR has some drawbacks when applied to complex problems, such as linearity assumption, sensitivity to outliers, multicollinearity, or limited predictive performance due to its simplicity. It is also noted that the error term is omitted for brevity in this and subsequent equations.

A.2. Quadratic polynomial regression

Quadratic polynomial regression (QPR) is an extension of linear regression, where the relationship between the independent variables and the dependent variable is assumed to be an order-2 polynomial. This method can capture more complex relationships between the input features and the target variable compared to linear regression by incorporating the quadratic terms of the independent variables. For a dataset with m input features, the quadratic polynomial regression model can be represented as:

$$y_i = \omega_0 + \sum_{j=1}^{m} \omega_j x_{ij} + \sum_{j=1}^{m} \sum_{k=j}^{m} \omega_{jk} x_{ij} x_{ik}. \tag{A.2}$$

In the above equation, the first summation term represents the linear contributions of each independent variable, while the second double summation term accounts for the interaction between the independent variables as well as their quadratic contributions.

A.3. Decision tree

Decision tree (DT) is a nonparametric, supervised-learning method that recursively partitions the input feature space to create a tree-like structure for making predictions. It is capable of capturing nonlinear relationships between the input features and the target variable, and it is interpretable due to its tree-like structure, which represents a set of hierarchical decisions. A DT is constructed by iteratively splitting the dataset into subsets based on the values of input features. At each node, the tree selects the best feature to split the data by minimizing a certain impurity criterion, such as the mean square error. This process continues until a stopping criterion is met, such as a maximum tree depth, a minimum number of samples per leaf node, or an improvement in the impurity measure below a certain threshold. Given a dataset with m input features, the decision tree prediction model can be represented as a series of decisions based on the values of the input features:

$$y_i = f(x_{i1}, x_{i2}, \dots, x_{im}), \tag{A.3}$$

where f is a function representing the hierarchical decisions made by the decision tree.

A.4. Random forest

Random forest (RF) is an ensemble-learning method that combines the predictions of multiple decision trees to improve the overall performance and stability of the model. By leveraging the power of multiple trees, random forests can capture complex nonlinear relationships between input features and the target variable while mitigating the overfitting issues commonly associated with single decision trees. A RF model consists of B individual decision trees, each trained on a bootstrap sample of the original dataset. During the training process, each tree is built by selecting a random subset of input features at each split, which further increases the diversity among the trees in the ensemble. The final prediction of the random forest model is obtained by averaging the predictions of all the individual trees:

$$y_i = \frac{1}{B} \sum_{b=1}^{B} f_b(x_{i1}, x_{i2}, \dots, x_{im}), \tag{A.4}$$

where f_b is the prediction function of the bth decision tree. RF has several advantages over single decision trees, such as improved predictive performance, reduced overfitting, or increased model stability. Despite its improved performance and robustness compared to a single decision tree, random forests still have some limitations, in that, they can be computationally expensive and slower to train, especially with large datasets and a high number of trees in the ensemble.

A.5. Gradient boosting

Gradient boosting (GB) is an ensemble-learning technique that builds a strong model by iteratively fitting weak learners, typically decision trees, to the residuals of the previous learners. The main idea behind gradient boosting is to minimize the loss function by adding new learners in a sequential manner, with each new learner aiming to correct the errors made by the previous learners. In gradient boosting, the final model is a weighted sum of the weak learners, and the prediction is given by:

$$y_i = \sum_{b=1}^{B} w_b h_b(\boldsymbol{x}_i), \tag{A.5}$$

where B is the number of weak learners in the ensemble, w_b is the weight assigned to the bth learner, and $h_b(x)$ is the prediction function of the bth weak learner. The training process of gradient boosting involves updating the weights of the weak learners and minimizing the loss function using gradient descent. Denoting the model at the bth step by F_b, then,

$$F_{b+1}(\boldsymbol{x}_i) = F_b(\boldsymbol{x}_i) + h_b(\boldsymbol{x}_i) = y_i, \tag{A.6}$$

or

$$h_b(\boldsymbol{x}_i) = y_i - F_b(\boldsymbol{x}_i). \tag{A.7}$$

From the observation that residuals $h_b(\boldsymbol{x}_i)$ for a given model are proportional to the negative gradients of the mean squared error:

$$L_{\text{MSE}} = \frac{1}{n} \sum_{i=1}^{n} \left[y_i - F_b(\boldsymbol{x}_i) \right]^2 \tag{A.8}$$

$$-\frac{\partial L_{\text{MSE}}}{\partial F_b(\boldsymbol{x}_i)} = \frac{2}{n} \left[y_i - F_b(\boldsymbol{x}_i) \right] = \frac{2}{n} h_b(\boldsymbol{x}_i). \tag{A.9}$$

A.6. Light gradient boosting machine

Light gradient boosting machine (LightGBM) is a gradient boosting framework developed to be more efficient and scalable than traditional gradient boosting methods. The main difference between LightGBM and other gradient-boosting methods lies in the way the trees are constructed. Instead of growing trees level-wise, LightGBM grows trees leaf-wise, which means that it adds new leaves to the existing tree that has the highest reduction in the loss function, rather than adding new leaves at each level. This leaf-wise growth strategy enables LightGBM to converge faster and obtain a more accurate model with fewer iterations.

A.7. Extreme gradient boosting

XGBoost, an acronym for extreme gradient boosting, represents a refined iteration of gradient boosting machines (GBM). This advanced algorithm incorporates a regularized learning objective that controls the model's complexity, thus mitigating the risk of overfitting. The regularization improves the model's generalization capabilities, making it more robust to noise and less prone to overfitting. Furthermore, XGBoost's efficient and parallelizable training process makes it a popular, efficient, and robust choice for large-scale machine learning tasks in data science applications.

A.8. Support vector regression

The main idea behind the support vector regression (SVR) is to find a function that approximates the relationship between the input features and the target variable, while allowing a predefined error tolerance ϵ. The objective of SVR is to minimize the following function:

$$F(\boldsymbol{w}, \varepsilon) = \frac{1}{2} \|\boldsymbol{w}\|^2 + C \sum_{i=1}^{n} \varepsilon_i, \tag{A.10}$$

where C determines the penalty assigned to ϵ. The optimization is subjected to the following constraint:

$$|y_i - (\boldsymbol{w}\boldsymbol{x}_i + w_0)| \leq \xi_i, \tag{A.11}$$

where ξ_i depends on magnitude of ε_i and a margin e:

$$\xi_i = \begin{cases} 0 & \text{if } |\varepsilon_i| \leq e \\ |\varepsilon_i| - e & \text{otherwise.} \end{cases} \tag{A.12}$$

In practice, the parameter C determines the trade-off between maximizing the margin and minimizing the training error. Larger values of C will result in a smaller margin and more accurate predictions on the training data, but may lead to overfitting. To solve the optimization problem, SVR employs the kernel trick, which allows the algorithm to operate in a higher-dimensional feature space without explicitly computing the feature mapping function. Common kernel functions used in SVR include linear, polynomial, radial basis function, and sigmoid kernels. However, SVR is computationally expensive, especially for the large datasets in this work. Therefore, linear kernel function is selected in this work.

A.9. k-nearest neighbor

The k-nearest neighbor (kNN) method is a widely used instance-based learning algorithm that was originally developed for pattern classification and has been later applied to regression problems. The advantages of kNN include but are not limited to: (1) simple and easy to understand—kNN is a straightforward algorithm that requires minimal tuning and can be easily explained; (2) nonparametric—kNN does not make assumptions about the underlying data distribution, making it well-suited for handling complex, nonlinear relationships; and (3) adaptive to local data structures—kNN can adapt to local variations in the data, providing accurate predictions in areas where data is densely sampled. Therefore, kNN has been widely employed in the domain of solar irradiance or forecasting research [41–43].

The kNN finds the k nearest training samples to a given input data point and making predictions based on the majority vote (in classification problems) or the average value (in regression problems) of these neighbors. Mathematically, the kNN regression prediction $\hat{y}_{n+1}$ for a new input $\boldsymbol{x}_{n+1}$ can be defined as:

$$\hat{y}_{n+1} = \frac{1}{k}\sum_{i=1}^{n} y_i \mathbb{I}_{\{i \in N_k(x)\}}, \tag{A.13}$$

where $\mathbb{I}$ is an indicator function, $N_k(x)$ represents the k nearest neighbors of x in the training data. The distance metric, usually being the Euclidean distance, is used to find the nearest neighbors.

In this work, several strategies are employed to optimize kNN models to ensure accurate predictions. First, kNN is sensitive to the choice of k. This work sets $k = 10$ based on the recommendation in the literature [44]. Moreover, the performance of kNN is affected by feature scaling. Therefore, we employ the standard normalization approach to scale the input data, ensuring all features have similar scales and improving the performance of the kNN model. Standard normalization, also known as Z-score normalization, transforms each feature to have a mean of 0 and a standard deviation of 1. This normalization step ensures that the distance metric in the kNN model is not dominated by features with larger numerical ranges, thus allowing for more accurate predictions.

A.10. Deep learning and artificial neural network

Deep learning, as a subfield of machine learning, can learn complex hierarchical representations from raw data. One of the most popular deep learning approaches is an artificial neural network (ANN) with three or more layers. ANN has been successfully applied to various tasks due to its ability to model complex nonlinear mappings [5,45]. A commonly used ANN structure is the fully connected network or multilayer perceptron, which is a popular data-driven tool for pattern recognition, data classification, and regression [43,46]. The ANN model consists of neurons organized into layers, with the layers between the input and output layers called hidden layers. Neurons take in the weighted sum of inputs through various layers and produce an output using an activation function. Before training the ANN model, input data is normalized using standard normalization techniques to ensure all features have similar scales. This preprocessing step is crucial for the effective training and performance of the ANN model. More ANN optimization strategy is referred to [5]. In this work, we use the adaptive moment estimation (ADAM) algorithm [47] as the optimizer for training the deep learning model. ADAM is a popular choice due to its ability to dynamically adjust learning rates for each parameter during training. The hyperparameters are set as: the exponential decay rates $\beta_1 = 0.9$, $\beta_2 = 0.999$, tolerance $\epsilon = 10^{-8}$, and the learning rate $\alpha = 0.001$.

References

[1] Yang D. Estimating 1-min beam and diffuse irradiance from the global irradiance: A review and an extensive worldwide comparison of latest separation models at 126 stations. Renew Sustain Energy Rev 2022;159:112195.
[2] Sengupta M, Xie Y, Lopez A, Habte A, Maclaurin G, Shelby J. The National Solar Radiation Data Base (NSRDB). Renew Sustain Energy Rev 2018;89:51–60.
[3] Mayer MJ, Yang D. Probabilistic photovoltaic power forecasting using a calibrated ensemble of model chains. Renew Sustain Energy Rev 2022;168:112821.
[4] Wang W, Yang D, Hong T, Kleissl J. An archived dataset from the ECMWF Ensemble Prediction System for probabilistic solar power forecasting. Sol Energy 2022;248:64–75.
[5] Inman RH, Pedro HTC, Coimbra CFM. Solar forecasting methods for renewable energy integration. Progr Energy Combust Sci 2013;39(6):535–76.
[6] Chu Y, Li M, Coimbra CFM, Feng D, Wang H. Intra-hour irradiance forecasting techniques for solar power integration: A review. iScience 2021;24(10).
[7] Mayer MJ. Influence of design data availability on the accuracy of physical photovoltaic power forecasts. Sol Energy 2021;227:532–40.
[8] Chu Y, Pedro HTC, Coimbra CFM. Hybrid intra-hour DNI forecasts with sky image processing enhanced by stochastic learning. Sol Energy 2013;98:592–603.
[9] Mayer MJ. Benefits of physical and machine learning hybridization for photovoltaic power forecasting. Renew Sustain Energy Rev 2022;168:112772.
[10] Mayer MJ. Influence of design data availability on the accuracy of physical photovoltaic power forecasts. Sol Energy 2021;227:532–40.
[11] Hollands KGT. A derivation of the diffuse fraction's dependence on the clearness index. Sol Energy 1985;35(2):131–6.
[12] Hollands KGT, Crha SJ. An improved model for diffuse radiation: Correction for atmospheric back-scattering. Sol Energy 1987;38(4):233–6.
[13] Gueymard CA, Ruiz-Arias JA. Extensive worldwide validation and climate sensitivity analysis of direct irradiance predictions from 1-min global irradiance. Sol Energy 2016;128:1–30.
[14] Engerer NA. Minute resolution estimates of the diffuse fraction of global irradiance for Southeastern Australia. Sol Energy 2015;116:215–37.
[15] Yang D. Temporal-resolution cascade model for separation of 1-min beam and diffuse irradiance. J Renew Sustain Energy 2021;13(5):056101.
[16] Yang D, Gu Y, Mayer MJ, Gueymard CA, Wang W, Kleissl J, et al. Regime-dependent 1-min irradiance separation model with climatology clustering. Renew Sustain Energy Rev 2024;189:113992.
[17] Every JP, Li L, Dorrell DG. Köppen–Geiger climate classification adjustment of the BRL diffuse irradiation model for Australian locations. Renew Energy 2020;147:2453–69.
[18] Starke AR, Lemos LFL, Boland J, Cardemil JM, Colle S. Resolution of the cloud enhancement problem for one-minute diffuse radiation prediction. Renew Energy 2018;125:472–84.
[19] Yang D. Solar radiation on inclined surfaces: Corrections and benchmarks. Sol Energy 2016;136:288–302.
[20] Driemel A, Augustine J, Behrens K, Colle S, Cox C, Cuevas-Agulló E, et al. Baseline Surface Radiation Network (BSRN): structure and data description (1992–2017). Earth Syst Sci Data 2018;10(3):1491–501.
[21] Forstinger A, Wilbert S, Kraas B, Peruchena CF, Gueymard CA, Collino E, et al. Expert quality control of solar radiation ground data sets. In: Solar world congress 2021. Virtual conference: International Solar Energy Society; 2021.
[22] Beck HE, Zimmermann NE, McVicar TR, Vergopolan N, Berg A, Wood EF. Present and future Köppen–Geiger climate classification maps at 1-km resolution. Sci data 2018;5(1):1–12.
[23] Mayer MJ, Yang D. Calibration of deterministic NWP forecasts and its impact on verification. Int J Forecast 2023;39(2):981–91.

[24] Yang D, Alessandrini S, Antonanzas J, Antonanzas-Torres F, Badescu V, Beyer HG, et al. Verification of deterministic solar forecasts. Sol Energy 2020;210:20–37, Special Issue on Grid Integration.
[25] Bright JM, Engerer NA. Engerer2: Global re-parameterisation, update, and validation of an irradiance separation model at different temporal resolutions. J Renew Sustain Energy 2019;11(3):033701.
[26] Ridley B, Boland J, Lauret P. Modelling of diffuse solar fraction with multiple predictors. Renew Energy 2010;35(2):478–83.
[27] Paulescu E, Blaga R. A simple and reliable empirical model with two predictors for estimating 1-minute diffuse fraction. Sol Energy 2019;180:75–84.
[28] Wilson AM, Jetz W. Remotely sensed high-resolution global cloud dynamics for predicting ecosystem and biodiversity distributions. PLoS Biol 2016;14(3):1–20.
[29] Yang D, Gueymard CA. Probabilistic merging and verification of monthly gridded aerosol products. Atmos Environ 2021;247:118146.
[30] Hersbach H, Bell B, Berrisford P, Hirahara S, Horányi A, Muñoz-Sabater J, et al. The ERA5 global reanalysis. Q J R Meteorol Soc 2020;146(730):1999–2049.
[31] Murphy AH, Winkler RL. A general framework for forecast verification. Mon Weather Rev 1987;115(7):1330–8.
[32] Yang D, Perez R. Can we gauge forecasts using satellite-derived solar irradiance? J Renew Sustain Energy 2019;11(2):023704.
[33] Yang D, Wang W, Bright JM, Voyant C, Notton G, Zhang G, et al. Verifying operational intra-day solar forecasts from ECMWF and NOAA. Sol Energy 2022;236:743–55.
[34] Yang D, Bright JM. Worldwide validation of 8 satellite-derived and reanalysis solar radiation products: A preliminary evaluation and overall metrics for hourly data over 27 years. Sol Energy 2020;210:3–19, Special Issue on Grid Integration.
[35] Yagli GM, Yang D, Gandhi O, Srinivasan D. Can we justify producing univariate machine-learning forecasts with satellite-derived solar irradiance? Appl Energy 2020;259:114122.
[36] Thornton C, Hutter F, Hoos HH, Leyton-Brown K. Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In: Proc. ACM SIGKDD int. conf. knowl. discovery data mining. 2013, p. 847–55.
[37] Feurer M, Klein A, Eggensperger K, Springenberg J, Blum M, Hutter F. Efficient and robust automated machine learning. In: Proc. advances neural inform. process. syst.. 2015, p. 2962–70.
[38] Karasu S, Altan A. Recognition model for solar radiation time series based on random forest with feature selection approach. In: 2019 11th International conference on electrical and electronics engineering. IEEE; 2019, p. 8–11.
[39] Montgomery DC, Peck EA, Vining GG. Introduction to linear regression analysis. John Wiley & Sons; 2021.
[40] Peng Z, Yoo S, Yu D, Huang D. Solar irradiance forecast system based on geostationary satellite. In: 2013 IEEE international conference on smart grid communications. IEEE; 2013, p. 708–13.
[41] Paoli C, Voyant C, Muselli M, Nivet M-L. Forecasting of preprocessed daily solar radiation time series using neural networks. Solar Energy 2010;84(12):2146–60.
[42] Pedro HTC, Coimbra CFM. Nearest-neighbor methodology for prediction of intra-hour global horizontal and direct normal irradiances. Renew Energy 2015;80:770–82.
[43] Pedro HTC, Coimbra CFM, David M, Lauret P. Assessment of machine learning techniques for deterministic and probabilistic intra-hour solar forecasts. Renew Energy 2018;123:191–203.
[44] Chu Y, Coimbra CFM. Short-term probabilistic forecasts for direct normal irradiance. Renew Energy 2017;101:526–36.
[45] Anagnostos D, Schmidt T, Cavadias S, Soudris D, Poortmans J, Catthoor F. A method for detailed, short-term energy yield forecasting of photovoltaic installations. Renew Energy 2019;130:122–9.
[46] Yang D, Kleissl J, Gueymard CA, Pedro HTC, Coimbra CFM. History and trends in solar irradiance and PV power forecasting: A preliminary assessment and review using text mining. Sol Energy 2018;168:60–101.
[47] Kingma DP, Ba J. Adam: A method for stochastic optimization. 2014, arXiv preprint arXiv:1412.6980.