40th European Photovoltaic Solar Energy Conference and Exhibition

This presentation was selected by the Sc. Committee of the EU PVSEC 2023 for submission of a full paper to one of the EU PVSEC's collaborating peer-reviewed journals.

Evaluation of Irradiance Decomposition Models and Their Predictors

Jacob K. Thorning*, Thøger K. Hass, Sergiu V. Spataru
Email: J. K. Thorning jkrtho@dtu.dk
Technical University of Denmark (DTU), Department of Electrical and Photonics Engineering
Frederiksborgvej 399, 4000 Roskilde, Denmark

ABSTRACT: This study presents a comprehensive approach to developing models for decomposing global solar radiation in to its diffuse and direct components, a critical step in solar energy applications. Utilizing data from multiple geographical regions and Köppen-Geiger climate zones, the research ensures a broad understanding of the environmental factors affecting solar radiation. The study meticulously selects predictors through multiple feature selection techniques, focusing on capturing essential information while minimizing redundancy. Single-predictor models are developed to provide insights into the relationships between selected predictors and the diffuse fraction, while two multi-predictor models demonstrates the potential for more accurate estimations by leveraging the collective predictive power of multiple variables. One of the developed models, called Nested, performs better than other state-of-the-art universal models, while performing slightly worse than the best climate-specific model in terms of three performance metrics. The research lays a solid foundation for future studies and practical applications in the solar energy sector, emphasizing the need for further testing of techniques and procedures for decomposition model development.
Keywords: solar radiation, diffuse fraction, predictors, feature selection, model development

1 INTRODUCTION

Decomposition of Global Horizontal Irradiance (GHI) into Diffuse Horizontal Irradiance (DHI) and Direct Normal Irradiance (DNI) components is crucial for various applications, including solar energy resource assessment, solar collector performance evaluation, and climate studies. Accurate estimation of diffuse and direct solar irradiance improves the design and efficiency of photovoltaic (PV) and concentrating solar power (CSP) systems. Understanding the behavior of these components enables better modeling of the Earth's radiation budget and its effects on the climate. Improved predictions facilitate lower uncertainty in investment in PV projects, more efficient planning and allocation of energy resources, contributing to cost-effectiveness. Furthermore, enhanced grid integration through better synchronization and balancing of supply and demand is vital for grid stability and reliability. Since G_{poa} has the most significant influence on the expected power output, it is essential to estimate it reliably. However, due to the high costs associated with measuring DHI or DNI, it is more common to measure only GHI [1]. Consequently, estimating G_{poa} requires two modeling steps: (1) decomposing GHI into DHI and DNI with a decomposition model, which generally follows the form of Equation 1, where f is a decomposition model, t is date-time, lat is latitude and lon is longitude.

$$DHI = GHI \cdot f(GHI, t, lat, lon) \qquad (1)$$

(2) transposing DHI, DNI, and ground-reflected irradiance into Global Tilted Irradiance (GTI) with a transposition model [2] and for more accuracy, shading simulations [1, 3]. These modeling steps often introduce the highest uncertainty in PV system performance simulations [4–6].

Over the past six decades, numerous decomposition models have been proposed to estimate diffuse and direct solar irradiance components from GHI measurements [7, 8]. These models have evolved to address various challenges, such as the need for higher accuracy, higher temporal resolution, improved performance under diverse climatic conditions, and incorporation of additional factors influencing solar irradiance. The earliest well recognized model for solar irradiance decomposition was proposed by Liu and Jordan [9]. This simple empirical model, based on the ratio of daily diffuse, k_d, to global solar radiation, laid the foundation for future developments in the field. In Skartveit, Olseth, and Tuft [10] a model for estimating the hourly diffuse fraction considering the variability in the atmospheric conditions and surface albedo was introduced. This model marked a significant improvement over previous approaches by being highly analytical and incorporating variability in the ratio GHI to extraterrestrial GHI, called clearness index k_t, and regional surface albedo, which affect k_d beyond k_t. A significant development came from Boland, Scott, and Luther [11], where the authors used a logistic function to map k_t values to k_d. The inspiration for using a logistic function was a visual impression through a k_d-k_t scatterplot. The "BRL" model, presented in Ridley, Boland, and Lauret [12], built upon previous research by incorporating multiple predictors in a logistic function to estimate k_d. This approach allowed for improved accuracy in modeling

10.4229/EUPVSEC2023/4CO.9.2
020362-001

k_d under various climatic conditions. The model from Engerer [13] was designed to estimate k_d with minute resolution for southeastern Australia. Being the first model developed with 1-minute average data, this model provided a higher temporal resolution, making it suitable for the accuracy and timescale of modern system performance simulations, and set the standard for decomposition model time resolution and accuracy going forward [7]. Starke et al. [14] proposed models for each category of Köppen-Geiger climate zones (KGCZs) based on a single framework, with focus on high performance during Cloud Enhancement (CE) events. In Yang and Boland [15] satellite-augmented models were proposed, leveraging DHI and GHI estimated from satellite imagery to improve the accuracy of the k_d estimation. In Yang [16], a temporal-resolution cascade model was introduced, where k_d is estimated with a time average of k_d as a predictor. Recently, two new models of different complexity were proposed in Paulescu and Paulescu [17], based on rational functions, named M1 and M2. M1 uses 8 predictors and M2 uses 3 predictors, included as a more accessible alternative to most recent models which use 6 or more predictors. In each work the authors found that at least one of the proposed models outperformed the established models.

In this study, we address the concerns raised by prior research regarding the potential for overfitting and the absence of a scientific foundation in solar irradiance decomposition models. In Gueymard and Ruiz-Arias [7] the authors observed that adding predictors generally improves model precision and repeatability; however, they also noted that there could be exceptions due to overfitting when incorporating numerous predictors without a clear basis. Similarly, Yang and Boland [15] emphasized the importance of understanding the atmospheric physics governing diffuse radiation, rather than merely modifying regression-based models or specifying alternative predictor sets.

The objective of this work is to establish a systematic procedure for predictor selection and model development. Through an in-depth analysis of the empirical relationships between predictors and the diffuse fraction, this study aims to establish a scientifically informed basis for the development of decomposition models. Employing feature selection techniques Variable Importance in Projection (VIP) of Partial Least Squares (PLS) regression, Maximum Relevance - Minimum Redundancy (mRMR), and Recursive Feature Elimination (RFE). This methodological approach is anticipated to contribute to the development of more accurate and reliable models, thereby enhancing the precision of solar irradiance predictions.

The rest of this paper is organized as follows: Section 2 describes the datasets used in this work, Section 3 presents the methods and results of predictor selection, Section 4 describes how single and multi-predictor models are developed, Section 5 presents a validation of two proposed models against models from the literature, Section 6 is a systematic discussion of the methods and results for each step of the process presented, and finally Section 7 presents the conclusions to this work.

2 DATA

The datasets used in this work are from 10 stations of the Baseline Solar Radiation Network (BSRN) network listed in Table I, where data from stations 1-5 are split randomly in to 80%/20% and used for model training and testing respectively, and data from stations 6-10 are used for model validation. Each station's data was sampled uniformly to match the station with the smallest sample size, ensuring equal representation across all stations. The training and testing datasets are used as a whole for the feature selection. Since the testing data is used in the iterative model development process, a fully independent dataset is required for validation. The sampling is performed after calculating the required predictor values. The training and testing data comprise 231662 samples from each of the 5 stations, amounting to a total of 1158310 samples. Similarly, for validation, each of the 5 stations contribute 232635 samples, yielding a total validation set of 1163175 samples. These post-Quality Control (QC) 1-minute datasets are from Gueymard et al. *BSRN data set for IEA-PVPS Task-16 Activity 1.4 Quality Control* [18]. The authors have made a significant contribution to the field by providing high-quality, reference datasets from various locations across the globe. Their efforts in collecting, curating, and sharing these valuable resources are gratefully acknowledged. The reader is referred to [19] for more information.

Using data from multiple locations with varying climates is crucial for developing a robust and generalizable model, as it captures a broad range of environmental and climatic conditions, enabling a model to identify underlying universal patterns and eliminate location-specific biases. This contributes to a more reliable predictor importance analysis and predictor selection process, as the methods will prioritize predictors that are significant across multiple climates and regions rather than those that are relevant only to a specific location.

3 PREDICTOR SELECTION

Table II compiles the symbols and expressions (or references) of predictors commonly employed in the literature [7, 8].

10.4229/EUPVSEC2023/4CO.9.2
020362-002

Table I: Selected BSRN stations representing multiple continents and Köppen-Geiger climate zones. Data from the first five stations are used for training and testing, while data from the last five stations are used for validation.

#	Code	Name	Country	Zone	Zone description
1	COC	Cocos Island	Australia	Aw	Equatorial savannah with dry winter
2	GOB	Gobabeb	Namibia	BWk	Desert climate, cold
3	GCR	Goodwin Creek	USA	Cfa	Warm temperate climate, fully humid, hot summer
4	TOR	Toravere	Estonia	Dfc	Snow climate, fully humid, cool summer and cold winter
5	NYA	Ny-Ålesund	Norway	ET	Tundra climate
6	BRB	Brasilia City	Brazil	Aw	Equatorial savannah with dry winter
7	FPE	Ford Peck	USA	BSk	Cold steppe climate
8	IZA	Tenerife	Spain	Csb	Warm temperate climate with dry summer
9	TIK	Tiksi	Russia	Dsd	Snow climate with dry summer, extremely continental
10	GVN	Dronning Maud Land	Antarctica	EF	Frost climate

Table II: Summary of predictors used in solar irradiance decomposition models. The table includes symbols and expressions (or references) of the predictors, which are derived from global horizontal irradiance, date, time of day, latitude, and longitude.

Symbol		Expression
Irradiance derivatives / k-indices		
k_t	=	E_{gh} / E_{0h}
k_{csi}	=	E_{gh} / E_{ghc}
k_{tc}	=	E_{ghc} / E_{0h}
Δk_{tc}	=	k_{tc} - k_t
k'_t	=	$\frac{k_t}{(1.031e^{-1.4/(0.9+9/AM)}+0.1)}$
k_{de}	=	max(0, 1 - E_{ghc}/E_{gh})
Derivatives of time and location		
θ_z		See [20, p 15]
Z	=	90 - θ_z
AM		See [21]
AST		See [20, p 11]
Time-series dependent		
K_t	=	$\frac{1}{60}\sum_{n=-30}^{30} k_{t,n}$
$\overline{K_t}$	=	$\frac{1}{1440}\sum_{n=0}^{1440} k_{t,n}$
V^{SO}	=	$[(k_{csi} - k_{csi-1})^2 + (k_{csi} - k_{csi+1})]^{1/2}$
V^P	=	$(\lvert k'_t - k'_{t-1}\rvert + \lvert k'_t - k'_{t+1}\rvert)/2$
ψ	=	$(k_{t,-1} + k_{t,+1})/2$
Cascade modelling		
k_d^s	=	${E_{dh}}^{sat}/{E_{gh}}^{sat}$
k_d^m		k_d estimated with e.g. model from Engerer [13]

We have limited our work to predictors derived from any combination of measured E_{gh}, date-time (year, day, and time of day), and geographical location (latitude, longitude), with E_{gh} being the sole measured variable.

3.1 Methods for predictor evaluation and selection

Partial Least Squares (PLS) regression is a method that combines the ideas of Principal Component Analysis (PCA) and linear regression to model the relationship between the predictor variables (features) and k_d (the target variable). PLS projects both the input features and the output target onto a new set of latent variables (which are analogous to the components of PCA), called PLS components. These components are chosen to maximize the covariance between the features and the target. VIP scores are a measure of the importance of each feature in the projection of the data using PLS regression [22]. The highest VIP scores indicate the most significant predictors in the PLS model.

Maximum Relevance - Minimum Redundancy (mRMR) is a filter-based method that aims to select predictors that are highly correlated with the target variable (maximum relevance) while minimizing the inter-predictor correlation (minimum redundancy). It does so by ranking predictors based on their relevance to the target variable and iteratively selecting predictors that contribute the most new information while minimizing the redundancy between them. The relevance of each predictor to k_d is assessed using the importance scores provided by a random forest regression model (from [23]). For each pair of predictors, the redundancy score is computed using Spearman correlation. Recursive Feature Elimination (RFE) operates by iteratively fitting the model, ranking the predictors based on their importance, and removing the least important predictor in each iteration. This process continues until the desired number of predictors is reached. In this work, decision trees are used as the regression model.

An aggregated average score is calculated by first normalizing the scores from each feature selection method individually, such that the best predictor has a score of 1 and all other predictors has a score less than 1 but greater than or equal to 0. The scores are combined by taking the average score for each predictor. Finally, the predictors are ranked based on

10.4229/EUPVSEC2023/4CO.9.2
020362-003

their average score across the three metrics.

3.2 k_d predictor selection results

A heatmap of the normalized scores derived from various feature selection methods is shown in Figure 1.

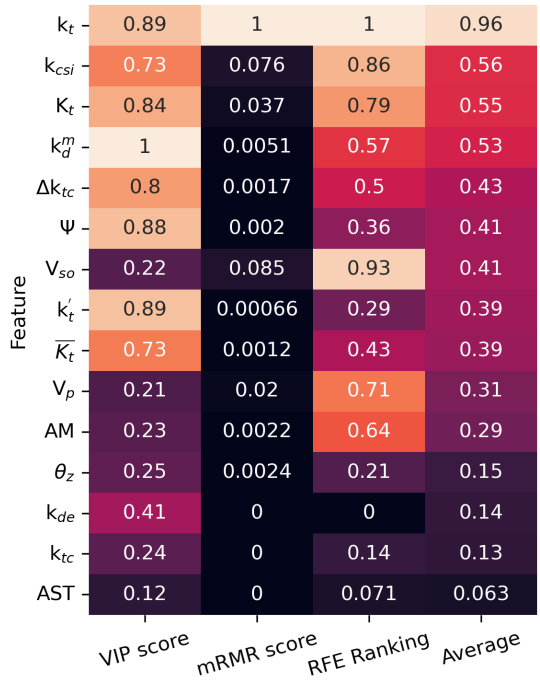

Figure 1: Heatmap of the normalized feature selection scores across different methods, sorted by their overall rank. The color intensity represents the relative importance of each predictor, with lighter colors indicating higher importance.

The top-ranked predictor, k_t, is identified as crucially important across all methods, with a high average score of 0.97. This is in line with our expectations [7, 8] k_t has the maximum mRMR score, indicating that it has the highest relevance to k_d. The low mRMR scores for the other predictors imply that these predictors either do not provide as much information about the outcome as k_t does, or, more likely, they may be somewhat redundant, sharing information that's already provided by other predictors in the predictor set. While other predictors might still be significant in their contribution to the model, their unique contributions are low. E.g. θ_z and Air Mass (AM) contain the same information, as they may be calculated as functions of each other. Due to the ranking being based on average scores, beyond k_t the only significant feature selection methods are VIP and RFE.

The second-ranked predictor, k_{csi}, also exhibits strong performance across all methods, when considering its mRMR score relative to the other predictors except k_t. The importance of K_t highlights the value of considering the temporal aspect of k_d estimation, reflecting the historical context and trends, capturing gradual changes in atmospheric conditions. ψ is ranked fourth, with an average score of 0.53. This predictor's importance also emphasizes the role of time-series information in predicting k_d, as ψ accounts for inertia in the atmosphere, similar to K_t. The fifth-ranked predictor, k_d^m, achieves the highest VIP score, but a relatively low RFE ranking, opposite the case of k_{csi}. This shows the importance of using multiple feature selection methods. The sixth-ranked predictor Δk_{tc} achieves an average score of 0.49, demonstrating a moderate level of importance across VIP and RFE. It is worth noting that some predictors, such as Apparent Solar Time (AST), exhibit consistently low scores across all methods, indicating that they may not contribute meaningfully to the prediction of k_d.

The analysis offers an understanding of the importance of different predictors in predicting k_d. By selecting the top six predictors, k_t, k_{csi}, K_t, k_d^m, Δk_{tc} and ψ, we have ensured a balance between capturing essential information and minimizing redundancy. This selection will provide a solid foundation for constructing models that accurately maps the relationships between these predictors and k_d.

4 DEVELOPMENT OF NEW MODELS

4.1 Single-predictor model development

In the pursuit to model k_d using the selected predictors individually, a modified logistic function is tailored. The basic logistic form in Equation 2 [12], where k_d is the target variable and x represents a predictor, is used as a starting point.

$$k_d = \frac{1}{1 + e^{(\beta_0 + \beta_1 x)}} \tag{2}$$

Recognizing that a more sophisticated approach may be advantageous, we started an iterative process to fine-tune this function, experimenting with replacing constants and adjusting parameter placement. The end result is Equation 3 where C, β_0, and β_1 are parameters optimized for each predictor.

$$k_d = C + \left| \frac{1}{1 + e^{(\beta_0 + \beta_1 x)}} - C \right| \tag{3}$$

In this formulation, C is a critical parameter positioned both within and outside of the absolute function. This placement provides the model with flexibility, allowing the function to pivot at a base level, C, and to start increasing with the predictor variables instead of the traditional decrease seen in basic logistic functions. Equation 3 is applied to k_t, Δk_{tc}, k_{csi}, K_t, and ψ independently, and each is fitted a set of optimized parameters. For the predictor k_d^m, we found a simple linear function with β_0 as intercept

10.4229/EUPVSEC2023/4CO.9.2
020362-004

and β_1 as slope to be the most fitting representation. The parameters are listed in Table III.

Table III: Optimized parameter values for the mathematical functions fitted for each predictor variable. The columns C, β_0, and β_1 represent the parameters of the logistic and linear functions as applicable. For the linear function applicable to k_d^m, C is not applicable (N/A).

Predictor	C	β_0	β_1
k_t	0.16130	-5.53387	9.42544
k_{csi}	0.19767	-7.48200	8.34717
K_t	0.13453	-5.62334	9.44307
ψ	0.16197	-5.47208	9.29277
k_d^m	N/A	-0.04510	1.00128
Δk_{tc}	0.20215	0.78571	-12.56507

Figure 2 displays the relationships between each predictor and k_d, visually justifying the functions chosen to map the predictors.

4.2 Multi-Predictor model development

In pursuit of achieving a more accurate estimation of the diffuse fraction, we incorporate multiple predictors simultaneously, leveraging their collective predictive capacity. Our approach has been two-fold.

In the first approach, we have selected the predictor of the single-predictor model with the highest R^2 value (coefficient of determination) as the principal predictor. Subsequently, additional predictors are incorporated into the logistic function iteratively, each being included in the exponential part of the denominator. For instance, the function to include one additional predictor with a logistic relationship is Equation 4.

$$k_d = C + \left| \frac{1}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}} - C \right| \quad (4)$$

For the inclusion of k_d^m, which has a linear relationship with k_d, it is added as a separate term as illustrated in Equation 5

$$k_d = C + \left| \frac{1}{1 + e^{(\beta_0 + \beta_1 x_1)}} - C \right| + \beta_2 x_2 \quad (5)$$

In these functions, x_1 and x_2 represent the predictors, while β_0, β_1, β_2, and C are the model parameters. The β_1 values in Table III serve as the initial guesses for each new parameter in this iterative process. After calculating R^2 of incorporating each so-far unused predictor individually, the predictor which provides the largest improvement in R^2 is added to the model and the process is repeated until all predictors are included. This model is referred to as Nested, due to the incorporation of the predictors in the argument of the same logistic function. This iterative process provides information about the value of including each of the predictors considered post feature selection, and it ensures that the parameter solution is globally optimal at all times.

The second approach, in contrast, utilizes a weighted sum of all the individually fitted single-predictor models, as shown in Equation 6 where w_i denotes the weight of the i-th function in the order shown in Table III, f_i is the function, x_i is the corresponding predictor values, and $\mathbf{p}_i$ represents the set of parameters for the function.

$$k_d = C + \sum_i (w_i \cdot f_i(x_i, \mathbf{p}_i)) \quad (6)$$

Each model function is associated with a weight parameter, which is determined through an optimization process. This ensemble-based approach aims to harness the collective predictive power of all the predictors by adjusting their individual influence based on the associated weights. This model is called Summed.

4.3 Model development results

The Nested model demonstrated a strong performance when used on the test data, with an R^2 score of 0.87034. The order of predictors included and the R^2 at the end of each iteration is shown in Table IV.

Table IV: Improvement in coefficient of determination (R^2) with sequential addition of predictors in the developed model. Each row represents a new iteration where a predictor is added to the model.

Iteration	Added predictor	R^2	Improvement
1	k_t	0.81477	-
2	Δk_{tc}	0.85071	4.41%
3	K_t	0.86471	1.65%
5	k_d^m	0.86730	0.30%
4	ψ	0.87659	1.07%
6	k_{csi}	0.87713	0.06%

To further improve its predictive power, we capped the predicted values at 1 and introduced an additional parameter, C_2, in the numerator of the function. These modifications increased the R^2 score to 0.88452. Since multi-predictor fitting is significantly different from single-predictor fitting, we tested fitting the parameters with three modifications: 1) changing the 1 in the denominator to a parameter, 2) removing C, 3) removing the absolute function and the C with a negative sign.

There was nearly no difference on the four functions in terms of R^2, which leads us to nominate the simplest version, Equation 7, which had an R^2 of 0.88525 when applied to the test data. The param-

10.4229/EUPVSEC2023/4CO.9.2
020362-005

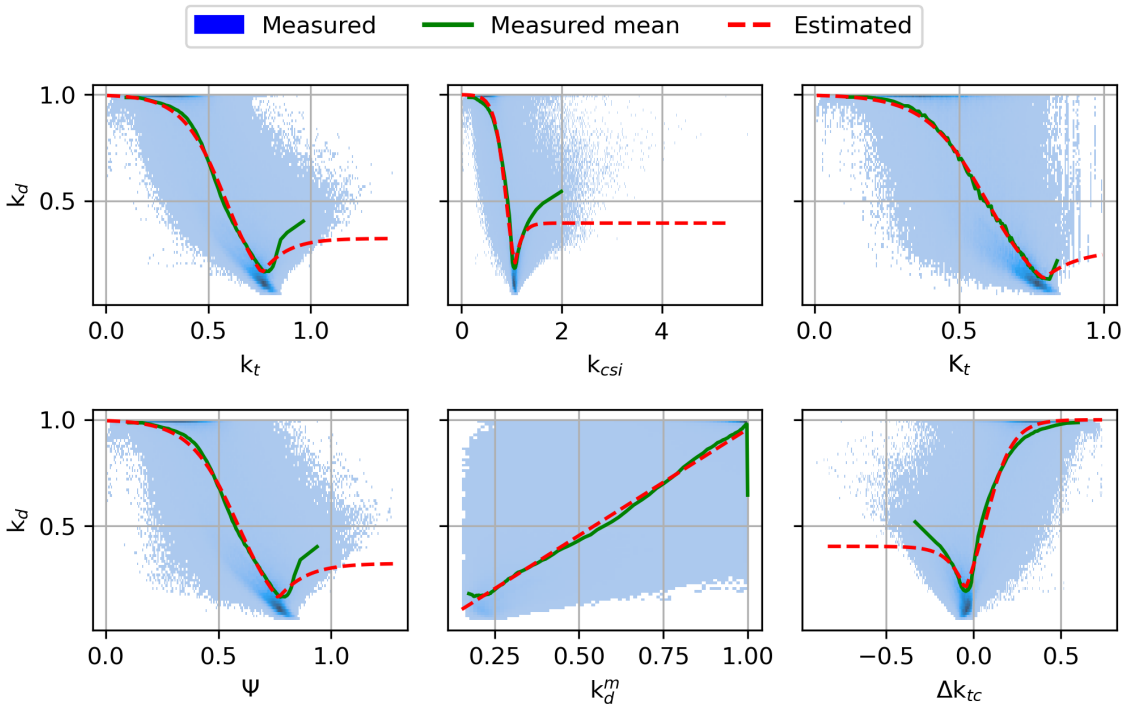

Figure 2: The relationships between the diffuse fraction (k_d) and each predictor variable. The blue histograms represent the measured data. The solid green line represents the mean trend, calculated within bins of the predictor variable. The dashed red line represents the estimated relationship derived from the respective functions for each predictor variable. The consistency between the green and red lines signifies the accuracy of our singe-predictor.

eters that produced the best fit are C: 0.45292, β_0: -13.49189, β_1: -6.48912, β_2: 1.48927, β_3: 12.03733, β_4: 4.07416, β_5: 8.08428, β_6: 0.53947.

The Summed model resulted in an R^2 score of 0.88167. The optimal weights are C: -0.09954, ω_0: 0.13129, ω_1: -0.00440, ω_2: 0.37380, ω_3: 0.16500, ω_4: 0.03460, ω_5: 0.51928.

5 MODEL VALIDATION

To assess the performance of the Nested and Summed models, five models from literature are selected. The models selected are Engerer2 [13], Engerer4 [24], Paulescu [25], Starke3 [26], Yang4 [16]. The selection is based on the best performing models in [8], from which the naming convention is also used. The recently proposed models from [17] are not included due to substantial discrepancies observed between the the first model's (M1) estimated k_d values and the corresponding measured values, perhaps caused by a typo in the parameters listed in the publication.

To understand the performance of each model, three metrics were employed: Normalized Mean Bias Deviation (nMBD), Normalized Root Mean Square Deviation (nRMSD), and coefficient of determination (R^2). The nMBD measures the average deviation of predicted values from the target values, normalized by the mean of the observations. It provides a measure of the overall bias in the predictions; a value of 0 indicates no bias. The nRMSD measures the average spread of the residuals, normalized by the mean of the observations, and thus provides a measure of the accuracy of the predictions; the lower the value, the better the prediction. The R^2 measures the proportion of the variance in the dependent variable that is predicted from the independent variables; a value close to 1 indicates a strong predictive power.

Tables V, VI, and VII present the R^2, nMBD, and nRMSD results respectively, for each model across each station, and the results are visualized in Figure 3. Performance was averaged across all stations to create a singular rating for each model per metric, and each model was accordingly assigned a rank. For nMBD the relative performance, averages and ranks are based on absolute values of the nMBD, i.e. mean absolute deviation. The relative performance of each model is indicated row wise in green.

The Nested model consistently presents excellent performance across the majority of stations and

$$k_d = \min\left(1, \frac{C}{1+\mathrm{e}^{(\beta_0+\beta_1 k_t+\beta_2 \Delta k_{tc}+\beta_3 K_t+\beta_4 \psi+\beta_5 k_{csi})}} + \beta_6 k_d^m\right) \quad (7)$$

10.4229/EUPVSEC2023/4CO.9.2
020362-006

Table V: R^2 values for different models across various zones. Each cell represents the R^2 value of the corresponding model at a specific station. The highest value in each row, indicating the best performing model for that station, is highlighted.

Data	Zone	Engerer2	Engerer4	Paulescu	Starke3	Yang4	Nested	Summed
BRB	A	0.937	0.945	0.940	0.954	0.937	0.950	0.948
FPE	B	0.951	0.969	0.965	0.971	0.953	0.975	0.974
GVN	E	0.804	0.810	0.783	0.893	0.806	0.828	0.838
IZA	C	0.964	0.979	0.976	0.979	0.966	0.986	0.977
TIK	D	0.960	0.956	0.945	0.963	0.965	0.963	0.964
Average	All	0.923	0.932	0.922	0.952	0.925	0.940	0.940
Rank	All	6	4	7	1	5	2	3

Table VI: nMBD [%] values for different models across various zones. Each cell represents the nMBD value of the corresponding model at a specific station. The relative performance of each model is indicated row wise in green. The relative performance, averages and ranks are based on absolute values of the nMBD, i.e. mean absolute deviation.

Data	Zone	Engerer2	Engerer4	Paulescu	Starke3	Yang4	Nested	Summed
BRB	A	5.606	-1.748	-2.274	-2.259	8.161	3.182	4.803
FPE	B	8.472	-4.065	-5.200	-6.527	9.579	-0.655	0.422
GVN	E	-23.496	-46.549	-55.449	-31.123	-24.151	-42.025	-41.186
IZA	C	9.094	4.361	-12.332	-8.372	7.923	-0.075	12.359
TIK	D	-6.960	-12.659	-9.479	-7.346	-5.169	-9.511	-7.387
Average	All	10.726	13.876	16.947	11.125	10.996	11.090	13.231
Rank	All	1	6	7	4	2	3	5

Table VII: nRMSD [%] values for different models across various zones. Each cell represents the nRMSD value of the corresponding model at a specific station. The relative performance of each model is indicated row wise in green.

Data	Zone	Engerer2	Engerer4	Paulescu	Starke3	Yang4	Nested	Summed
BRB	A	40.591	38.014	39.645	34.652	40.769	36.311	36.794
FPE	B	35.709	28.595	30.158	27.378	35.107	25.393	25.997
GVN	E	71.696	70.715	75.462	52.891	71.444	67.286	65.308
IZA	C	30.896	23.390	25.253	23.610	29.876	19.111	24.733
TIK	D	32.533	34.147	38.113	31.255	30.463	31.001	30.571
Average	All	42.285	38.972	41.726	33.957	41.532	35.820	36.681
Rank	All	7	4	6	1	5	2	3

climatic zones. Compared to other universal models (i.e. not KGCZ-specific) it displays superior ability to generalize across different climate conditions in terms of all three metrics, confirming its robust nature. Overall the best performing model is Starke3, which utilizes parameters fitted for each KGCZ, which makes the Nested the best performing universal model based on the specific metrics and validation dataset. The poor performance of Yang4 in terms of R^2 and nRMSD is surprising, as it has recently been crowned the best performing model in a major study utilizing data from 126 different stations [8].

6 DISCUSSION

6.1 Predictor selection

The predictor selection process employs a combination of methods, including PLS regression, VIP scores, mRMR, and RFE. While the integration of multiple methods is great for its comprehensiveness, there are several aspects that warrant critical discussion. While PLS regression is effective in handling multi-collinearity, its reliance on linear relationships may not capture complex non-linear interactions between predictors and the target variable. Additionally, VIP scores, which are derived from PLS, may not always provide clear-cut distinctions between relevant and irrelevant features, especially in cases

10.4229/EUPVSEC2023/4CO.9.2
020362-007

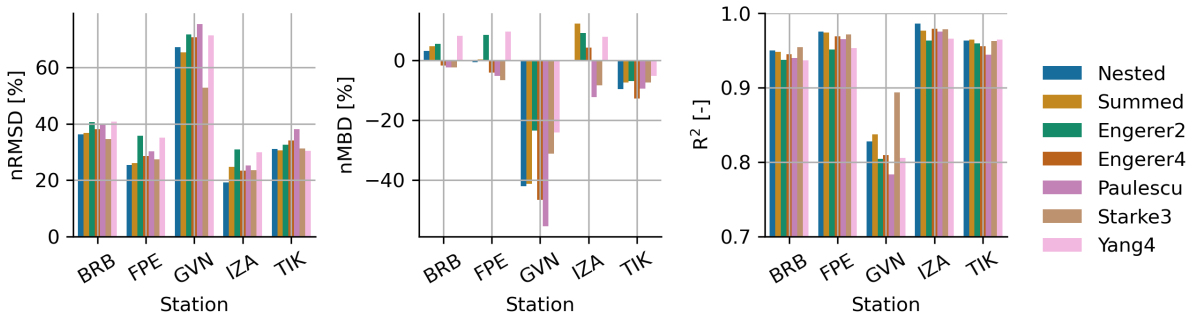

Figure 3: Station-specific model performance evaluated across different metrics. The bars represent model performance at each station in terms of Normalized Root Mean Square Deviation (nRMSD), Normalized Mean Bias Deviation (nMBD), and Coefficient of Determination (R^2).

where the number of features is large. mRMR does not take into account potential interactions between features, which can be critical in complex systems such as solar radiation prediction. While RFE is a powerful feature selection technique, it is computationally expensive, especially for large datasets. Additionally, by iteratively removing features, there is a risk of losing information that might have been captured by subsets of features. Using machine learning for feature selection may not account for the fundamental mathematical relationships in regression models. While machine learning focuses on optimizing performance through complex relationships, regression models often seek to capture simpler, interpretable mathematical patterns. This discrepancy necessitates careful feature selection in regression modeling.

The feature selection process is not validated against a separate dataset, so it is unknown if the results are specific to the combination of climates which the 5 stations belong to, i.e. the selected predictors may be overfitted to the training dataset.

6.2 Model development

While the logistic function is traditionally used for classification problems, for decomposition models it has been adapted for regression since its first use [11]. The logistic function inherently has an S-shape, which might not be the best representation for the relationships between predictors and k_d in all cases, particularly for k_t values above 0.7, where CE events occur. To circumvent this an absolute term was introduced, which appeared a good fit for single-predictor regression, but in the Nested model the absolute term was actually not utilized, which became apparent as fitted parameter values were identical with and without the absolute term. While introducing additional parameters outside of the exponential part of the denominator such as C or C_1 and C_2 may improve the fit, the introduction of additional parameters without a theoretical basis can lead to overfitting, where the model may fit the noise in the training data rather than capturing the underlying trend.

In terms of evaluating the models during development, reliance on R^2 as the sole metric might not be sufficient. R^2 can sometimes be a misleading indicator of model performance, especially in cases with noise, data clustering, or when the model is overfit. Other metrics such as nMBD or nRMSD could have been used in addition to R^2, however this was avoided for simplicity. The Summed model takes a more ensemble-based approach. It is opportune that this approach tries to harness the collective predictive power of all single-predictor models. However, it is essentially a linear combination of the outputs of single-predictor models. This linearity could limit the model's ability to capture complex, non-linear relationships between predictors and the target variable, and a different ensemble method could prove superior.

6.3 Validation

Since Yang [8], Yang4 has been anticipated to exhibit superior performance among the models. However, this expectation was not met in the current study. Interestingly, a similar discrepancy was observed in the work of Paulescu and Paulescu [17]. In Paulescu and Paulescu [17], Yang4 was found to be among the least effective models across the majority of the 36 stations included in their analysis, with Engerer2 performing the poorest. This observation aligns with the findings of the present study. In Yang [8], the stations utilized for validation are anonymized and denoted using codes such as A1, A2, B1, B2, etc., where the letter represents the major KGCZs and the number serves as a unique identifier within that zone. In contrast, [16, 17, 24, 26] provides explicit station codes, allowing for the identification of the specific stations from which the evaluation metrics are derived. The lack of transparency in station identification in Yang [8] precludes a direct comparison of results with those of Paulescu and Paulescu [17] or the present study. This is particularly regrettable given that both the current study and

10.4229/EUPVSEC2023/4CO.9.2
020362-008

Paulescu and Paulescu [17] observe a suboptimal performance of Yang4, contrary to expectations. The inability to directly compare results hampers the assessment of whether the observed discrepancies are attributable to the limited number of stations or other factors.

Since the validation dataset is chosen to be rather small relative to recent works on decomposition models, it is pertinent to discuss the contrasting methodologies employed for model validation, specifically focusing on the utilization of data from meteorological stations across various KGCZs. Traditionally [7, 8, 17, 26] model validation is conducted using data from numerous different stations spanning various KGCZs. The number of samples per station and the number of stations per climate zone are typically dictated by the data available, with authors leveraging all accessible data to ensure as extensive a validation as possible. Such an approach has advantages, as utilizing a large dataset from numerous stations provides a more comprehensive overview of the model's performance across diverse conditions. The extensive dataset can help in identifying patterns and trends that may not be evident with a smaller dataset, thereby contributing to the robustness of the validation. However, the varying number of samples per station can lead to inconsistencies in the validation results, and some climate zones may be over-represented due to the availability of more data, which can bias the validation results.

In contrast, in this work, a different approach is adopted where a uniform number of samples from five different stations, one station from each major KGCZ, is used for model validation. By using a uniform number of samples, the validation process is standardized, which can lead to more consistent and comparable results, and selecting one station from each major KGCZ ensures a balanced representation of different climate zones, preventing any single zone from dominating the validation. This approach is not perfect either, as using only five stations may limit the scope of the validation and may not capture the full range of variability within each climate zone, and with a smaller dataset, the validation results may be more sensitive to outliers or anomalous data points. A judicious combination of both approaches, ensuring both breadth and uniformity, and using a standard set of stations, could be considered for more comparable model validation in the future, and rule out cherry picking.

Employing a more appropriate climate classification scheme than the Köppen-Geiger climate classification could improve the consistency in model validations, as also concluded in Starke et al. [26]. Utilizing a classification system with greater relevance could facilitate more meteorologically adaptive modeling.

7 CONCLUSIONS

In the last decade, the literature on decomposition modeling has largely lacked data- or physics driven basics for the methods used for developing new models. The work presented in this paper addresses this gap by offering a systematic approach to the model development process.

The development of single-predictor models provided insights into the relationships between individual predictors and the diffuse fraction. However, it was the multi-predictor model, Nested, that showcased the potential for more accurate estimations by leveraging the collective predictive power of multiple variables. Notably, the Nested model outperformed other recent universal models during validation.

While this study represents an advancement in modeling the diffuse fraction of solar radiation, it is important to recognize its limitations. The models are based on the limited data used, and their performance may vary with different datasets, especially from locations where they have not been validated.

Furthermore, it is crucial to acknowledge that the methods employed in this work are not beyond scrutiny. Critical evaluation and transparent discussion of the choices made in model development are essential for the advancement of scientific knowledge and the development of models that are both robust and interpretable. This, in turn, contributes to the credibility and utility of models in practical applications and decision-making.

In conclusion, this study has laid a foundation for the modeling of the diffuse fraction and has opened avenues for further research. It is encouraged that others aid in the search for better methods of predictor selection, model development and standardization of model validation. We hope that this work will serve as a reference, ultimately benefiting solar energy applications.

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.

APPENDIX, NESTED PYTHON CODE

```
def nested(kt,dktc,Kt,
    psi,kdm,kcsi):
    b0 = -13.49189475455595
    b1 = -6.489121933141542
    b2 = 1.4892743162295365
    b3 = 12.037329137008419
    b4 = 4.074156126269681
    b5 = 8.08428258256891
    b6 = 0.5394717724137723
```

10.4229/EUPVSEC2023/4CO.9.2
020362-009

```
    C = 0.45292421404358085

    EXP = b0+b1*kt+b2*dktc+\
            b3*Kt+b4*psi+b5*kcsi
    denum = 1 + np.exp(EXP)
    kd = C/denum + b6*kdm
    return np.minimum(kd,1)
```

References

[1] C.A. Gueymard. "Solar Radiation Resource: Measurement, Modeling, and Methods". In: *Reference Module in Earth Systems and Environmental Sciences*. Elsevier, 2021. ISBN: 978-0-12-409548-9. DOI: https://doi.org/10.1016/B978-0-12-819727-1.00101-1. URL: https://www.sciencedirect.com/science/article/pii/B9780128197271001011.

[2] Dazhi Yang. "Solar radiation on inclined surfaces: Corrections and benchmarks". In: *Solar Energy* 136 (2016), pp. 288–302. ISSN: 0038-092X. DOI: https://doi.org/10.1016/j.solener.2016.06.062. URL: https://www.sciencedirect.com/science/article/pii/S0038092X16302432.

[3] Marc A. Anoma et al. "View Factor Model and Validation for Bifacial PV and Diffuse Shade on Single-Axis Trackers". eng. In: *2017 Ieee 44th Photovoltaic Specialist Conference (pvsc)* (2017), pp. 1549–1554. ISSN: 01608371.

[4] Martin János Mayer and Gyula Gróf. "Extensive comparison of physical models for photovoltaic power forecasting". In: *Applied Energy* 283 (2021), p. 116239.

[5] Giorgio Belluardo et al. "Evaluation of uncertainty in PV project design: Definition of scenarios and impact on energy yield predictions". In: *2017 IEEE 44th Photovoltaic Specialist Conference (PVSC)*. IEEE. 2017, pp. 3360–3365.

[6] Clifford W Hansen and Curtis E Martin. "Photovoltaic system modeling: Uncertainty and sensitivity analyses". In: *Sandia Report (SAND2015-6700)* (2015), p. 82.

[7] Christian A. Gueymard and Jose A. Ruiz-Arias. "Extensive worldwide validation and climate sensitivity analysis of direct irradiance predictions from 1-min global irradiance". In: *Solar Energy* 128 (2016). Special issue: Progress in Solar Energy, pp. 1–30. ISSN: 0038-092X. DOI: https://doi.org/10.1016/j.solener.2015.10.010. URL: https://www.sciencedirect.com/science/article/pii/S0038092X15005435.

[8] Dazhi Yang. "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". In: *Renewable and Sustainable Energy Reviews* 159 (2022), p. 112195.

[9] Benjamin YH Liu and Richard C Jordan. "The interrelationship and characteristic distribution of direct, diffuse and total solar radiation". In: *Solar energy* 4.3 (1960), pp. 1–19.

[10] Arvid Skartveit, Jan Asle Olseth, and Marit Elisabet Tuft. "An hourly diffuse fraction model with correction for variability and surface albedo". In: *Solar Energy* 63.3 (1998), pp. 173–183. ISSN: 0038-092X. DOI: https://doi.org/10.1016/S0038-092X(98)00067-X. URL: https://www.sciencedirect.com/science/article/pii/S0038092X9800067X.

[11] John Boland, Lynne Scott, and Mark Luther. "Modelling the diffuse fraction of global solar radiation on a horizontal surface". In: *Environmetrics: The official journal of the International Environmetrics Society* 12.2 (2001), pp. 103–116.

[12] Barbara Ridley, John Boland, and Philippe Lauret. "Modelling of diffuse solar fraction with multiple predictors". In: *Renewable Energy* 35.2 (2010), pp. 478–483. ISSN: 0960-1481. DOI: https://doi.org/10.1016/j.renene.2009.07.018. URL: https://www.sciencedirect.com/science/article/pii/S0960148109003012.

[13] N.A. Engerer. "Minute resolution estimates of the diffuse fraction of global irradiance for southeastern Australia". In: *Solar Energy* 116 (2015), pp. 215–237. ISSN: 0038-092X. DOI: https://doi.org/10.1016/j.solener.2015.04.012. URL: https://www.sciencedirect.com/science/article/pii/S0038092X15001905.

[14] Allan R. Starke et al. "Resolution of the cloud enhancement problem for one-minute diffuse radiation prediction". In: *Renewable Energy* 125 (2018), pp. 472–484. ISSN: 0960-1481. DOI: https://doi.org/10.1016/j.renene.2018.02.107. URL: https://www.sciencedirect.com/science/article/pii/S0960148118302593.

[15] Dazhi Yang and John Boland. "Satellite-augmented diffuse solar radiation separation models". In: *Journal of Renewable and Sustainable Energy* 11.2 (2019), p. 023705. DOI: 10.1063/1.5087463. eprint: https://doi.org/10.1063/1.5087463. URL: https://doi.org/10.1063/1.5087463.

[16] Dazhi Yang. "Temporal-resolution cascade model for separation of 1-min beam and diffuse irradiance". In: *Journal of Renewable and Sustainable Energy* 13.5 (2021), p. 056101.

[17] Eugenia Paulescu and Marius Paulescu. "Minute-Scale Models for the Diffuse Fraction of Global Solar Radiation Balanced between Accuracy and Accessibility". In: *Applied Sciences* 13.11 (2023), p. 6558.

[18] Christian A Gueymard et al. *BSRN data set for IEA-PVPS Task-16 Activity 1.4 Quality Control*. data set. 2022. DOI: 10.1594/PANGAEA.939988. URL: https://doi.org/10.1594/PANGAEA.939988.

[19] Anne Forstinger et al. "Expert quality control of solar radiation ground data sets". In: *ISES Solar World Congress*. 2021.

[20] John A. Duffie. *Solar Engineering of Thermal Processes, Photovoltaics and Wind, Solar Engineering of Thermal Processes Photovoltaics and Wind*. eng. John Wiley & Sons, 2020.

[21] Fritz Kasten and Andrew T Young. “Revised optical air mass tables and approximation formula”. In: *Applied optics* 28.22 (1989), pp. 4735–4738.

[22] Tahir Mehmood et al. “A review of variable selection methods in partial least squares regression”. In: *Chemometrics and intelligent laboratory systems* 118 (2012), pp. 62–69.

[23] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[24] Jamie M. Bright and Nicholas A. Engerer. “Engerer2: Global re-parameterisation, update, and validation of an irradiance separation model at different temporal resolutions”. In: *Journal of Renewable and Sustainable Energy* 11.3 (2019), p. 033701. DOI: 10.1063/1.5097014. eprint: https://doi.org/10.1063/1.5097014. URL: https://doi.org/10.1063/1.5097014.

[25] Eugenia Paulescu and Robert Blaga. “A simple and reliable empirical model with two predictors for estimating 1-minute diffuse fraction”. In: *Solar Energy* 180 (2019), pp. 75–84. ISSN: 0038-092X. DOI: https://doi.org/10.1016/j.solener.2019.01.029. URL: https://www.sciencedirect.com/science/article/pii/S0038092X19300386.

[26] Allan R. Starke et al. “Assessing one-minute diffuse fraction models based on worldwide climate features”. In: *Renewable Energy* 177 (2021), pp. 700–714. ISSN: 0960-1481. DOI: https://doi.org/10.1016/j.renene.2021.05.108. URL: https://www.sciencedirect.com/science/article/pii/S0960148121007916.