Renewable Energy 189 (2022) 259–272

Contents lists available at ScienceDirect

Renewable Energy

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

Improved turbidity estimation from local meteorological data for solar resourcing and forecasting applications

Shanlin Chen [a], Mengying Li [a, b, *]

[a] Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong Special Administrative Region
[b] Research Institute for Smart Energy, The Hong Kong Polytechnic University, Hong Kong Special Administrative Region

ARTICLE INFO

Article history:
Received 16 October 2021
Received in revised form
8 January 2022
Accepted 24 February 2022
Available online 3 March 2022

Keywords:
Clear-sky irradiance
Turbidity estimation
Meteorological measurements
Machine learning methods

ABSTRACT

This work presents a new method to estimate atmospheric turbidity with improved accuracy in estimating clear-sky irradiance. The turbidity is estimated by machine learning algorithms using commonly measured meteorological data including ambient air temperature, relative humidity, wind speed and atmospheric pressure. The estimated turbidity is then served as the Linke Turbidity input to the Ineichen-Perez clear-sky model to estimate clear-sky global horizontal irradiance (GHI) and direct normal irradiance (DNI). When compared with the original Ineichen-Perez model which uses interpolated turbidity from the monthly climatological means, our turbidity estimation better captures its daily, seasonal, and annual variations. When using the improved turbidity estimation in the Ineichen-Perez model, the root mean square error (RMSE) of clear-sky GHI is reduced from 24.02 W m^{-2} to 9.94 W m^{-2}. The RMSE of clear-sky DNI is deceased from 76.40 W m^{-2} to 29.96 W m^{-2}. The presented method is also capable to estimate turbidity in partially cloudy days with improved accuracy, evidenced by that the corresponding estimated clear-sky irradiance has smaller deviation from measured irradiance in the cloudless time instants. In sum, the proposed method brings new insights about turbidity estimation in both clear and partially cloudy days, providing support to solar resourcing and forecasting.

1. Introduction

Solar radiation reaching the Earth surface is either absorbed or scattered by the atmosphere based on the types and concentrations of the participating constituents and their radiative optical properties [1]. For solar energy conversion systems such as photovoltaics (PV) and concentrating solar power (CSP), ground level irradiance assessment and forecasting are crucial for their design and operation [2–5]. The attenuation of ground level solar irradiance is mainly caused by clouds, aerosols, water vapor, carbon dioxide and ozone [6], where clouds are the major modulator followed by aerosols and water vapor. However, the high temporal and spatial variations of the three major modulators as well as sensing difficulties of their concentrations [7] have posed considerable challenges for solar resourcing and forecasting applications. Therefore, a variety of clear-sky models have been developed over the years to estimate time varying ground level global horizontal irradiance (GHI), direct normal irradiance (DNI) and diffuse horizontal irradiance (DHI) if there were no clouds in the sky. The clear-sky models have been used extensively to quantify the effects of local aerosols and water vapor, as well as to facilitate cloud identification and analysis for forecasting applications [1,8–10].

As summarized in Refs. [10,11], clear-sky models with different complexity and performance can be broadly classified into two main groups: physical models and empirical models. Physical models apply radiative transfer models (RTMs) to estimate the irradiance attenuation effect of atmospheric constituents, and the ground level solar irradiance can be obtained through integration of the attenuation caused by different atmospheric components [10]. Empirical models are based on simplified parameterizations of the attenuation processes [10], which estimate the clear-sky irradiance using some atmospheric parameters, such as the aerosol optical depth (AOD) and precipitable water in simplified Solis model [12], and the Linke turbidity (T_L) in Ineichen-Perez model [13].

Physical models perform detailed analysis of the atmospheric attenuation processes, which generally lead to higher accuracy [11,14]. However, they require many inputs about local atmospheric conditions, some of them are not widely available. For instance, the

* Corresponding author. Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong Special Administrative Region.
E-mail address: mengying.li@polyu.edu.hk (M. Li).

https://doi.org/10.1016/j.renene.2022.02.107
0960-1481/

REST2 model [15] has been verified as one of the most accurate clear-sky models [11,14], but the required information about the atmospheric constituents, such as AOD at 550 nm, column amount of ozone, nitrogen dioxide and precipitable water are difficult to obtain for most locations [14,16]. Yang [14] discussed the choice of clear-sky models in solar forecasting applications from the perspectives of accessibility, forecast performance and statistical properties. It is found that high-fidelity physical models like REST2 are not frequently used for solar forecasting due to its complexity, and no evidence suggests that physical models can lead to more accurate forecast results when compared with empirical models [14].

As a member of the empirical model family, the Ineichen-Perez model is extensively used in solar forecasting due to its simplicity [14]. The main input of the Ineichen-Perez model is the T_L factor, which is defined as the number of clean and dry atmospheres that produce the same attenuation equivalent to the real atmosphere [17]. The T_L factor quantifies the attenuation of aerosols and water vapor [18], which typically varies between 1 and 10 [10]. The T_L factor is available worldwide as monthly climatology value from the SoDa database [19]. In PVLIB [20], linear interpolations of the monthly values are applied to build daily T_L time series for each location when using the Ineichen-Perez clear-sky model.

The T_L factor is also directly used in other clear-sky models [10,21]. However, the invariant T_L factor based on the monthly climatology value and its linear interpolation cannot account for the short-term [22] and long-term variations [23] of atmospheric aerosols and water vapor concentrations, resulting in unsatisfying estimation of clear-sky irradiance [24]. The discrepancy of clear-sky irradiance obtained from T_L based clear-sky models and from measurements are observed in the studies by Moldovan et al. [21] and Polo et al. [22], and also noticeable when comparing the clear-sky solar irradiance measurements from Dessert Rock, Nevada (DRA) with PVLIB clear-sky model output (see Fig. 1).

Therefore, some studies were conducted to estimate T_L factor by different means with the aim to investigate its variations or to improve the estimation accuracy. Chaâbane et al. [7] adopted pyrheliometric measurements for the calculation of T_L factor in Tunisia during three summer months, where diurnal and monthly variations of T_L factor are observed. Polo et al. [22] estimated the daily T_L factor for clear days by using global irradiance measurements at solar noon and monthly mean T_L values. Using the estimated T_L to recalculate clear-sky solar irradiance results in a reduced root mean squared deviation (RMSD) when compared with using monthly mean values. The relative RMSD (rRMSD) decreases from 17.1% to 14.2% for the dataset of Baseline Surface Radiation Network (BSRN). For the dataset from Spanish Meteorological Agency (AEMet), the rRMSD reduces from 24.4% to 16.8%. Hove and Manyumbu [25] calculated the T_L factor based on daily GHIcs and ESRA clear-sky model [26], which typically has a lower value than the monthly mean. Inman et al. [27] reported a method for daily average T_L estimation using broadband DNI measurements under cloudless skies, and then applied the estimated T_L in DNI forecasting during cloud-free periods under the assumption of a persistence of daily averaged T_L within the forecasting horizon. The relative RMSE (rRMSE) and relative mean bias error (rMBE) are smaller than 5% for both historical and forecasted values, which are much smaller than the error range (10–20%) of SoDa monthly means. Behar et al. [28] used ambient temperature and relative humidity to estimate T_L and solar irradiance via the estimated optical thickness of clean-dry atmosphere, water vapor and aerosol. The T_L estimation has a rRMSE of 10.22% and a rMBE of 1.31%, the rRMSE and rMBE of corresponding DNI estimate are 5.21% and 0.91%, respectively. Moldovan et al. [21] applied time dependent interpolation polynomials instead of a constant daily T_L factor to improve the clear-sky model. Two different interpolation polynomials are obtained for the T_L factor in warm and cold seasons, respectively.

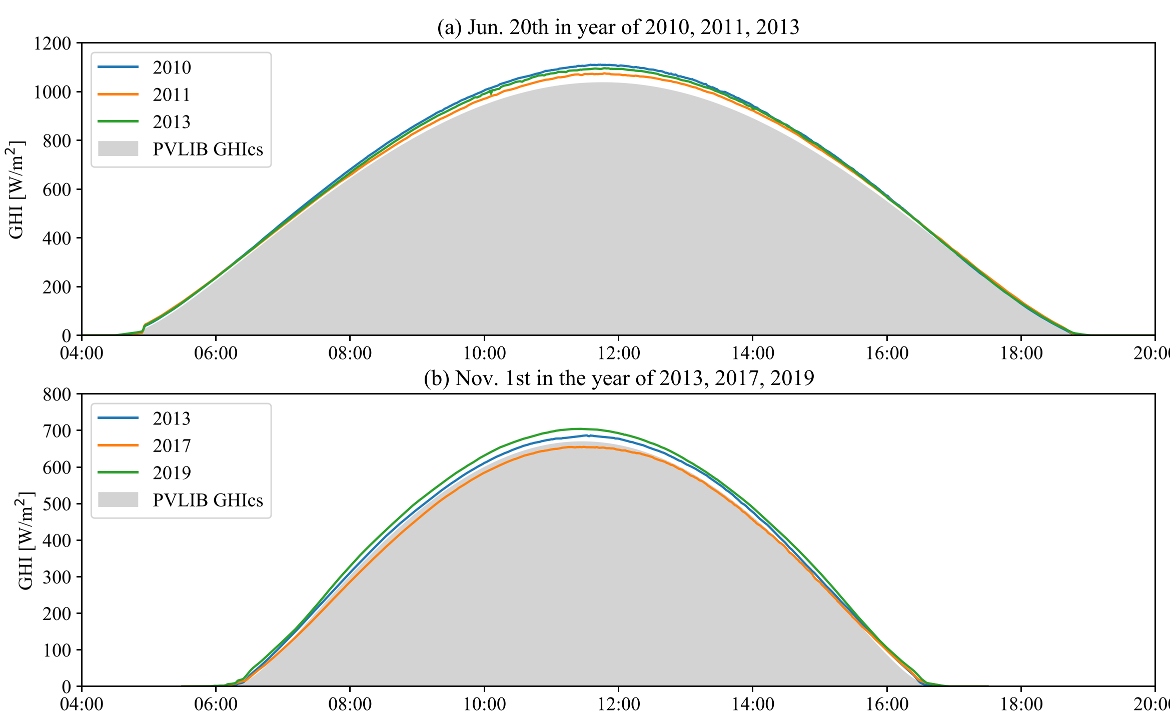

Fig. 1. Comparison of measured clear-sky GHI (GHIcs) with PVLIB GHIcs of the same day in different years. PVLIB uses the T_L factor from its look-up table, which is based on constant monthly climatology value. The PVLIB GHIcs remains the same on the same day of different years, while the measured GHIcs are not, indicating T_L factor also varies on a long term (i.e., yearly) basis.

The result shows that the relative error is reduced from 8.12% to 4% in the warm season and from 5.02% to 4.15% in the cold season.

The derivation of T_L based on irradiance and meteorological measurements offers a simpler way to estimate T_L without the detailed information about aerosol and water vapor contents. However, the methods summarized above still have some shortcomings to overcome. For example, the clear-sky irradiance measurements are not available in a cloudy day, and only using the GHIcs at the solar noon presented by Polo et al. [22] may lead to errors in estimating the clear-sky irradiance in other periods such as solar mornings, evenings and cloudy days. The method also cannot be used for locations without irradiance measurements. For the method presented by Behar et al. [28], using ambient temperature and relative humidity to estimate perceptible water and AOD may result in error accumulations in estimating T_L. For the study by Modlovan et al. [21], the T_L interpolation polynomials for warm and cold seasons are not capable of accounting for year-to-year T_L variation as shown in Fig. 1.

Therefore, we propose a new T_L estimation method to estimate high-fidelity T_L with the consideration of its short-term and long-term variations, and without the data dependence on local real-time irradiance measurements. The T_L factor is proposed to be estimated using local meteorological data by machine learning (ML) algorithms. In the following sections, data processing and proposed methodology are presented in Section 2. Section 3 presents the results and discussions of T_L and corresponding clear-sky irradiance estimations. The key findings and recommendations are summarized in Section 4.

2. Methodology of turbidity estimation

This section presents the data and methods used for T_L derivation and estimation. The T_L derivation is performed by applying Ineichen-Perez clear-sky model (PVLIB) reversely, i.e., taking the 1-min averaged GHIcs as the input to compute the 'ground truth' T_L. Then the derived minute-wise T_L time series is further averaged on the basis of daily, hourly and 5-min as the ML model training targets. The input meteorological data is also averaged with the same time basis for model training, tuning and testing. Finally, the trained model is applied to estimate the T_L for GHIcs estimation. The flowchart of the method for estimating the T_L and clear-sky irradiance is shown in Fig. 2. The T_L derivation and estimation can also be applied to clear-sky DNI (DNIcs), which will be discussed in Section 3.3.

2.1. Data selection

The data used in this work is from DRA, one of the Surface Radiation Budget Network (SURFRAD) stations [29]. DRA has a latitude of 36.62373°N, a longitude of 116.01947°W, an elevation of 1007 m, and a time zone of UTC-8 (8 h difference than coordinated universal time (UTC)). High resolution solar irradiance and meteorological data collected from year 2000–2020 are used in this work. Among the diverse variables in the comprehensive dataset, measurements of the downwelling global solar irradiance (GHI), direct normal irradiance (DNI), ambient air temperature (T_a), relative humidity (ϕ), wind speed (V) and local atmospheric pressure (P_a) are selected to build and test the proposed T_L estimation model. The selected data has high temporal resolutions (3-min averaged from year 2001–2008, and 1-min averaged from year 2009–2020) and its quality is carefully controlled.

DRA is chosen among the seven SURFRAD stations due to its high occurrence of cloudless days, which could provide adequate learning samples for the development and validation of the proposed T_L estimation model. The same methodology can be applied to other locations if sufficient data is given.

2.2. Selection of clear-sky days

The clear-sky irradiance is defined as the incident radiation at the Earth's surface under the conditions that would occur under a perfect "cloudless sky" [30]. The presence of clouds in the sky, especially when clouds obscure the Sun disc, will greatly affect the surface solar radiation, resulting in irradiance fluctuations. Since T_L is a factor that quantifies the attenuation of solar irradiance by atmospheric constituents (especially water vapor and aerosols) under cloud-free conditions, we only select clear-sky days for model development and validation.

The clear-sky days are selected following the approach developed by Long and his collaborators [31–33], and the clear-sky labels provided by RadFlux algorithm [31,33] are publicly available on the website of SURFRAD network. Specifically, the days will be labeled as "clear-sky day" if most of the time instants within the day are "clear" as detected by solar shortwave irradiance measurements (λ < 4 μm), or detected by atmospheric longwave irradiance

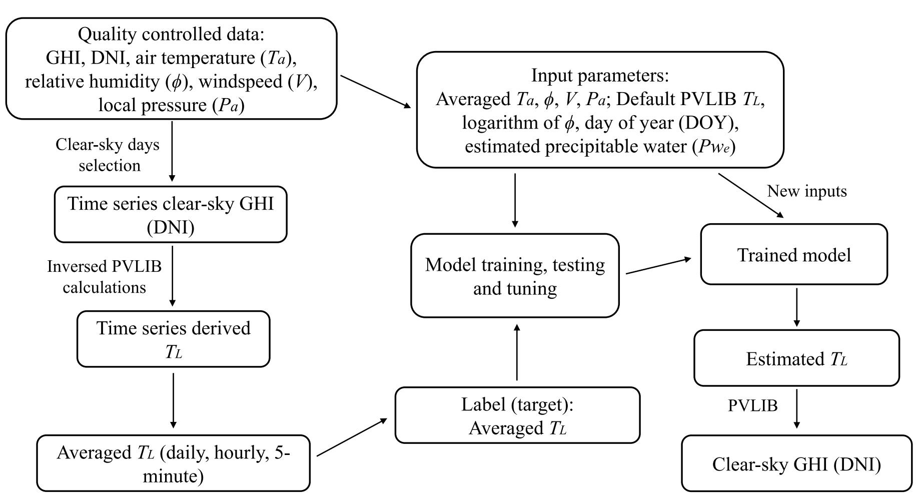

Fig. 2. An overview of the method to derive and estimate T_L. The model for estimating the daily, hourly and 5-min T_L is trained independently, and the input meteorological data is also averaged on the same time basis.

measurements ($\lambda > 4$ μm). The shortwave clear-sky detection algorithm has a 160° field of view, so the clear-sky instants with high solar zenith angles could not be detected [32]. During the daytime, the presence of clouds is more noticeable in the shortwave spectrum when compared with the longwave spectrum [33]. Therefore, most of the clear-sky instants detected by the longwave RadFlux algorithm [33] are in the nighttime. To verify the RadFlux clear-sky labels, we performed an additional manual check by comparing the measured GHI, clear-sky GHI estimated by RadFlux [34] and Ineichen–Perez clear-sky model from PVLIB [20]. Then, clear-sky detection for periods with high solar zenith angles are performed, and wrongly labeled clear-sky days are removed, as demonstrated in Fig. 3. Our training and testing datasets only contain the clear-sky days that pass both the RadFlux and manual checks (examples are presented in Fig. 3).

2.3. Derivation of 'ground truth' turbidity for model training

We adapt the methodology documented in PVLIB [20] to derive 'ground truth' T_L factor for model development. At each clear-sky time instance, the 'ground truth' T_L factor is derived from measured GHIcs values by inverting the following equation provided in PVLIB (proposed in Ref. [13]),

$$GHI_{cs} = c_1 \cdot I_0 \cdot cos(\theta) \cdot \exp(-c_2 \cdot AM \cdot (f_1 + f_2 \cdot (T_L - 1)))$$

Then the derived T_L based on GHIcs measurement is,

$$T_L = \left[\ln\left(\frac{GHI_{cs}}{c_1 \cdot I_0 \cdot cos(\theta)}\right) \Big/ (-c_2 \cdot AM) - f_1 \right] \Big/ f_2 + 1 \tag{1}$$

T_L could also be derived from DNIcs by inverting the following equations from PVLIB,

$$B_1 = I_0 \cdot b \cdot exp(-0.09 \cdot AM \cdot (T_L - 1))$$

$$B_2 = GHI_{cs} \cdot \left[\left(1 - \frac{(0.1 - 0.2 \cdot exp(-T_L))}{(0.1 + 0.882/f_1)}\right) \Big/ \cos(\theta) \right]$$

$$DNI_{cs} = \mathrm{Minimum}(B_1, B_2)$$

Then the derived T_L based on DNIcs measurements is,

$$T_L = \ln\left(\frac{DNI_{cs}}{I_0 \cdot b}\right) \Big/ (-0.09 \cdot AM) + 1, (\text{when } B_1 < B_2) \tag{2}$$

$$T_L = -\ln\left[\left(0.1 - \left(1 - \frac{DNI_{cs}}{GHI_{cs}} \cdot cos(\theta)\right) \cdot (0.1 + 0.882 / f_1)\right) \Big/ 0.2 \right], (\text{when } B_1 > B_2) \tag{3}$$

with:

$$AM = \left(\frac{1}{\cos(\theta) + 0.50572 \cdot \left(6.07995 + (90 - \theta)^{-1.6364}\right)}\right) \cdot \frac{P_a}{101325}$$

$$c_1 = 5.09 \cdot 10^{-5} \cdot h + 0.868$$

$$c_2 = 3.92 \cdot 10^{-5} \cdot h + 0.0387$$

$$f_1 = \exp(-h / 8000)$$

$$f_2 = \exp(-h / 1250)$$

$$b = 0.664 + 0.163/f_1$$

where GHI_{cs} [W m^{-2}] is the measured clear-sky GHI. DNI_{cs} [W m^{-2}] is the measured clear-sky DNI. B [W m^{-2}] is the normal beam clear-sky radiation. c_1, c_2, f_1, f_2, b are altitude-dependant coefficients, I_0 [W m^{-2}] is the solar constant, θ[°] represents the solar zenith angle, AM is the absolute airmass, T_L is the Linke Turbidity factor, P_a [Pa] is the local atmospheric pressure, and h [m] is local altitude.

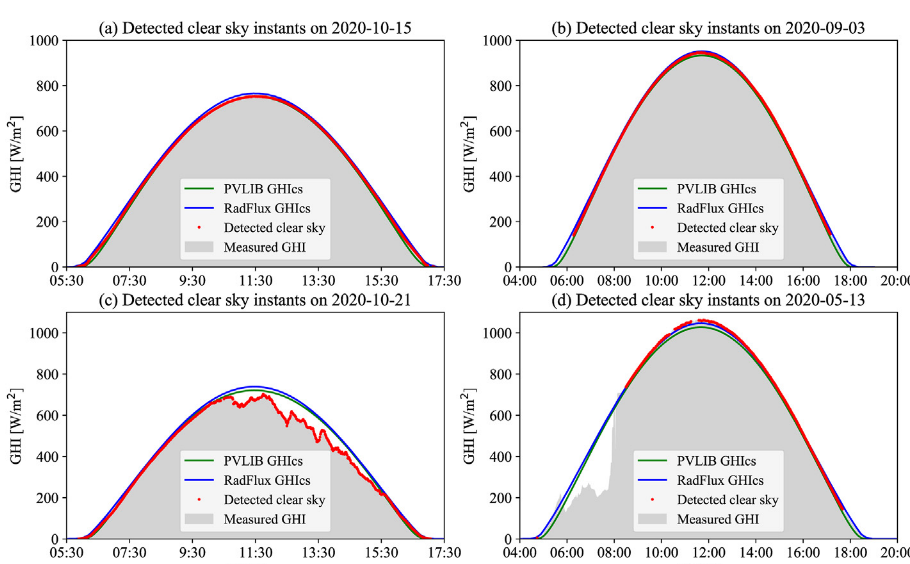

Fig. 3. Examples of clear-sky days selection in DRA. Measured GHI data is from SURFRAD, GHIcs are computed by both RadFlux and PVLIB. The clear-sky labels are from RadFlux. (a) A detected full clear-sky day. (b) A detected clear-sky day with high solar zenith periods not labeled. (c) A wrongly labeled clear-sky day which is removed by manual check. (d) A typical partly cloudy day.

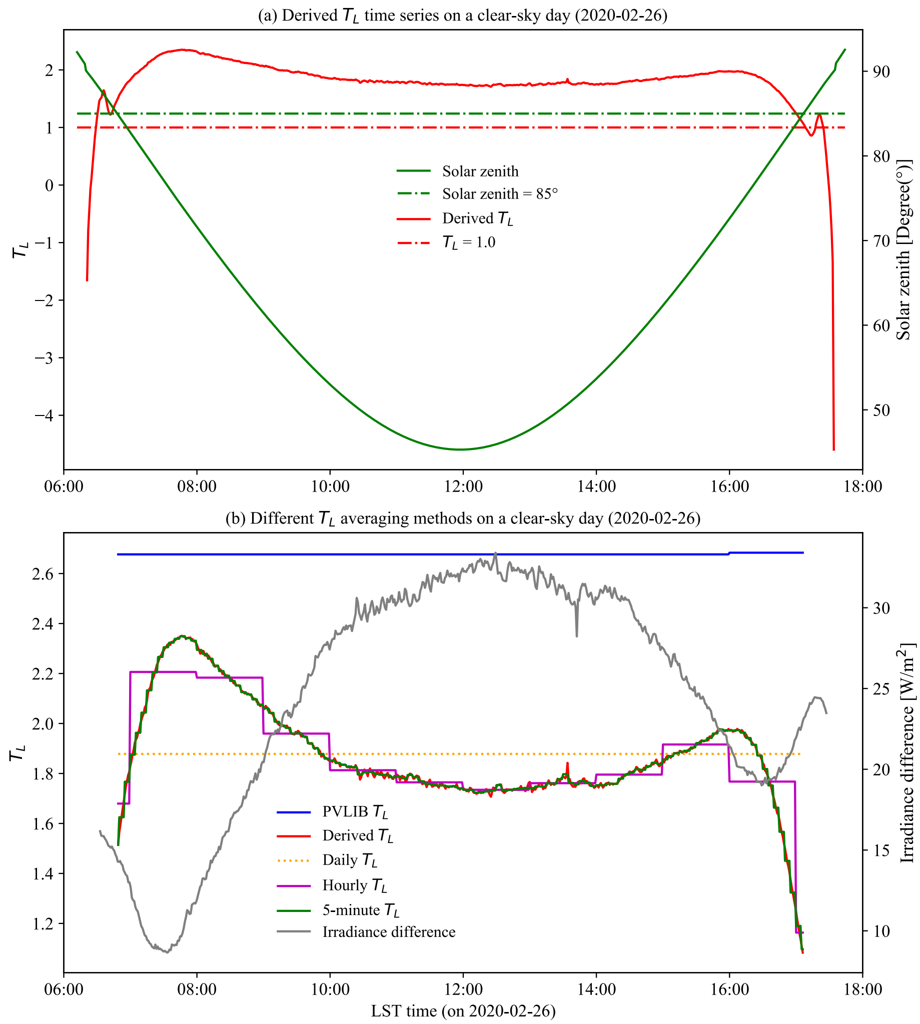

Fig. 4. Derived T_L time series with different averaging modes for a clear-sky day with respect to local standard time (LST). (a) Derived T_L time series on 2020-02-26. The derived T_L shows high variations and unrealistic values in the periods with high solar zenith angles. (b) Averaged T_L on different time basis and the irradiance difference between measured GHIcs and PVLIB GHIcs during the day.

Fig. 4 (a) illustrates the GHIcs-derived T_L in a randomly selected clear-sky day (one can find similar results in any other clear-sky days). Unlike the T_L factor used in PVLIB default calculations, the derived T_L factor is not a constant but varies during the day. Note that for periods with large solar zenith angle (greater than 85°), the derived T_L has large variations and unrealistic values (less than 1.0 and even negative), which are resulted from the applicable limitations of Eq. (1). Therefore, in the following sections when derived T_L time series are temporally averaged, the instances from the period when the solar zenith is greater than 85° are not included. Fig. 4 (b) shows the T_L time series averaged on different time basis, where the daily averaged T_L is much lower than the value used in PVLIB. The asymmetry of estimated T_L with respect to the zenith angle is observed in Fig. 4 (b), especially during morning and evening periods. This is possibly due to the high airmass effect, where the small difference in measured clear-sky irradiance will result in large discrepancy in the derived T_L values. In addition, the profile of clear-sky irradiance is not perfectly symmetric as shown in Fig. 4 (b) that the measured GHIcs in the morning (e.g., 6:00–8:00) is smaller than the ones near the evening (e.g., 16:00–18:00). Meanwhile, the water vapor content in the atmosphere is usually higher in the morning, which results in higher T_L values and thus lower GHIcs.

The averaged T_L derivations are then used to recalculate the 1-min averaged GHIcs using PVLIB, which shows noticeable improvement in estimating GHIcs, as shown in Fig. 5. In the clear-sky days of 2019 and 2020 (a total of 84 days are identified as clear), using PVLIB T_L generally underestimates the GHIcs with a mean bias error (MBE) of −20.48 W m^{-2} and a root mean square error (RMSE) of 24.02 W m^{-2}, when computed using 1-min averaged data when solar zenith angle is smaller than 85°. Using derived daily mean T_L yields a GHIcs estimation with overall MBE of 0.34 W m^{-2} and RMSE of 6.74 W m^{-2}, a 98.3% reduction in MBE and 71.9% decrease in RMSE. As the time resolution increases, the recalculated results become better as expected. Hourly mean T_L produces a RMSE of 2.81 W m^{-2} and a MBE of 0.05 W m^{-2} for GHIcs estimation. The 5-min averaged T_L gives an estimation of GHIcs with the lowest MBE of 0.01 W m^{-2} and lowest RMSE of 0.55 W m^{-2}. In general, using daily mean T_L can successfully correct the bias in estimating GHIcs and reduce RMSE by 71.9%. Using temporally finer hourly and 5-min averaged T_L can further reduce the RMSE in GHIcs estimations, but they also substantially increased the size of training data in the following ML based T_L estimation models.

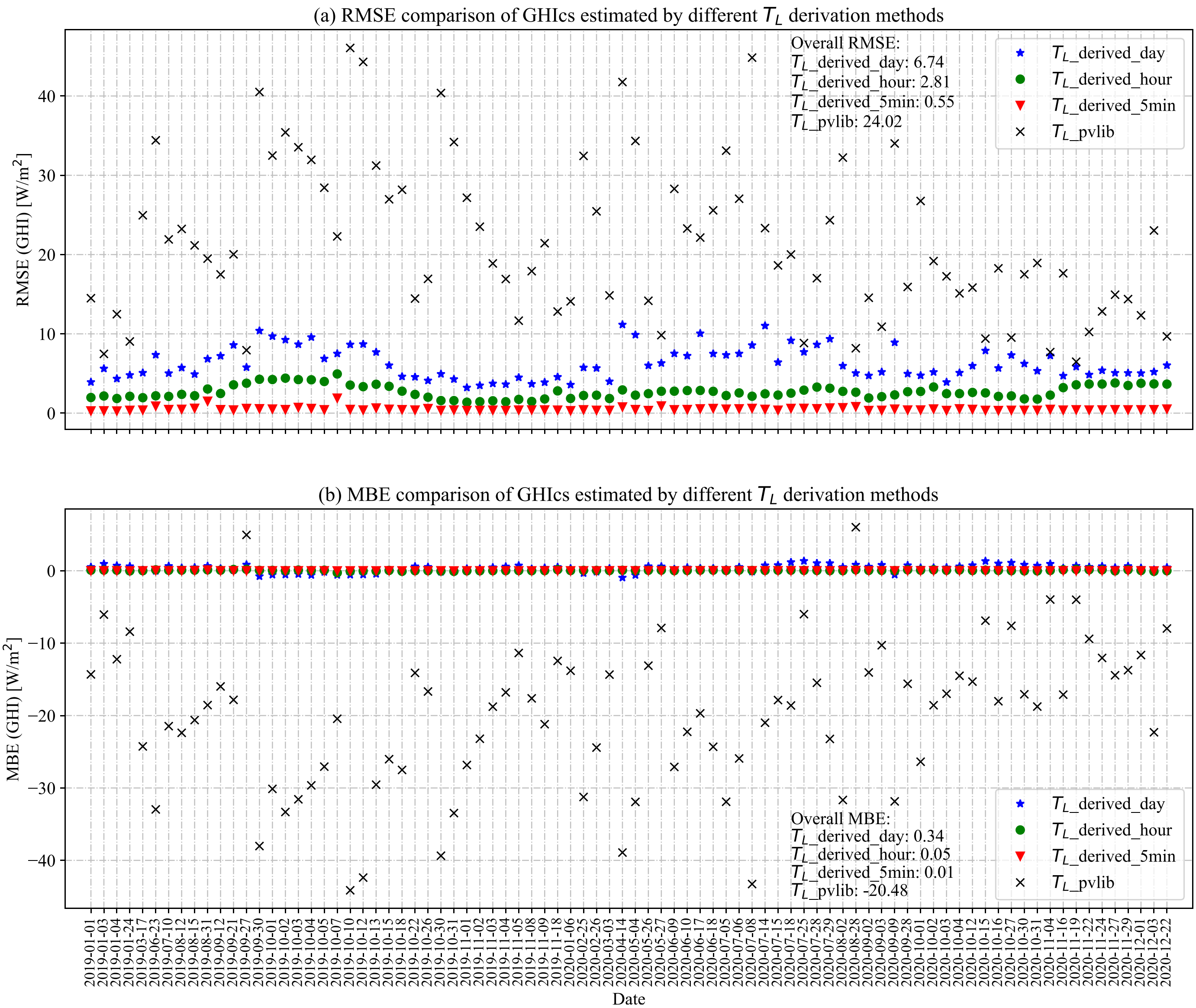

Fig. 5. Comparison of GHIcs estimation using different T_L averaging modes for the clear-sky days in year 2019 and 2020. (a) Daily RMSE of 1-min averaged GHIcs estimation. (b) Daily MBE of 1-min averaged GHIcs estimation. All the derived T_L regardless of averaging mode produce more accurate GHIcs than PVLIB.

2.4. Turbidity estimation from local meteorological data

The previous section demonstrates that improving T_L estimations could substantially improve the accuracy of GHIcs estimations. However, the T_L is derived from GHIcs measurements, which is not known as a priori in real-time applications. Therefore, we propose to use ML methods with widely available meteorological measurements to estimate local T_L.

We use three independent ML models for daily, hourly and 5-min averaged T_L factors. The label (target) is the averaged T_L derivations from Section 2.3, and the input parameters are: the default PVLIB T_L, ambient air temperature T_a, relative humidity ϕ (and its logarithm), wind speed V, atmospheric pressure P_a, day of year (DOY), and estimated precipitable water Pw_e. The meteorological time series (i.e., air temperature, relative humidity, wind speed, pressure) are averaged on the same time basis as the T_L. The PVLIB T_L is adapted to the corresponding time resolution as well. The logarithm of relative humidity is based on its averaged value, and the estimated precipitable water is calculated from the averaged temperature and the averaged relative humidity using the empirical model proposed by Gueymard [35,36] with the following equations.

$$Pw_e = 0.1 \cdot H_v \cdot \rho_v$$

$$H_v = 0.4976 + \frac{1.5265 \cdot T_a}{273.15} + \exp\left(\frac{13.6897 \cdot T_a}{273.15} - 14.9188 \cdot \left(\frac{T_a}{273.15}\right)^3\right)$$

$$\rho_v = 216.7 \cdot \phi \cdot e_s / T_a$$

$$e_s = \exp\left(22.330 - 49.140\cdot\frac{100}{T_a} - 10.922\cdot\left(\frac{100}{T_a}\right)^2 - 0.390\,15\frac{T_a}{100}\right)$$

where Pw_e [cm] is the estimated precipitable water. H_v [km] is the apparent water vapor scale height. ρ_v [g m^{-3}] is the surface water vapor density. ϕ [%] is the relative humidity. e_s [millibar] is the saturation water vapor pressure. T_a [°C] is the ambient air temperature.

ML technique is a powerful tool in regression modelling, which can model the relations between input features and target especially when the representation is complicated. ML algorithms have been widely used in classification, prediction and pattern recognition applications [37]. Here, we apply and compare Linear Regression (LR), Random Forest (RF), and Multilayer Perceptron (MLP) for T_L estimation, which are three commonly available and extensively used methods in real applications.

LR involves a linear combination of the input variables, which may have significant limitations for pattern recognition, particularly for problems with high dimensionality [37]. Therefore, linear model is extended by considering linear combinations of fixed nonlinear functions (basis function) of the input variables. Polynomial (powers of input variables) regression is one example of the extended linear models [37]. Although linear models are considered relatively simple and might not be suitable for high-dimensional problems, they have good analytical properties and form the fundamental for more advanced models [37]. Here we apply LR as a reference method in estimating T_L.

RF regressor is an ensemble method that combines several randomized regression decision trees to achieve a better performance [38]. RF is a bagging technique, all the involved decision trees are built in parallel and depend on the random vectors sampled from the training dataset. The predictions are averaged using bootstrap aggregation, which is one of the most computational-efficient methods to improve stability of the estimates [38]. RF models have been demonstrated to be robust predictors for both small sample sizes and data with high dimensionality [38].

MLP is also known as feed-forward neural network, which consists of an input layer, one or more hidden layers and one output layer [37]. MLP networks have high flexibility in approximation and can easily extend the structure by adding more hidden layers. MLP networks are trained and the parameters are obtained by back propagation [37]. There are different nonlinear activation functions of hidden layer(s), which could differ for different applications.

Data from 2000 to 2018 is used as the training set (20% of which is for validation) and data from 2019 to 2020 is used for testing. The model hyperparameters are tuned by using tenfold cross-validation method. The error evaluation metrics are MBE, RMSE and their normalized counterparts. All the above-mentioned ML models are adapted from Scikit-learn [39] and PyCaret [40], where more details regarding the applied algorithms can be found.

Table 1
Comparison of 1-minute averaged GHIcs recalculations and estimations using derived and estimated T_L for clear-sky days in the year of 2019 and 2020. PVLIB results are presented here for reference.

T_L	GHIcs recalculations[a]		GHIcs estimations[b]	
	RMSE [W m^{-2}]	MBE [W m^{-2}]	RMSE [W m^{-2}]	MBE [W m^{-2}]
Daily mean	6.74	0.34	9.94	2.09
Hourly mean	2.81	0.05	9.62	1.45
5-min mean	0.55	0.01	10.28	−0.01
PVLIB[c]	24.02	−20.48	24.02	−20.48

[a] GHIcs recalculations are based on the averaged T_L factors derived from GHIcs.
[b] GHIcs estimations are based on the estimated T_L factors from the ML (MLP) models with meteorological parameters as input.
[c] PVLIB uses the daily interpolated T_L based on the monthly climatological T_L map.

3. Results and discussion

The best ML model is selected separately for daily, hourly and 5-min averaged T_L estimation, and the overall corresponding 1-min averaged GHIcs estimation results for clear-sky days in year 2019 and 2020 are presented in Table 1. Compared with the GHIcs recalculation, GHIcs based on the estimated T_L yields slightly larger MBE and RMSE. Although 5-min averaged T_L has the best performance for GHIcs recalculations, but the fine temporal resolution does not show much superior results in GHIcs estimation. Estimating hourly averaged T_L results to better GHIcs estimation with an MBE of 1.45 W m^{-2} and a RMSE of 9.62 W m^{-2}. Using daily averaged T_L achieves a comparable result with slightly larger MBE of 2.09 W m^{-2} and RMSE of 9.94 W m^{-2}. Given that less complexity and computational resource are required for using daily averaged T_L, the subsequent results and discussion are based on the daily averaged T_L and the associated model.

3.1. Estimations of daily turbidity and 1-minute averaged GHIcs in clear-sky days

When compared with the monthly climatology mean of T_L, the derived daily T_L generally has a lower value and has a much higher fluctuation (see Fig. 6 (a)). The T_L values of 2020 is different from the year of 2019, which indicates the T_L also has a yearly variation. This long-term fluctuation of T_L is possibly caused by pollution [23] and the dynamics of aerosols and water vapor in the atmosphere [22]. Among the applied ML algorithms, MLP regressor gives the best results as shown in Table 2 and Fig. 6, with comparatively lower testing RMSE and MBE values.The normalized RMSE (nRMSE) of T_L estimation from all the ML models are around 10%. The learning curve of MLP regressor is shown in Fig. 7.

Fig. 8 (a) presents the sensitivity analysis of the meteorological inputs for the MLP model. The estimated T_L increases when the temperature and relative humidity become higher, while the increases in wind speed and pressure lead to a drop in the T_L estimation. Wind speed is the least sensitive parameter, so its impact on the T_L estimation is limited. Relative humidity and temperature have comparatively larger influence than wind speed, and temperature is a more crucial input for T_L estimation compared with relative humidity. Regards to local pressure, it does not have large variance as shown in Fig. 8 (b), so either increase or decrease pressure by 10% would lead it to be out of its min-max range. Since the MLP model is trained based on data samples with a small range of pressure variation, the out-of-range pressure will produce unrealistic T_L estimation. This is why the pressure shows the relatively larger sensitivity. In practical applications, the pressure of a certain place has limited variation, so its influence on T_L estimation also remains limited.

When using the estimated daily T_L from the MLP model to estimate 1-min averaged GHIcs, most of the tested clear-sky days in the year of 2019 and 2020 show noticeable improvements when compared with the PVLIB GHIcs in terms of RMSE and MBE (see Fig. 9). The overall RMSE of GHIcs estimation using the MLP-estimated T_L in 2019 and 2020 is 9.94 W m^{-2}, which is slightly higher than the RMSE (6.74 W m^{-2}) of GHIcs recalculation from the derived T_L, but much lower than the RMSE of 24.02 W m^{-2} from PVLIB. Note that there are some cases of model underperformance,

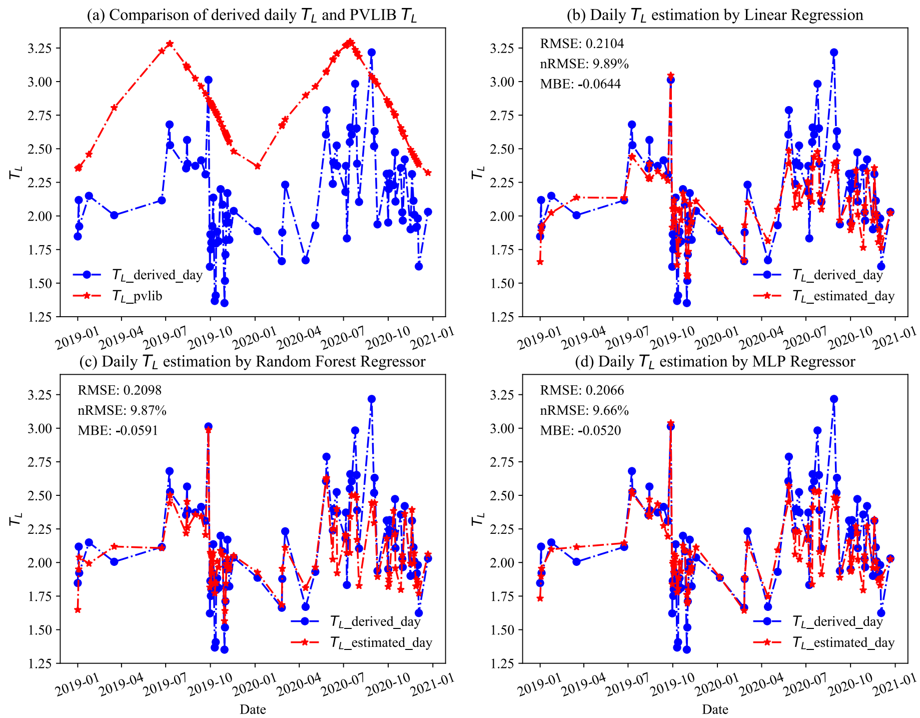

Fig. 6. Comparison of the derived daily T_L and PVLIB T_L and the performance of applied ML methods. (a) The comparison of derived T_L and PVLIB T_L. The comparison between derived T_L and estimated T_L from different methods (b) Linear Regression (c) Random Forest Regressor and (d) MLP Regressor.

Table 2
Training and testing errors of the applied ML algorithms for T_L estimation.

ML algorithm	Training		Testing	
	RMSE	MBE	RMSE	MBE
LR	0.2452	0.0000	0.2104	−0.0644
RF	0.2127	−0.0004	0.2098	−0.0591
MLP	0.2339	0.0044	0.2066	−0.0520

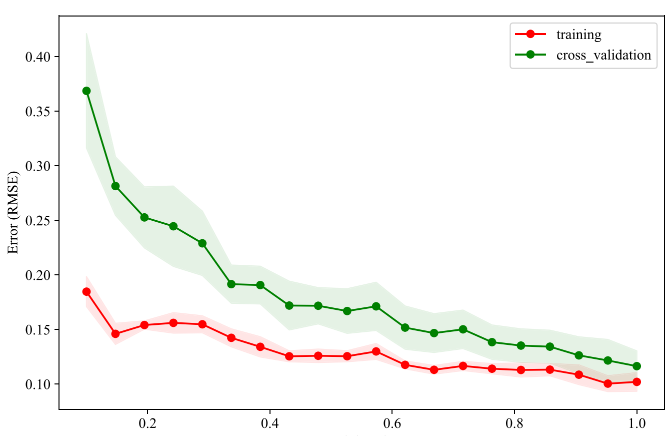

Fig. 7. Learning curve of MLP regressor with tenfold cross validation. The dots represent mean values, and the related shadows reflect the standard derivation.

which are likely due to that PVLIB T_L is already close to the derived T_L. Nevertheless, the GHIcs estimation using the estimated T_L factor has an overall better performance compared with PVLIB, which uses unmodified T_L based on the monthly climatology values, especially when the PVLIB GHIcs and measured GHIcs have large discrepancy.

3.2. Estimations of daily turbidity and 1-minute averaged GHIcs in partially clear days

Furthermore, we test our T_L estimation model in partially cloudy days when not all periods are cloudless throughout the day. As shown in Fig. 10, the model is also applicable to estimate T_L in this case and the corresponding 1-min averaged GHIcs estimation shows better agreement when compared with PVLIB for the clear-sky instants during the day. The potential explanation to this phenomenon is that the presence of clouds in partially clear days has limited effect on local meteorological parameters as well as ground level aerosols and water vapor concentrations. Accordingly, using local meteorological measurements (e.g., temperature, relative humidity) to estimate GHI with the presence of clouds may not be effective. Note that the phenomenon could be different in fully overcast days as the meteorological parameters might be affected, which needs further investigation. Nevertheless, the trained model works for the partially cloudy days, which would provide more accurate clear-sky irradiance during those periods for solar resourcing and forecasting applications. In addition, since the ML model can estimate T_L in both clear-sky and partially cloudy days, the derived T_L from the clear-sky instants in the partially cloudy days as well as corresponding meteorological variables can be included in the dataset for model development and testing. Which in turn can provide more data for ML model training and could potentially improve the model accuracy.

3.3. Estimations of daily turbidity and 1-minute averaged DNIcs

The same method is applied to estimate 1-min averaged DNIcs using the improved T_L estimations. Since PVLIB uses the same T_L value for calculating GHIcs and DNIcs, we use the GHIcs-estimated T_L to estimate DNIcs, as shown in Fig. 11. Both the recalculations and estimations have better overall performance than PVLIB, the RMSE is reduced from 76.40 W m^{-2} to 47.16 W m^{-2} and 50.77 W m^{-2},

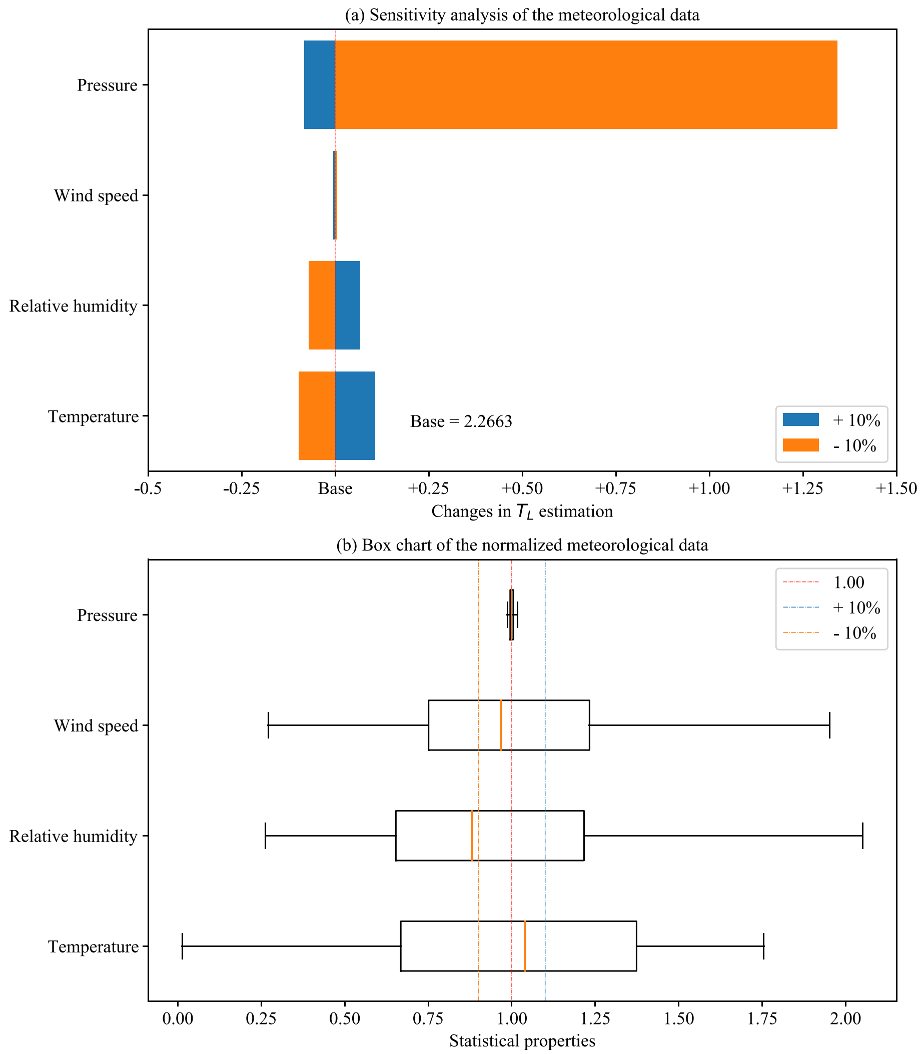

Fig. 8. Sensitivity analysis and statistical properties of the meteorological inputs for the MLP model. (a) Sensitivity analysis based on the changes of a sole parameter, where the base is the mean value. (b) Box chart of the normalized meteorological measurements.

respectively. The MBE is decreased from −62.45 W m^{-2} to 32.34 W m^{-2} for recalculations, and to 39.93 W m^{-2} for estimations. However, the error reduction is not as effective as GHIcs estimation, as it is noticed that the derived T_L from GHIcs could not always lead to better DNIcs estimations than PVLIB. Consequently, the estimated T_L could potentially lead to large errors by accumulating uncertainties in T_L estimation, as shown in Fig. 11.

To further improve the accuracy of DNIcs estimations, we derive T_L from DNIcs and develop separate ML models for T_L estimation following the similar strategy as described in Section 2. T_L is derived using Eqs. (2) and (3) from measured clear-sky DNI. A comparison among different T_L modelling methods for DNIcs estimation is shown in Table 3. All the improved T_L factors for DNIcs recalculations and estimations have superior results than default PVLIB. The 5-min averaged T_L gives the lowest RMSE of 5.74 W m^{-2} and a MBE of −1.36 W m^{-2} for recalculating DNIcs, while daily mean T_L generates a RMSE of 18.93 W m^{-2} and a MBE of 1.75 W m^{-2}. However, the developed MLP models for daily, hourly and 5-min T_L estimation yield comparable results for estimating DNIcs, which means averaging T_L on smaller time basis has limited potential to improve the DNIcs estimation accuracy. Since using daily mean T_L can generate comparable DNIcs estimations with less complexity, a detailed comparison of DNIcs estimation by suing estimated daily T_L and default PVLIB T_L is shown in Fig. 12.

From the perspective of atmospheric radiative transfer, DNI is comparatively more sensitive than GHI to the variations of atmospheric constituents and cloud dynamics, as GHI is the sum of DHI and the horizontal projection of DNI (GHI = DNI $\cdot$ $cos(\theta)$ + DHI, where $\theta[°]$ is the solar zenith angle). The rapidly changing DNIcs in the solar morning and evening also makes the DNIcs estimation more challenging than GHIcs. As demonstrated by our results, the default PVLIB T_L yields a RMSE of 76.40 W m^{-2} and a MBE of −62.45 W m^{-2} for DNIcs estimation, which is about three times of the RMSE (24.02 W m^{-2}) and MBE (-20.48 W m^{-2}) for estimating GHIcs.

Compared with GHIcs estimation from derived and estimated T_L factors, DNIcs estimation generally has comparatively larger errors of RMSE and MBE (see Tables 1 and 3). Using the 5-min averaged T_L factor almost produce a "perfect" GHIcs recalculation with the RMSE of 0.55 W m^{-2} and MBE of 0.01 W m^{-2}, while the RMSE is 5.74 W m^{-2} and MBE is −1.36 W m^{-2} for DNIcs recalculation. When it comes to estimation, the ML model (MLP is chosen) estimated daily T_L for DNIcs estimation has a RMSE of 29.96 W m^{-2}, which is nearly three times of the RMSE (9.94 W m^{-2}) of estimating GHIcs.

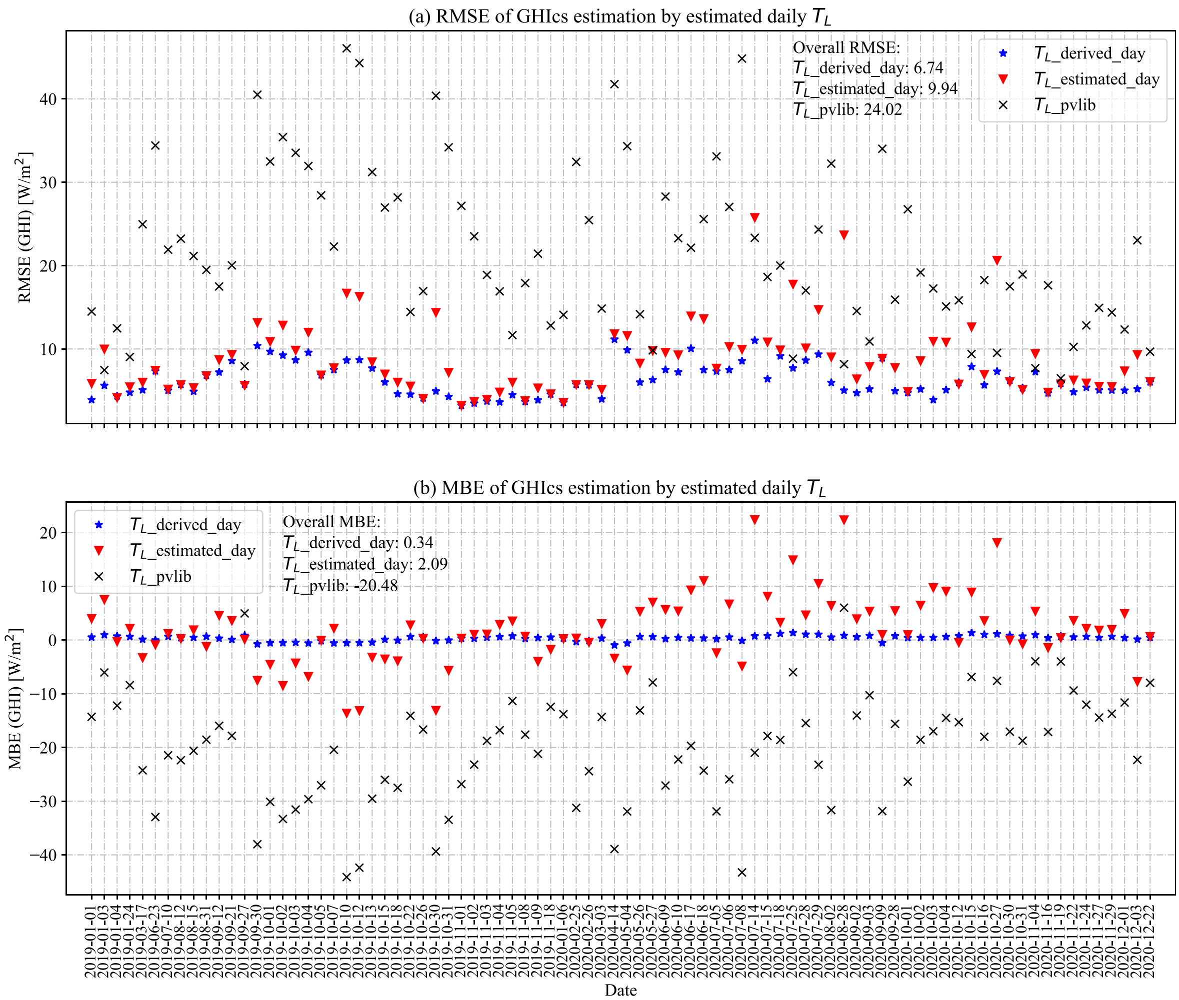

Fig. 9. The comparison of GHIcs estimation based on derived and estimated daily T_L factors. (a) Daily RMSE of GHIcs estimation. (b) Daily MBE of GHIcs estimation. Generally, the estimated T_L performs better than the default PVLIB T_L factor.

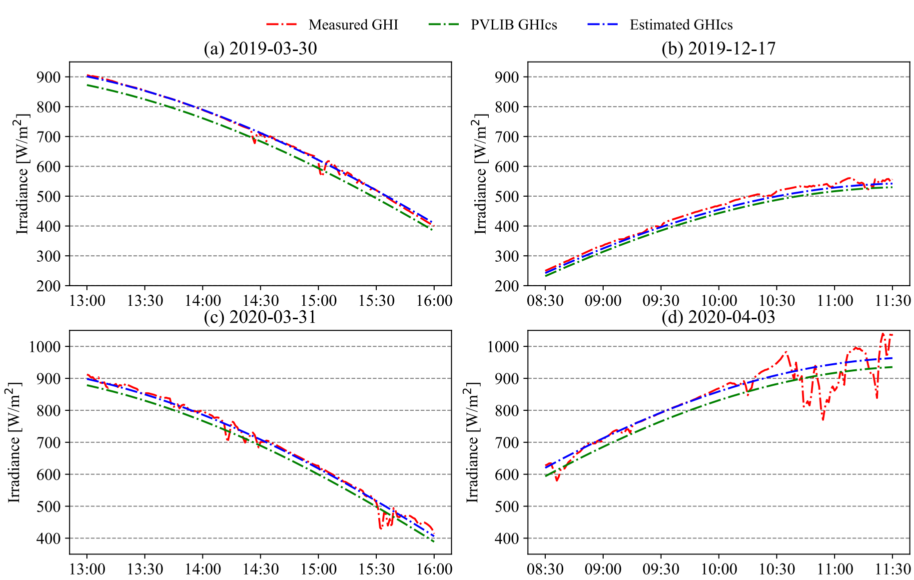

Fig. 10. Examples of GHI in partially cloudy days during (a) 2019-03-30 (b) 2029-12-27 (c) 2020-03-31 (d) 2020-04-03. The GHIcs calculated from the estimated T_L shows a higher accuracy than PVLIB when compared with measured GHI in the clear-sky instants.

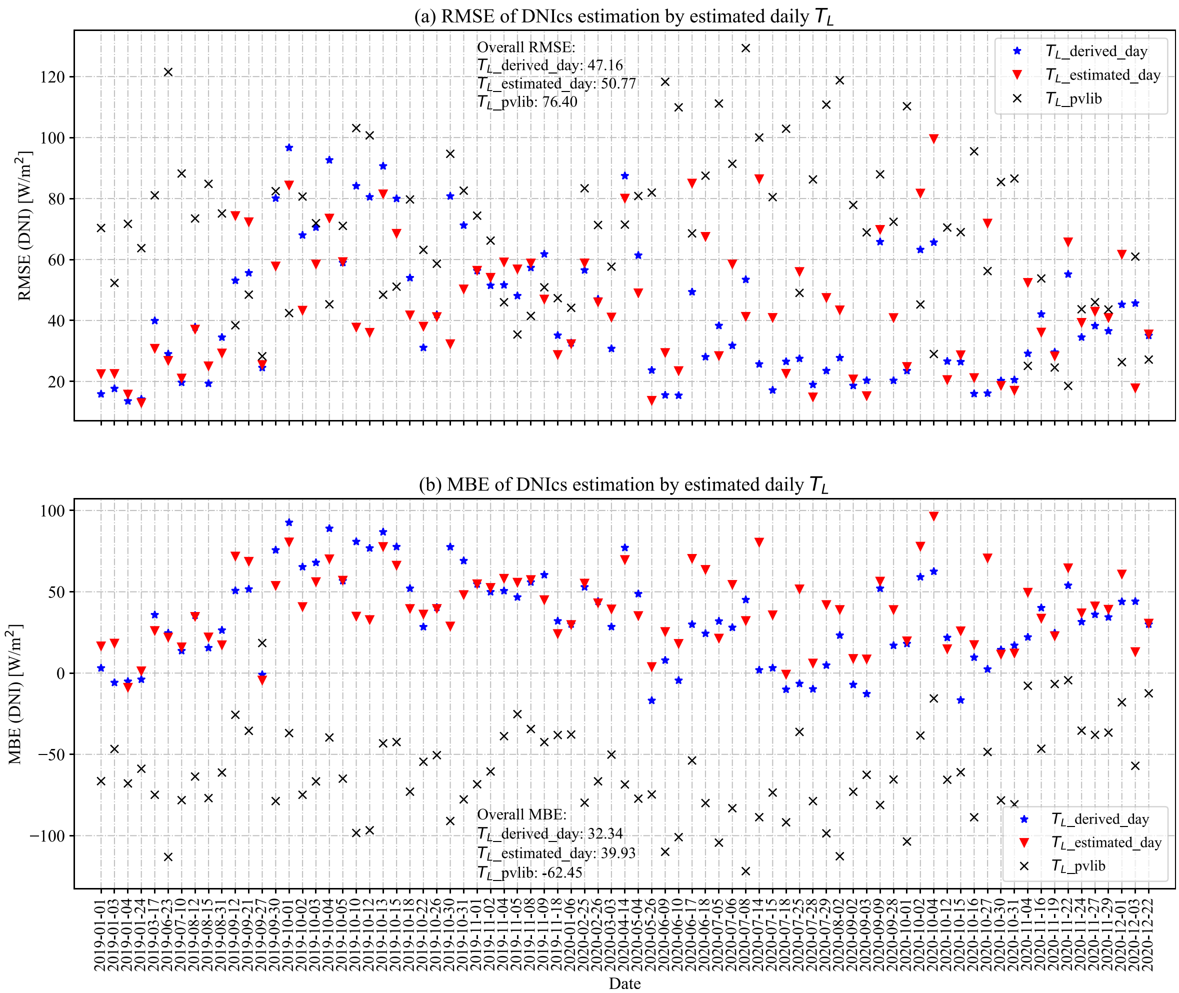

Fig. 11. The RMSE and MBE of DNIcs estimation using the GHIcs-based derived and estimated T_L. Both recalculations and estimations have lower overall RMSE and MBE than PVLIB but with some exceptions.

Table 3
Comparison of DNIcs recalculations and estimations using derived and estimated T_L for clear-sky days in the year of 2019 and 2020. Although 5-minute averaged T_L has the lowest RMSE for DNIcs recalculations, the 1-min averaged DNIcs estimations based on daily, hourly and 5-min averaged T_L show little difference.

T_L	DNIcs recalculations[a]		DNIcs estimations[b]	
	RMSE [W m^{-2}]	MBE [W m^{-2}]	RMSE [W m^{-2}]	MBE [W m^{-2}]
Daily mean	18.93	−1.75	29.96	2.68
Hourly mean	8.96	−1.46	30.75	−0.04
5-min mean	5.74	−1.36	31.96	−1.24
PVLIB[c]	76.40	−62.45	76.40	−62.45

[a] DNIcs recalculations are based on the averaged T_L factors derived from DNIcs.
[b] DNIcs estimations are based on the estimated T_L factors from the ML (MLP) model developed from the derived T_L.
[c] PVLIB uses the daily interpolated T_L based on the monthly climatological T_L map.

For partially cloudy days, the proposed DNIcs estimating method also outperforms PVLIB, but the degree of error reduction is smaller than those of GHIcs estimation, as demonstrated by Fig. 13. In sum, DNIcs estimation is more challenging than GHIcs and often has larger discrepancies, the applications that rely heavily on accurate DNIcs estimation is recommended to adopt the methods to improved DNIcs estimation (such as the one presented in this work).

3.4. Generic applicability of the proposed method

Here we apply the proposed methodology at other SURFRAD stations with limited occurrences of clear-sky days to demonstrate the generic applicability of the proposed method. The results of 1-min averaged GHIcs estimation for all the SURFRAD stations using the T_L estimation models developed using both clear-sky and partially clear days are presented in Table 4. Compared with the default PVLIB calculations, the proposed method generally produces better GHIcs estimations for all SURFRAD stations.

4. Conclusions

In this work, we present a new method to estimate turbidity factor T_L using common meteorological data by ML algorithms. The model inputs are: the default PVLIB T_L, ambient air temperature, relative humidity (and its logarithm), wind speed, atmospheric pressure, day of year (DOY), and estimated precipitable water. The model output is estimated T_L, which has the same temporal resolution as the input parameters. The training target of the ML algorithms is the T_L derived from measured clear-sky GHI or DNI. When tested using data from Desert Rock, Nevada, the new method successfully captures both the short-term and the long-term temporal variations of T_L by inferring from the local meteorological

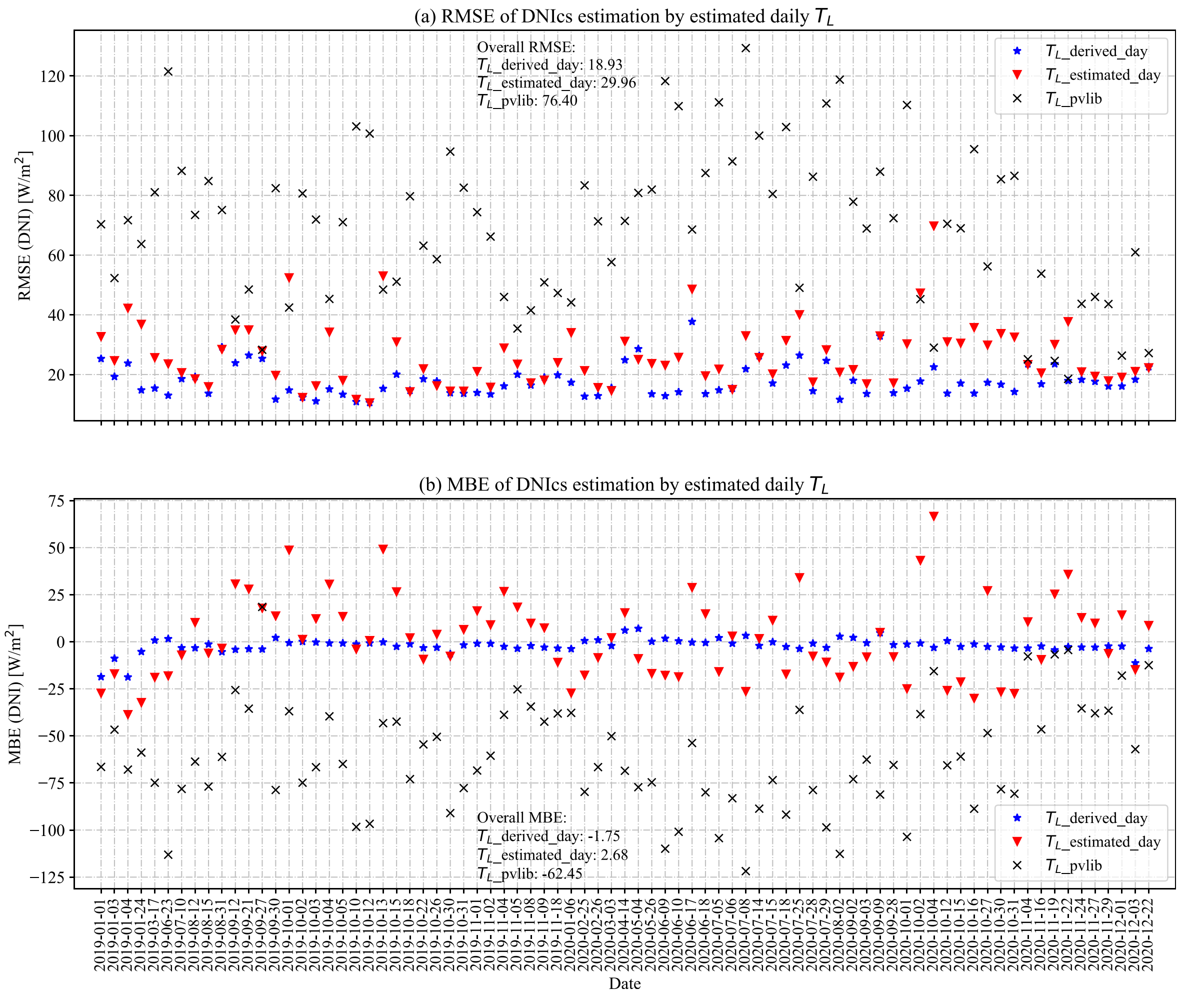

Fig. 12. The comparison of DNIcs estimation based on derived and estimated daily T_L factors. (a) RMSE of DNIcs estimation. (b) MBE of DNIcs estimation. Generally, the estimated T_L performs better than the default PVLIB T_L factor in terms of RMSE and MBE.

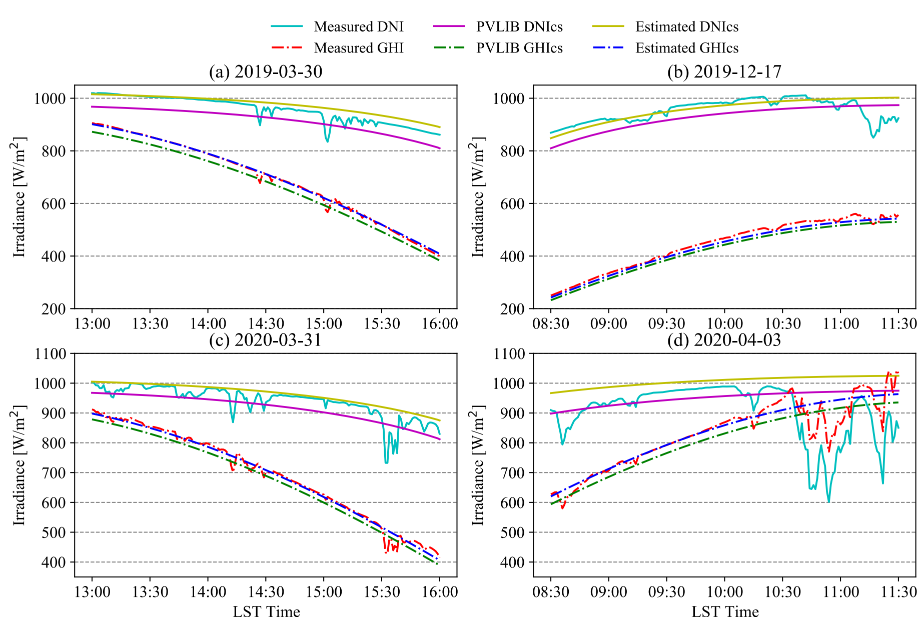

Fig. 13. DNI and GHI time series in partially cloudy days during (a) 2019-03-30 (b) 2029-12-27 (c) 2020-03-31 (d) 2020-04-03. Compared with DNIcs estimation in partially cloudy days, GHIcs estimated from improved T_L factor has higher accuracy in the clear-sky instants.

Table 4
Results of 1-min averaged GHIcs estimations using estimated T_L for clear-sky and partially clear days in 2019 for all the SURFRAD stations. PVLIB results from default T_L are presented in brackets for reference.

Stations	Clear-sky days		Partially clear days		Clear-sky and partially clear days	
	nRMSE [%]	nMBE [%]	nRMSE [%]	nMBE [%]	nRMSE [%]	nMBE [%]
BON	3.81 (10.68)	−1.90 (−9.08)	3.16 (9.24)	−0.96 (−7.80)	3.38 (9.16)	−0.37 (−7.58)
DRA	1.42 (4.09)	−0.20 (−3.33)	1.48 (3.79)	−0.17 (−3.02)	1.52 (3.80)	−0.13 (−2.99)
FPK	2.90 (7.03)	−1.91 (−4.83)	2.53 (5.32)	−0.78 (−2.97)	2.62 (5.09)	−0.48 (−2.59)
GWN	3.10 (8.52)	−1.19 (−7.52)	3.12 (7.06)	−0.64 (−5.27)	3.24 (7.00)	−0.44 (−4.97)
PSU	1.73 (8.47)	−0.28 (−7.81)	2.29 (7.01)	−0.25 (−6.07)	2.62 (7.11)	−0.35 (−6.04)
SXF	1.69 (7.24)	−0.09 (−6.66)	2.85 (6.13)	−0.03 (−5.01)	3.05 (6.29)	−0.16 (−5.14)
TBL	2.50 (2.73)	1.15 (−0.75)	2.37 (2.45)	1.40 (−0.47)	2.63 (2.66)	1.39 (−0.24)

measurements, thus leading to substantial accuracy improvement in estimating clear-sky irradiance. The major findings and recommendations are:

- We perform T_L estimation on the averaging basis of a day, an hour and every 5-min. Although 5-min averaged T_L can better represents its temporal variation, using daily or hourly averaged T_L to estimate GHIcs or DNIcs has no significant reduction in accuracy. Therefore, we recommend using the improved daily-averaging T_L (with less complexity and less computational resource requirement) for GHIcs or DNIcs estimations.
- Although the default Ineichen-Perez clear-sky model uses the same turbidity factor for GHIcs and DNIcs estimations, we found that using the same values would deteriorate DNIcs estimation. Therefore, we recommend using two separately trained ML models to generate different T_L values, one for GHIcs estimation and one for DNIcs estimation.
- During clear days, when compared with the default PVLIB T_L, the RMSE of GHIcs estimation based on the improved daily T_L decreased from 24.02 W m^{-2} to 9.94 W m^{-2}, a 58.6% reduction of error. The RMSE of DNIcs estimation is reduced from 76.40 W m^{-2} to 29.96 W m^{-2}, a 60.8% reduction of error. The default PVLIB generally underestimates the GHIcs and DNIcs with an MBE of −20.48 W m^{-2} and −62.45 W m^{-2}, respectively. The bias are corrected when using the improved daily T_L, yielding an MBE of 2.09 W m^{-2} for estimating GHIcs, and 2.68 W m^{-2} for estimating DNIcs, respectively.
- The daily T_L estimation method is also tested in partially cloudy days (with partial clear periods and partial cloudy periods). It is observed that the corresponding GHIcs and DNIcs estimations show better agreement with clear-sky irradiance measurements during cloudless time instances, when compared with default PVLIB results. The results indicate that the presence of clouds does not significantly change local air temperature and relative humidity, as well as water vapor and aerosol concentrations. Furthermore, the results demonstrate the potential of the proposed method in assisting solar irradiance modelling and forecasting in partially cloudy conditions, especially for cloud identification applications.

In sum, our proposed method offers a simpler way for T_L estimation without priori knowledge of aerosol and water vapor content in the atmosphere. The estimated T_L can substantially improve the accuracy of clear-sky GHI and DNI estimations when used in an empirical clear-sky model. Our results also imply that local meteorological data such as air temperature and relative humidity can represent column water vapor and aerosol concentrations with high accuracy during both clear and partially cloudy days. Solar resourcing and forecasting applications are expected to be improved when the proposed method is used to estimate clear-sky irradiance with higher accuracy.

CRediT authorship contribution statement

Shanlin Chen: Methodology, Software, Investigation, Writing – original draft. **Mengying Li:** Conceptualization, Resources, Supervision, Writing – review & editing.

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.

Acknowledgement

This work is partially funded by The Hong Kong Polytechnic University Grant P0035016. All the data used in this work are from the SURFRAD network. We would like to thank all the researchers and organizations who have made great effort in providing the high quality and publicly available data.

References

[1] J. Kleissl, Solar Energy Forecasting and Resource Assessment, Academic Press, 2013.
[2] Y. Chu, M. Li, H.T.C. Pedro, C.F.M. Coimbra, Real-time prediction intervals for intra-hour DNI forecasts, Renew. Energy 83 (2015) 234–244.
[3] M. Li, Y. Chu, H.T.C. Pedro, C.F.M. Coimbra, Quantitative evaluation of the impact of cloud transmittance and cloud velocity on the accuracy of short-term DNI forecasts, Renew. Energy 86 (2016) 1362–1371.
[4] Y. Chu, M. Li, C.F.M. Coimbra, D. Feng, H. Wang, Intra-Hour Irradiance Forecasting Techniques for Solar Power Integration: A Review, iScience, 2021, p. 103136.
[5] D.P. Larson, M. Li, C.F.M. Coimbra, Scope, Spectral cloud optical property estimation using real-time GOES-R longwave imagery, J. Renew. Sustain. Energy 12 (2) (2020), 026501.
[6] K.-N. Liou, An Introduction to Atmospheric Radiation, Elsevier, 2002.
[7] M. Chaâbane, M. Masmoudi, K. Medhioub, Determination of Linke turbidity factor from solar radiation measurement in northern Tunisia, Renew. Energy 29 (13) (2004) 2065–2076.
[8] R.H. Inman, H.T.C. Pedro, C.F.M. Coimbra, Solar forecasting methods for renewable energy integration, Prog. Energy Combust. Sci. 39 (6) (2013) 535–576.
[9] Y. Chu, M. Li, C.F.M. Coimbra, Sun-tracking imaging system for intra-hour DNI forecasts, Renew. Energy 96 (2016) 792–799.
[10] F. Antonanzas-Torres, R. Urraca, J. Polo, O. Perpiñán-Lamigueiro, R. Escobar, Clear sky solar irradiance models: a review of seventy models, Renew. Sustain. Energy Rev. 107 (2019) 374–387.
[11] X. Sun, J.M. Bright, C.A. Gueymard, X. Bai, B. Acord, P. Wang, Worldwide performance assessment of 95 direct and diffuse clear-sky irradiance models using principal component analysis, Renew. Sustain. Energy Rev. 135 (2021) 110087.
[12] P. Ineichen, A broadband simplified version of the Solis clear sky model, Sol. Energy 82 (8) (2008) 758–762.
[13] P. Ineichen, R. Perez, A new airmass independent formulation for the Linke turbidity coefficient, Sol. Energy 73 (3) (2002) 151–157.
[14] D. Yang, Choice of clear-sky model in solar forecasting, J. Renew. Sustain.

Energy 12 (2) (2020), 026101.
[15] C.A. Gueymard, REST2: high-performance solar radiation model for cloudless-sky irradiance, illuminance, and photosynthetically active radiation—validation with a benchmark dataset, Sol. Energy 82 (3) (2008) 272–285.
[16] X. Zhong, J. Kleissl, Clear sky irradiances using REST2 and MODIS, Sol. Energy 116 (2015) 144–164.
[17] F. Linke, Transmission coefficient and turbidity factor, J Beitraege Phys Fr Atom 10 (2) (1922) 91–103.
[18] P. Ineichen, Conversion function between the Linke turbidity and the atmospheric water vapor and aerosol content, Sol. Energy 82 (11) (2008) 1095–1097.
[19] J. Remund, L. Wald, M. Lefèvre, T. Ranchin, J. Page, Worldwide Linke turbidity information, ISES Solar World Congress 2003 400 (2003) 13. International Solar Energy Society (ISES).
[20] W.F. Holmgren, C.W. Hansen, M.A. Mikofski, Pvlib python: a python package for modeling solar energy systems, J. Open Source Software 3 (29) (2018) 884.
[21] C.L. Moldovan, R. Păltănea, I. Visa, Improvement of clear sky models for direct solar irradiance considering turbidity factor variable during the day, Renew. Energy 161 (2020) 559–569.
[22] J. Polo, L.F. Zarzalejo, L. Martin, A.A. Navarro, R. Marchante, Estimation of daily Linke turbidity factor by using global irradiance measurements at solar noon, Sol. Energy 83 (8) (2009) 1177–1185.
[23] L. Narain, S.N. Garg, Estimation of Linke turbidity factors for different regions of India, Int. J. Environ. Waste Manag. 12 (1) (2013) 52–64.
[24] X. Sun, J.M. Bright, C.A. Gueymard, B. Acord, P. Wang, N.A. Engerer, Worldwide performance assessment of 75 global clear-sky irradiance models using principal component analysis, Renew. Sustain. Energy Rev. 111 (2019) 550–570.
[25] T. Hove, E. Manyumbu, Estimates of the Linke turbidity factor over Zimbabwe using ground-measured clear-sky global solar radiation and sunshine records based on a modified ESRA clear-sky model approach, Renew. Energy 52 (2013) 190–196.
[26] C. Rigollier, O. Bauer, L. Wald, On the clear sky model of the ESRA—European Solar Radiation Atlas—with respect to the Heliosat method, Sol. Energy 68 (1) (2000) 33–48.
[27] R.H. Inman, J.G. Edson, C.F.M. Coimbra, Impact of local broadband turbidity estimation on forecasting of clear sky direct normal irradiance, Sol. Energy 117 (2015) 125–138.
[28] O. Behar, D. Sbarbaro, A. Marzo, L. Moran, A simplified methodology to estimate solar irradiance and atmospheric turbidity from ambient temperature and relative humidity, Renew. Sustain. Energy Rev. 116 (2019) 109310.
[29] J.A. Augustine, J.J. DeLuisi, C.N. Long, SURFRAD—A national surface radiation budget network for atmospheric research, Bull. Am. Meteorol. Soc. 81 (10) (2000) 2341–2358.
[30] C.A. Gueymard, J.M. Bright, D. Lingfors, A. Habte, M. Sengupta, A posteriori clear-sky identification methods in solar irradiance time series: review and preliminary validation using sky imagers, Renew. Sustain. Energy Rev. 109 (2019) 412–427.
[31] C.N. Long, T.P. Ackerman, Identification of clear skies from broadband pyranometer measurements and calculation of downwelling shortwave cloud effects, J. Geophys. Res. Atmos. 105 (D12) (2000) 15609–15626.
[32] C.N. Long, K.L. Gaustad, The Shortwave (SW) Clear-Sky Detection and Fitting Algorithm: Algorithm Operational Details and Explanations, Pacific Northwest National Laboratory, 2004.
[33] C. N. Long, D. D. Turner, A method for continuous estimation of clear-sky downwelling longwave radiative flux developed using ARM surface measurements, J. Geophys. Res. Atmos. 113 (D18).
[34] L.D. Riihimaki, K.L. Gaustad, C.N. Long, Radiative flux analysis (RADFLUXANAL) value-added product: retrieval of clear-sky broadband radiative fluxes and other derived values, in: Tech. rep., ARM Data Center, Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States), 2019.
[35] C.A. Gueymard, Analysis of monthly average atmospheric precipitable water and turbidity in Canada and northern United States, Sol. Energy 53 (1) (1994) 57–71.
[36] C.A. Gueymard, Assessment of the accuracy and computing speed of simplified saturation vapor equations using a new reference dataset, J. Appl. Meteorol. Climatol. 32 (7) (1993) 1294–1300.
[37] C. M. Bishop, Pattern recognition, Mach. Learn. 128 (9).
[38] G. Biau, E. Scornet, A random forest guided tour, Test 25 (2) (2016) 197–227.
[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[40] M. Ali, PyCaret: an Open Source, Low-Code Machine Learning Library in python, PyCaret version 2.