Forecasting Photosynthetic Photon Flux Density Under Cloud Effects: Novel Predictive Model Using Convolutional Neural Network Integrated With Long Short-term Memory Network

Ravinesh C Deo (✉ ravinesh.deo@usq.edu.au)
University of Southern Queensland https://orcid.org/0000-0002-2290-6749

Richard H Grant
Purdue University

Ann Webb
The University of Manchester

Sujan Ghimire
University of Southern Queensland

Damien P. Igoe
University of Southern Queensland

Nathan J. Downs
University of Southern Queensland

Mohanad S Al-Musaylh
Southern Technical University

Alfio V. Parisi
University of Southern Queensland

Jeffrey Soar
University of Southern Queensland

Research Article

Keywords: Photosynthetic radiation, deep learning, stochastic cloud effects, solar radiation modelling, photosynthetic photon flux density, risk evaluation

Posted Date: November 18th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1069113/v1

Forecasting photosynthetic photon flux density under cloud effects: *novel predictive model using convolutional neural network integrated with long short-term memory network*

Ravinesh C Deo[1*], Richard H Grant[2], Ann Webb[3], Sujan Ghimire[1], Damien P. Igoe[1], Nathan J. Downs[1], Mohanad S Al-Musaylh[4], Alfio V. Parisi[1], Jeffrey Soar[5]

[1] School of Sciences, University of Southern Queensland, Australia

[2] Dep. of Agronomy, Purdue Univ., West Lafayette, IN, United States

[3]Dep. of Earth and Environmental Sciences, Faculty of Science and Engineering, University of Manchester, Manchester, M13 9PL, United Kingdom

[4] Department of Information Technologies, Management Technical College, Southern Technical University, Basrah 61001, Iraq

[5] School of Business, University of Southern Queensland, Australia

[*]Corresponding Author (Prof Ravinesh Deo): ravinesh.deo@usq.edu.au

Abstract

Forecast models of solar radiation incorporating cloud effects are useful tools to evaluate the impact of stochastic behaviour of cloud movement, real-time integration of photovoltaic energy in power grids, skin cancer and eye disease risk minimisation through solar ultraviolet (UV) index prediction and bio-photosynthetic processes through the modelling of solar photosynthetic photon flux density (PPFD). This research has developed deep learning hybrid model (*i.e.*, CNN-LSTM) to factor in role of cloud effects integrating the merits of convolutional neural networks with long short-term memory networks to forecast near real-time (*i.e.*, 5-minute) *PPFD* in a sub-tropical region Queensland, Australia. The prescribed CLSTM model is trained with real-time sky images that depict stochastic cloud movements captured through a Total Sky Imager (TSI-440) utilising advanced sky image segmentation to reveal cloud chromatic features into their statistical values, and to purposely factor in the cloud variation to optimise the CLSTM model. The model, with its competing algorithms (*i.e.*, CNN, LSTM, deep neural network, extreme learning machine and multivariate adaptive regression spline), are trained with 17 distinct cloud cover inputs considering the chromaticity of red, blue, thin, and opaque cloud statistics, supplemented by solar zenith angle (SZA) to predict short-term *PPFD*. The models developed with cloud inputs yield accurate results, outperforming the *SZA*-based models while the best testing performance is recorded by the objective method (*i.e*., CLSTM) tested over a 7-day measurement period. Specifically, CLSTM yields a testing performance with correlation coefficient $r = 0.92$, root mean square error $RMSE = 210.31$ µ mol of photons m^{-2} s^{-1}, mean absolute error $MAE = 150.24$ µ mol of photons m^{-2} s^{-1}, including a relative error of $RRMSE = 24.92\%$ *MAP*E = 38.01%, and Nash Sutcliffe's coefficient $E_{NS} = 0.85$, and Legate & McCabe's Index $LM = 0.68$ using cloud cover in addition to the *SZA* as an input. The study shows the importance of cloud inclusion in forecasting solar radiation and evaluating the risk with practical implications in monitoring solar energy, greenhouses and high-value agricultural operations affected by stochastic behaviour of clouds. Additional methodological refinements such as retraining the CLSTM model for hourly and seasonal time scales may aid in the promotion of agricultural crop farming and environmental risk evaluation applications such as predicting the solar UV index and direct normal solar irradiance for renewable energy monitoring systems.

Keywords Photosynthetic radiation; deep learning; stochastic cloud effects; solar radiation modelling; photosynthetic photon flux density; risk evaluation

1.0 Introduction

The global solar radiation used by plants in photosynthesis spans about 400-700 nm wavelength, which is a relatively narrow part of the entire solar spectrum, but one containing only about half the solar energy. Within this limits can be defined both the energy available for photosynthesis, the Photosynthetically Active Radiation (*PAR*, Wm^{-2}) or alternatively, the Photosynthetic Photon Flux Density (*PPFD*; µ mol of photons $m^{-2}s^{-1}$) [1] that will now be the subject of this paper. Lipid proteins, forming the building block of terrestrial and marine food webs, contribute to global biomass derived from agricultural animal and plant products that continue to be a growing source of worldwide energy production. Currently, green biofuels account for 11% of the world's total energy supply [2] coming from primary plant and vegetable oil crops, secondary lignocellulosic by-products [3, 4], and third generation, enriched lipid microalgae bioproducts.

Significant research has focused on the optimisation of biofuel production particularly through the efficient production of microalgae photo-bioreactors (PBR) that can optimise the light, temperature, nutrient loads, and continuity of microalgae species [5-7]. Recent research works concentrated on the genetic modification of microalgae species for optimal acclimation to the environment. These are aimed at enhancing the overall output efficiency of the targeted microalgae products [8-10]. Alternative energy resources for PBR have also been investigated by including artificial light or organic fluorescent dyes to maximise solar conversion into optimal photosynthetic radiation bands [4]. Costs of artificial light sources have to date restricted the development of PBRs that do not retain enough access to reliable sources of photosynthetic-active solar radiation. Importantly, the availability of open-air setups utilising natural sunlight continues to be the most economically viable solution to farm microalgae and develop sustainable bio-products. These systems are by far the most prevalent, roughly occupying 90% of all third-generation commercial biofuel production facilities [11]. They are however dependent on both long and short-term fluctuations in localized-scale solar radiation where production can be improved by monitoring farms with robust forecasting efforts especially in real-time scales.

Solar radiation, affected by season, latitude and temporal variations in cloud cover, ozone, and atmospheric aerosols, influences the optimal utilisation of light at any given biomass production system, including its effect on plant growth or overall health. Typically, tropical environments that produce consistently high levels of solar insolation at the earth's surface are ideal [12]. However tropical climates are frequently affected by strong seasonal precipitation patterns resulting in fluctuations in solar light intensity. Cloud cover alone can drop the available Photosynthetic Photon Flux density (*PPFD*), which can reach 2000 μmol of photons m^{-2} s^{-1} at noon, by as much as 80% [5, 13]. Broken cloud can bring about short-term cloud enhancement of solar radiation (up to ~20%) and such conditions can bring about rapid fluctuation of solar radiation both above and below the clear sky values. Yet, ideally, efficient biomass production requires a steady and reliable supply and monitoring of *PPFD* [15].

As net primary productivity is strongly influenced by climatic factors, much effort has been expended on measuring (and subsequently monitoring) the *PPFD*. A review of literature shows some limitations in terms of current predictive approaches where most methods have used monitoring rather than real-time forecasting approaches. Remote sensing platforms have been used to determine vegetation net production efficiency [16] and as a result can be used to determine the best locations for establishing farms, greenhouses or other high value agricultural hubs [12, 17-19]. Satellite remote sensing methods inherently must approximate the geometric absorption, scattering and transmission of clouds from relatively low resolution single-direction reflectance [20]. The most important environmental predictors to determine the global *PPFD* on the earth's surface are the annual precipitation, monthly cloud fraction, bioclimate layer information and month [21-25].

Having identified the best location for crops, the next step would be to forecast solar radiation conditions so that crops are protected and their growth is optimised. The seasonal and climatic factors which can be readily sourced from public datasets have been employed in previous AI-based approaches too, particularly to accurately predict agricultural crop yield, drought indices and rainfall in Pakistan [28, 29], China [30], USA [31] and Australia [32-34]. Such AI-based approaches are becoming useful tools to derive agricultural and biomass product efficiency mapping on a much broader scale where accurate surface instrumentation and local climate records are not available. Hemispherical photographs have been used to estimate *PPFD* with limited success[35]. Another approach has been artificial neural network (ANN) models that map out the available global surface *PPFD* using remote satellite products as predictor variables. This model, however, is based on an ANN approach that requires environmental predictors to produce an accurate forecast system [36].

Biomass productivity is not only dependent on total *PPFD* but also the diffuse fraction of *PPFD [37]*. Methods for retrospective *PPFD* estimation employ a mixture of remote satellite products, global reanalysis of climate information [38] and local surface instrumentation [39] to model both direct and diffuse photosynthetic-active radiation and output biomass for a range of ecological and agricultural applications have also been developed [21].

In respect to solar energy, monitoring or integration into electricity grids, intermittencies in power production are highly driven by cloud variations [40]. However, the ability to develop reliable models to predict short-term (*e.g*., 5-10 minute) solar radiation can provide a future solar system real-time monitoring capability to resolve clean energy challenges by better capturing cloud cover, lifetime, spread or stochastic movements. Also, the option to capture cloud cover variations in a solar ultraviolet index (UV Index) model such as the one developed previously by Deo *et al*., [41] can help in skin cancer and eye disease risk mitigation. Developing a PPFD prediction model trained with cloud images may provide useful insights into UV index, solar power production or energy demand monitoring.

In a previous study, the near real-time PPFD prediction model of Deo *et al*. [39] was based on an adaptive neuro-fuzzy inference system to predict *PPFD* over 5-minute horizons in Queensland (Australia), using time lagged *SZA* data under cloud-free conditions. Utilising the local solar zenith angle (*SZA*) as the only input variable, they demonstrated good accuracy in predicting the real-time *PPFD* with changes in *SZA* for 5 minute and hourly forecasts. Such studies that model real-time solar photosynthetic energy can play a pivotal role in helping explore regional development of the agricultural sector. However, the inclusion of cloud cover (which is vital for the control of plant growth, was not considered in previous studies. The development of an AI-based model to predict the influence of cloud variations at near real-time, and how the cloud properties (derived from image chromic information) might control the amount of ground-based photosynthetic-active radiation is yet to be explored.

This paper develops an artificial intelligence (AI)-approach that considers the total sky conditions, addressing the role of cloud cover variations to accurately model *PPFD* at 5-minute time scales. The contribution and novelty are to build a first deep learning AI method for real-time *PPFD* forecasting, capturing the influence of cloud properties on measured photosynthetic-active radiation.

A deep learning-based methodology utilising whole sky image characteristics of both the cloud and cloud-free conditions typical to local farming environments incorporates data features from high temporal resolution images such as those captured by Total Sky Imager (TSI) or geo-stationary satellites *e.g*., Himawari 8 or 9 providing inter-minute level sky images. The objectives are as follows. (1) To process *TSI*-based cloud images corresponding to *PPFD* measured at 5-minute intervals through a custom-built cloud segmentation algorithm [42] applied to each image, and produce descriptive statistics based on the blue, red, thin and opaque cloud chromatic features (*i.e*., means, standard deviations, differences, ratios). These are then used to build an optimal set of model inputs (*i.e*., cloud image properties) against a target (*i.e*., *PPFD*). (2) To develop deep learning-based convolutional neural network and long short-term memory network (CLSTM) model following our earlier study [43], implemented for near real-time PPFD forecasting. (3) To benchmark the CLSTM model *w.r.t* conventional machine learning (MARS, ELM) and deep learning LSTM, CNN and DNN methods tested on the same training and testing subsets. To pursue the objectives, the present study has utilised data from a local TSI as a proof of concept. The parameters employed are cloud fraction, cloud type and the red-green-blue cloud chromatic properties derived from segmented sky images, with respect to simultaneous *PPFD* measurement at the subtropical location of Toowoomba (27.6°S), Australia.

2.0 Theoretical Overview

The theoretical details of deep learning (*i.e*., CNN, LSTM, DNN) and conventional machine learning (ELM and MARS) methods are described elsewhere [43-48]). The CLSTM model, constructed by integrating CNN and LSTM, had been used elsewhere in natural language processing where emotions were analysed with text inputs [49], in speech processing where voice search tasks were performed using CLDNN combining CNN, LSTM and DNN [50], in video processing with CNN and Bi-directional LSTM models built to recognize human actions in video sequences [51], in the medical area where the CNN-LSTM method was developed to detect arrhythmias in electrocardiograms [52] and in industrial areas where a convolutional bi-directional LSTM model was designed to predict tool wearing [53]. Other studies with CLSTM are evident, for example, time series application for prediction of residential energy consumption [54] [55], solar radiation prediction [43, 56-58] and wind speed prediction [59-61] as well as stock market applications in the prediction of share prices [62, 63]. In the solar radiation forecasting area, the study of Ghimire *et al.* [43] has developed a CLSTM model and compared its performance against the CNN, LSTM and DNN-based models, showing that the CLSTM model outperformed the standalone version of both CNN and LSTM models.

<Fig. 1>

Following earlier implementations [43], in this study we integrate CNN and LSTM to produce a hybrid system that ensures most prevalent data features are extracted using CNN prior to the sequential modelling of real-time photosynthetic radiation at 5-minute intervals. This objective model is depicted by a simplified schematic architecture in Figure 1. Generally, a CNN system is known to extract local trends or other features as well as common features recurring in time series at different intervals [64] and then used to serve as further inputs to LSTM model's architecture. LSTM is able to capture both the short- and the long-term dependencies in data patterns (*e.g.*, linking PPFD variability against time-based cloud movements) to learn the time sequential relationships among predictors and a target [65, 66]. First introduced for object recognition in image processing [67], the CNN model has a prominent structure composed of many convolution layers, pooling layers and one or more fully connected layer [62]. The primary building block applies a convolution filter (*i.e.*, a kernel function) for input data to generate a feature mapping scheme [68]. Using different filters, many sets of convolutions are performed in order to create different feature maps [69]. These are eventually combined to produce the convolution layer's final output. In the pooling layer, each feature map's dimension is reduced through down-sampling thereby mitigating the risks of model overfitting and reducing the model's training time [70]. The fully-connected layer at the end of the CNN is replaced with LSTM via the flattening layer to produce the hybrid CLSTM predictive model [71].

Other than the CLSTM model, the present study has utilised a standalone LSTM as a variation on Recurrent Neural Network (RNN) composed of memory cells coupled through layers, rather than the neurons in a conventional ANN-type model [72]. The RNN is generally considered to be somewhat incompetent in describing long-term dependences due to the gradient vanishing phenomenon [73]. Because of this, LSTM was developed by Hochreiter and Schmidhuber in 1997 [74] and enhanced by Graves in 2013 [75]. In contrast to the classic RNN where gradients back-propagate exponentially, the LSTM model allows for gradients to flow unchanged by employing a cell memory. By using input gate, a forget gate, and an output gate, the LSTM unit can decide what to remember and what to forget and is therefore capable of addressing long-term dependencies. [76]. In general, an LSTM block is made of the sigmoid (*σ*) and hyperbolic tangent (*tanh*) layers, and two operations including pointwise summation (⊕) and multiplication (⊗) operations, as shown schematically in Figure 1. Mathematically, these processes can be defined by equations 1-6 [43].

Input gate i_t:

$$i_t = \sigma(w_i x_t + R_i h_{t-1} + b_i) \quad <1>$$

Forget gate f_t:

$$f_t = \sigma(w_f x_t + R_f h_{t-1} + b_f) \quad <2>$$

Output gate y_t:

$$y_t = \sigma(w_y x_t + R_y h_{t-1} + b_y) \quad <3>$$

Cell c_t:

$$c_t = f_t c_{t-1} + i_t \bar{c_t} \quad <4>$$

$$\bar{c_t} = \sigma(w_c x_t + R_c h_{t-1} + b_c) \quad <5>$$

Output vector h_t:$h_t = \mathrm{y}_t \sigma(c_t)$ <6>

where, *σ* and *tanh* are activation functions in the range [0,1] and [1,1] respectively,

Sigmoid function: $\sigma(\gamma) = \frac{1}{1+e^{-\gamma}}$ <7>

Hyperbolic-tangent function: $\sigma(\gamma) = \frac{e^{\gamma}-e^{-\gamma}}{e^{\gamma}+e^{-\gamma}}$. <8>

b_i, b_f, b_y denote the input, forget, and output gate bias vectors, respectively;

c_{t-1} and h_{t-1} are the previous cell and its output vector;

h_t is the output vector;

x_t denotes the input vector;

w_i, w_f, *and* w_y are the matrix of weights from the input, forget, and output gates to the input, respectively; and

R_i, R_f, *and* R_y define the matrix of weights from the input, forget, and output gates to the input, respectively.

3.0 Materials and Method

3.1 Experimental Apparatus and Data Acquisition System

Photosynthetic photon flux density, *PPFD*, was measured with corresponding cloud cover images at the Toowoomba Campus of The University of Southern Queensland 120 km west of Brisbane, Australia. Fig. 2(a) shows the geographic location of the study site. At the University's Atmospheric and Solar Ultraviolet Radiation Laboratory, a quality-controlled monitoring station measured *PPFD* and weather conditions since 2011 (Fig. 2b). Located at an elevation of 690 m *ASL*, Toowoomba is a regional city with a high solar energy potential and is also classified as a regional centre for agricultural activities that makes the *PPFD* forecast models an advantageous tool for practical applications in agricultural sectors. The specific study site also has a relatively large number of full sunshine days and a clear hemispheric view of the solar horizon [77] that also makes it an ideal site to implement the CLSTM model for real-time forecasting of photosynthetic-active radiation.

<Fig 2(a-d)>

To build the proposed CLSTM predictive model, high-quality, yet cloud-influenced measurements of *PPFD* were acquired over the austral summer solstice period (01 to 31 Mar 2013). The data were collected using a Quantum sensor (LI-190R; LI-COR, Lincoln, USA) connected to a CR100 Campbell Scientific data logger (Logan, USA) (Fig. 2). The LI-190R automated system was installed on an unobstructed rooftop site to continuously monitor the photosynthetic-active radiation at 5-minute intervals over a 24-hr period. Employed in several other research works [39, 78, 79], the LI-190R system is mainly designed for long-term, outdoor usage with a manufacturer-stated uncertainty of ±5% traceable to the US National Institute of Standards and Technology. In this paper, the *PPFD* time series for the daytime period 07.00 AM—05.00 PM were used, considering that solar irradiance is mainly intercepted by plants during daytime, and that the night level of photosynthetic energy is practically zero.

<Fig 3(a-b)>

Figure 3(a) shows the temporal patterns in measured *PPFD* time series sampled at 5-minute intervals, ranging from 0 to 2300 μ mol of photons m^{-2} s^{-1} but this variation over entire diurnal cycles is different for different days or times. This is perhaps due to cloud cover or atmospheric conditions (*e.g.*, ozone, aerosols, water vapor). Fig. 3(b) shows a sample of five cloud images with their respective *PPFD* and solar zenith angle. It is noticeable that even for a similar value of *SZA* (28-29°) at 10.55 AM (10 Mar) and 12.55 PM (15 Mar), the value of *PPFD* varies by almost 28%. Similar observation can be made for the data on 01 March (06.55 AM) and 30 March (16.55 PM) measuring the *PPFD* values of 54 μ mol of photons m^{-2}s^{-1} and 333 μ mol of photons m^{-2}s^{-1}. Meanwhile here is rather similar PPFD for March 30th and March 5th even though SZA changes considerably. This illustrates how cloud fraction is an important modulator of SZA-controlled photosynthetic-active radiation, including cloud height and depth that are not considered in this analysis.

3.2 Sky Image Processing and Cloud Segmentation

A quick and efficient self-adaptive Python-based tool called the *TSI Analyser* developed in earlier work [42] is employed for sky image segmentation and extraction of cloud chromatic properties from images obtained by Total Sky Imager (*TSI*) instrument (serial number: 175). Details of the *TSI Analyser* algorithm are described elsewhere [42] but in principle, it is able to produce cloud cover-based statistical properties for *every* image that is associated with a measured *PPFD* value. This aims to capture the overall sky conditions, particularly, to include the contributory role of cloud cover variations in training the proposed CLSTM predictive model. To do this, we refer to comparisons between red and blue intensities in clouds, red-blue ratios, and red-blue difference. We also segmented each image into the normalized red-blue-ratio that was undertaken in our earlier paper [36] based on the commonly used red-blue ratio [80] such that the *TSI440*-based pixel values of each of the red and blue channels were determined. It is noteworthy that the normalized ratios are consistent with conventional cloud detection methods with practical importance in cloud segmentation [81]. It is also important to note that the red (*R*) to the blue (*B*) ratio maintains a higher relative resolution despite the down sampling that occurs when the images are saved in *.jpeg* format. To acquire images, the *TSI440* enables a user defined threshold for opaque and thin clouds [82] with the latter cloud type presenting a difficulty in cloud segmentation especially when aerosols are present [83], which is not further considered in this study, assuming everything captured by the user threshold to be thin cloud.

The *TSI Analyser* was applied to a 1-month dataset with 5-minute interval cloud images considering over 200,000 images collected at a 480 × 320 spatial resolution. These whole-sky images have been captured using *TSI440* [84-86] used in previous research (*e.g.*, [82, 87, 88]). The *TSI440* instrument consists of a reflective dome with a camera suspended above it [89, 90] pointing downwards to generate a *.jpeg* format colour image of the whole sky. A non-corrupted sky image array is then read using commands from the *NumPy* library in Python [91] by means of the OpenCV library's *imread* command [92]. This is converted from OpenCV's blue-green-red (*BGR*) to red-green-blue (*RGB*) format for further image processing.

Table 1 summarises the data for cloud chromatic properties derived from segmented images including the descriptive statistics (*i.e.*, mean, standard deviation, difference, & ratio) based on the blue, red, thin, and opaque cloud (pixelized) features *per* image. The segmentation algorithm produced the average of the whole sky blue (B_{av}), whole sky red (R_{av}), as well as the statistical features based on standard deviation, ratios, or differences of the blue (B) and red (R) pixel values for clouds that represent the estimated proportion of pixelized cloud features likely to be a function of the photosynthetic-active radiation received at a measuring sensor. To analyse the degree of associations between cloud movement and an instantly measured *PPFD* value, a cross correlation analysis is performed to determine the covariance measured by r_{cross} prior to developing the proposed CLSTM model. Table 1 includes the r_{cross} used to determine the order of our model input combinations, presented in Table 2. It is evident that the average of whole sky-blue pixel in a total sky image appears to generate the largest value of r_{cross} ~ -0.747, followed by the standard deviation of the blue cloud pixel (r_{cross} ~0.640). This exceeds an r_{cross} value of -0.631 computed for solar zenith angle that is traditionally used as the only predictor variable of photosynthetic-active radiation as per other studies (*e.g.*, [39]). This analysis also shows that the covariance of the whole sky-blue average and the standard deviation of the blue cloud pixels are more strongly correlated with *PPFD* compared with the *SZA* dataset.

<Table 1>

To corroborate the findings in Table 1 we now inspect visually the covariance in cloud chromatic properties against measured photosynthetic-active radiation. Figure 4 displays a scatterplot of the cloud cover statistics as well as *SZA* data that are regressed against the measured *PPFD* in the model training phase. The whole sky-blue average is seen to attain the highest coefficient of determination ($r^2 = 0.549$) with respect to the *PPFD* values. The other significant predictor variables are found to be the blue cloud pixel standard deviation ($r^2 = 0.403$), solar zenith angle ($r^2 = 0.403$) and the standard deviation of the whole sky-blue ($r^2 = 0.365$). It is especially notable that the ratio of red to blue sky and the difference between the blue and red pixels in a whole sky image appears to be weakly correlated with *PPFD* data series, and therefore, may not contribute significantly towards improving the proposed CLSTM model. Taken together, the present analyses clearly ascertain that at least two of the cloud chromatic properties (*i.e.*, whole sky blue & blue cloud pixel averages associated with measured *PPFD*) are more strongly correlated with *PPFD*, compared with the solar zenith angle used in earlier studies. This deduction confirms that the inclusion of cloud cover properties may be a crucial task used to improve earlier models for photosynthetic-active radiation (*e.g.*, [39]).

<Fig 4>

A comparison of the *PPFD* data series within the first 7 days of model training data is made against cloud-image derived predictor series in Figure 5. Note that here, the first 847 points are employed to demonstrate the association of *PPFD* and cloud property before developing the proposed CLSTM predictive model. While the changes in *PPFD* are not well-represented by *SZA* due to the solar zenith angle presenting a much smoother variation over any given diurnal cycle, there is a clear temporal correspondance between the magnitude of *PPFD* with many of the cloud-image statistical features. This correspondance is especially pronounced on the *x*-axis scale from the datum point 363 to 847 for image pixels representing the whole sky blue average and its standard deviation, and the standard deviation of the blue cloud pixels. Interestingly, for the whole sky red average pixels, the standard deviation of the red cloud pixels, the average of blue cloud pixels, the whole sky red-blue ratio, the standard deviation of the whole sky red and the difference of red-blue pixels are also demonstrating a good degree of harmony in terms of their temporal variation against the *PPFD* timeseries. While the direct association between some of the cloud chromatic properties is not so clear, as expected, there does appear to be a moderating effect in terms of the jumps in *PPFD* against any cloud property. This indicates that the subtle, yet non-linear effects of cloud movements on photosynthetic-active radiation should be captured in a *PPDF* forecast model.

<Fig 5>

<Table 2>

3.3 Predictive Model Design

To develop the objective hybrid model (*i.e.*, CLSTM) and benchmark (or comparative) models using deep learning (LSTM, CNN, DNN) and machine learning (ELM & MARS) algorithms, both the python [93] and the MATLAB-based [94] scripts were implemented on Intel *i*7 computer with 3.40 GHz processor running on 32GB memory. Figure 6 illustrates the model development stage and Table 2 lists the input combinations used in all designated models together with the details of data partitioned in the training (53.3%), validation (23.3%), and testing (23.3%) subsets.

<Fig. 6>

To build an accurate CLSTM model that can consider the role of cloud cover variations, particularly by using cloud chromatic properties to generate near real-time photosynthetic-active radiation forecasts, an optimal arrangement of the model's inputs is firstly deduced. A sequential ordering approach (*e.g.*, [95]) is adopted where ranked cross-correlation coefficients r_{cross} deduced from the respective predictor variable as illustrated Table 1 (*i.e.*, cloud-based time series, or solar zenith angle derived from an empirical method [96]. This proposed method led to the first predictive model (M_1) being constructed using the average of whole sky blue (B_{av}) pixels, followed by the second model (M_2) with both the B_{av} and the standard deviation of blue cloud pixels (BC_{sd}) pixels and the third model (M_3) having B_{av}, BC_{sd} and solar zenith angle (*SZA*) as enunciated by Table 2.

By inclusion of cloud properties, this study advances earlier work [39, 88] where *SZA* was the only predictor used to forecast *PPFD* and solar UV index ignoring cloud variations. This study advances the standard approaches [39, 88] that utilize only *SZA* neglecting the role of clouds in modulating *PPFD*. It is noteworthy that successive addition of series based on r_{cross} concurs with earlier prediction problems [95] aimed at evaluating potential improvements in CLSTM model. To evaluate the utility of a cloud-free model, a standard approach used in photosynthetic-active radiation [39], solar UV index [88] and global solar models [95]), a CLSTM model designated as M_{18}, with only the *SZA*, was constructed as a reference model without any inclusion of cloud cover properties. Overall, the model design process resulted in 18 distinct predictive models, as stated Table 2.

<Fig. 7>

As this study's intent is to build a forecast model that can accurately predict the photosynthetic-active radiation at a future timescale over near real-time (5-minute) intervals, we have further explored the cross correlation between cloud chromatic properties and photosynthetic-active radiation (or *PPFD*) using a time-lagged correlogram. Figure 7 identifies the covariance between *PPDF* (*i.e.*, target) and *SZA*, along with all of the other cloud-image derived predictor variable data in the model training phase. Evidently, the lagged series show a strong (±) serial correlation exceeding the statistically significant region at the 95% confidence which is indicated by a blue line. Interestingly, the correlation coefficient in terms of the time-shifted cloud properties for non-zero lag (*i.e.*, occurring for an input that was regressed on a target at a different timescale) is also prominent for some of the inputs (*e.g.*, thick clouds, average of red pixel values in the cloud cover, difference between whole sky red and the blue pixels, and the ratio of red to the blue pixel values in the clouds). This indicates a strong non-linear association between cloud chromatic properties and photosynthetic-active radiation, potentially indicating the need for a non-linear modelling approach to forecast photosynthetic-active radiation. To construct the proposed CLSTM model, all of the cloud chromatic properties and the *SZA* measured over a time lag of 5 minutes is used:

$$PAR(t+1) = f\{\mathrm{X}(t)\} \qquad (9)$$

where *PAR* (*t* + 1) denotes the photosynthetic photon flux density (*PPFD*, µmol of photons $m^{-1}s^{-1}$) at a next time interval of 5-minute time horizon, *X* (*t*) is the relevant input and *t* is the time scale. Prior to the modelling process, all inputs and the target were scaled to be between [0, 1] where:

$$X_N = \frac{X-\hat{X}}{\hat{X}-\breve{X}} \qquad (10)$$

where

X_N = Normalized values of a variable X

X = Actual value of a variable X

$\hat{X}$ = Maximum value of a variable X

$\breve{X}$ = Minimum value of a variable X

To identify the contributory effects of cloud variations in forecasting 5-minute photosynthetic-active radiation, this study firstly develops a 3-layered convolutional neural network (CNN) and long short term memory network (LSTM) with a 4-layered deep neural network (DNN), and multivariate regression spline (MARS) and extreme learning machine (ELM) models. Following the benchmark methods, CNN and LSTM algorithms were integrated in accordance with earlier study [43] to generate a 4-layered objective model (denoted as hybrid CLSTM). For model development parameters, see Appendix (Table A1 a-c). In general, for the CLSTM architecture, the first half comprised of the CNN used for feature extraction whereas the second half comprised of the LSTM algorithm used to forecast *PPFD* by incorporating these CNN-grained input features.

3.3.1 Common Hyperparameters for Deep Learning (DL) Models

Open-source DL Python libraries, Scikit-Learn [97] and Keras[98, 99] were used to implement CNN, LSTM and DNN algorithms. Hyperparameters of all benchmark models were deduced through grid search. In this study, the DL models share the following four common hyperparameters.

- *Activation functions*: Except for the output layer, all of the network layers relied on the same activation function, which accords to the other studies [100, 101] so we have used the rectified linear unit (*ReLU*) [102].
- *Dropout*: This is considered as a potential regularization to minimize overfitting issues in order to improve the training performance [103]. The dropout aims to select a fraction of the neurons (defined as a real hyperparameter over the range 0 to 1) at each model iteration and prevent them from retraining [104-106]. For this study, this fraction of neurons was maintained to be 0.1.

- *Two statistic regularization*. This included L1 (*i.e.*, least absolute deviation) and L2 (*i.e.*, least square error) applied together with the dropout. It is imperative to mention that the role of L1 and L2 penalization type parameters is to minimize the sum of the absolute differences and the sum of the square of the differences between the forecasted and target PPFD values, respectively [107-109]. Also, the addition of a regularization to the loss is to encourage smooth network mapping in the DL network, particularly by penalizing the large parameters values to reduce the level of nonlinearity in the network models [110, 111].
- *Early stopping*: The issue of overfitting can be further addressed by introducing an early stopping (ES) phase in Kera [98, 112] so that the mode is set to a minimum while the patience is set to 30 [110, 113, 114]. This is done to also ensure that the training process will terminate when the decrease in the validation loss has stopped for a number of patience-specified epochs [115-117].

3.2.2 CNN Hyperparameters and Hybrid CNN-LSTM Model Development

The CNN model's hyperparameters were also optimised that included the following options.

- *Filter size*: The size of the convolution operation filter was optimised.
- *Number of convolutions*: The number of convolutional layers in each CNN was optimised.
- *Padding*: This study has utilized the same padding in order to ensure that the input feature map and output feature map dimensions were identical [118].
- *Pool-size*: A pooling layer was used between each convolution layer to avoid further overfitting. This pooling layer also helps decrease the number of parameters and network complexity [119]. In this study, we have utilized a pool-size of 2 between the layer 1 and 2 of the CNN model.

Finally, the hybrid CNN-LSTM model comprised of 3 convolutional layers, with pooling operations where a selection of the convolutional layer channels was based on grid search process. In the model's architecture, the outputs of flattening layer served as the inputs of LSTM recurrent layer while the LSTM recurrent layer was directly linked to the final output.

3.4 Non-deep Learning Benchmark Models

This study develops ELM and MARS models (as benchmark methods) considering their relative success in solar predictive problems [88]. The ELM architecture composes of a single hidden layer system with 17 input neurons (to enable cloud cover and *SZA*-based inputs to be fed in) (Table 3c), a maximum of 1000 hidden neurons and 1 output neuron allocated to the forecasted *PPFD*. To optimise the ELM model, this study tests several activation functions (*i.e.*, sine, hard limit, radial basis, triangular basis, logarithmic sigmoid & tangent sigmoid equations) following earlier approach [88] with an optimal model achieved using logarithmic sigmoid equation indicated in Table 3(c). To identify an optimal ELM architecture, the hidden neuron was varied from 1 to 1000 with each architecture then evaluated on a validation dataset (25% in this study) to identify the optimal architecture. As ELM requires random initialization of hidden layer parameters, the model was run 1000 times with the lowest root mean square error (*RMSE*) over all hidden nodes used to select the optimal ELM model. The optimal ELM was denoted as 10–23–1 (input–hidden–output) which included 10 predictor variables and 23 hidden neurons to attain the most accurate forecasts of PPFD data.

For the MARS model, an ARESLab-based MATLAB toolbox (ver. 1.13.0) [120] is adopted. Out of the two basis functions (*i.e.*, cubic & linear) within its piecewise equation, the cubic form is adopted [121] given its capacity to handle multiple predictors. The generalized recursive partitioning regression (RPR) is also employed as an adaptive algorithm for function approximation [122] with the process including a forward and backward deletion process to reach the optimal MARS equation. In the forward phase, a ‘naïve’ model with just the intercept term is used with iterative addition of the reflected pair(s) of basis functions to generate the maximum decrease in the model training error based on *RMSE*. the model with the lowest Generalized Cross-Validation statistic was selected.

Table 3(c) also lists the optimal MARS model equation. For greater details about ELM and MARS, readers can consult earlier References [88].

3.5 Predictive Model Performance Evaluation

The study adopts the model performance metrics recommended by American Society for Civil Engineers [123] to evaluate the hybrid CLSTM (and all the other benchmark) models. By appraising the degree of agreement between *PPFD*for and *PPFD*obs the computed metrics include correlation coefficient (*r*), mean absolute error (*MAE*, mmol m^{-2}s^{-1}), root mean square error (*RMSE*, mmol m^{-2}s^{-1}), including the relative % magnitudes of *RMSE* and *MAE*, Legate & McCabe's (*LM*) and the Nash Sutcliffe's coefficient (E_{NS}). Mathematically, these are as follows [43, 124-126]:

$$r = \left(\frac{\sum_{i=1}^{N}\left(PPFD_{for,i} - P\overline{PFD}_{obs,i}\right)\left(PPFD_{for,i} - P\overline{PFD}_{obs,i}\right)}{\sqrt{\sum_{i=1}^{N}\left(PPFD_{for,i} - P\overline{PFD}_{obs,i}\right)^{2}}\sqrt{\sum_{i=1}^{N}\left(PPFD_{for,i} - P\overline{PFD}_{obs,i}\right)^{2}}} \right) \qquad \text{<11>}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\left(PPFD_{for,i} - PPFD_{obs,i}\right)\right| \qquad \text{<12>}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(PPFD_{for,i} - PPFD_{obs,i}\right)^{2}} \qquad \text{<9>}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\left(PPFD_{for,i} - PPFD_{obs,i}\right)}{PPFD_{obs,i}}\right| \times 100 \qquad \text{<14>}$$

$$RRMSE = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(PPFD_{for,i} - PPFD_{obs,i}\right)^{2}}}{\frac{1}{N}\sum_{i=1}^{N}\left(PPFD_{obs,i}\right)} \times 100 \qquad \text{<15>}$$

$$LM = 1 - \left[\frac{\sum_{i=1}^{N}\left|PPFD_{obs,i} - PPFD_{for,i}\right|}{\sum_{i=1}^{N}\left|PPFD_{obs,i} - P\overline{PFD}_{obs,i}\right|}\right],\ 0 \le LM \le 1 \qquad \text{<16>}$$

$$E_{NS} = 1 - \left[\frac{\sum_{i=1}^{N}\left(PPFD_{obs,i} - PPFD_{for,i}\right)^{2}}{\sum_{i=1}^{N}\left(PPFD_{obs,i} - P\overline{PFD}_{obs,i}\right)^{2}}\right],\ -\infty \le E_{NS} \le 1 \qquad \text{<10>}$$

where $PPFD_{obs}$ and $PPFD_{for}$ are the observed and forecasted i^{th} value in test period, $\overline{PPFD}_{obs}$ and $\overline{PPFD}_{\text{for}}$ are the observed and forecasted means and N is the number of datum points within a test set.

The present study adopts several performance measures for a robust evaluation of the forecast models specially to overcome the constraints of any single metric. Diagnostic tools and graphical representations utilising scatterplots and error distribution are used in conjunction with statistical indices to test the versatility of 5-minute forecasts models.

4.0 Results and Discussion

In this section the results generated by the hybrid CLSTM predictive model, including the other deep learning-based (LSTM, CNN, DNN) and machine learning-based (ELM, MARS) models are appraised by checking the degree of congruence between measured and forecasted photosynthetic-active radiation at a 5-minute temporal scale. A careful evaluation of the results emanating from the cloud cover-based models using various input combinations (*i.e.*, Table 2) and a reference model utilising only the solar zenith angle is also made, to identify the contributory role of cloud variations in modelling photosynthetic photon flux density (*PPFD*). Figure 8 shows a scatterplot of the tested data where the performance of CLSTM (and comparative models) is evaluated in terms of the degree of agreement between observed and forecasted *PPFD*. Also included are the results of deep learning-based LSTM, CNN and DNN, as well as the other machine learning-based (MARS & ELM) model. Note that in here, only the optimally trained model (out of the 17 designated input combinations, Table 2) considering the influence of cloud variations on 5-minute *PPFD*, are shown.

While the performance of the newly proposed CLSTM model seems to exceed that of the other predictive models, as evidenced by the largest r^2 (~0.846), the gradient (representing the forecasted and observed *PPFD*) closest to unity, and the smallest bias constant, it also had a capped maximum forecasted *PPFD*. (Fig. 8a), the most accurate prediction differs significantly for the different model types and their input combinations. For example, the best performance of the CLSTM model (Fig. 8a) is attained through M_8: $PAR = f\{B_{av}, BC_{sd}, SZA, B_{sd}, OC, R_{av}, RC_{sd}, BC_{av}\}$. This means that the CLSTM model requires cloud segmented properties based on the whole sky blue average, standard deviation of the blue pixels, blue cloud average pixels, standard deviation of the blue cloud pixels, opaque cloud pixels, standard deviation of the red cloud pixels, whole sky red average pixels, and the *SZA* time series yielded the most accurate performance. For the case of the LSTM model (Fig. 8b), the best performance is attained through M_{13}: $PAR = f\left\{B_{av}, BC_{sd}, SZA, B_{sd}, OC, R_{av}, RC_{sd}, BC_{av}, \frac{R_{av}}{B_{av}}, R_{sd}, RBC_{diff}, BRC_{diff}, TC\right\}$ with this model using the eight input variables that are already used in CLSTM as well as the time series of $\frac{R_{av}}{B_{av}}, R_{sd}, RBC_{diff}, BRC_{diff}$ and TC to generate the best performance. A similar deduction is made for CNN, ELM and MARS models where the designated model M_{11}, M_{10} and M_{12} is seen to generate the highest coefficient of determination compared with a lower r^2 value for the other input combinations specified in Table 2. When the best input combination for the DNN model is deduced by progressively adding the cloud cover properties one by one, the model M_5 generates the best performance (r^2 = 0.810) with an input combination $PAR = f\{B_{av}, BC_{sd}, SZA, B_{sd}, OC\}$. Note that in this case, only five input series (*i.e.*, whole sky-blue average and standard deviation of whole sky blue including the standard deviation of blue cloud pixels, solar zenith angle, and opaque clouds) are required. However, it is also noteworthy that the performance of the DNN model is relatively lower than CLSTM model (*i.e.*, r^2 = 0.810 *vs*. 0.846). The analysis reveals that, while the hybrid CLSTM model integrating the LSTM and CNN methods used to emulate 5-minute *PPFD* far exceeds the performance of all other comparative models, their inputs combinations (based on cloud properties and *SZA*) appear to be unique indicating the different capabilities for feature extraction required to accurately predict the photosynthetic-active radiation.

<Figure 8>

In congruence with previous results shown in Figure 8, the frequency of the absolute value of predicted error distribution in the testing phase generated by the *optimal* CLSTM and the *optimal* benchmark models, are shown in Figure 9. It is notable the newly proposed CLSTM model (*i.e.*, M_8)

generated almost 75% of all predictive errors within the smallest error bracket *i.e*., ±200 μ mol of photons m^{-2} s^{-1} band compared with LSTM, M_{13} (~72%), DNN, M_5 (~69%), CNN, M_{11} (~69%), ELM, M_{10} (~71%) and MARS, M_{12} (~63%). The largest frequency of predictive errors within the smallest error bracket no doubt concurs with a smaller frequency of redistributed forecast errors, albeit within a larger error band exceeding ±200 μ mol $m^{-1}s^{-1}$. For example, we note that ~17% of all predictive errors attained by CLSTM are located within the ±(200–400) μ mol of photons m^{-2} s^{-1} whereas those for LSTM, DNN, CNN, ELM and MARS are seen to record ~21%, 22%, 21%, 20% and 27% of all predictive errors, respectively.

<Figure 9>

Next, we investigate the overall statistical score metrics computed over the last 7 days of tested data (*i.e*., 24-03-2013 to 31-03-2013) using 5-minute *PPFD*. Table 3 presents both the optimal model developed using various input combinations (M_1–M_{17}), as well as the reference model (M_{18}) developed using traditional approach (*i.e*., solar zenith angle only) as per earlier studies [39]. Interestingly, the best performance among all tested models is attained by different input combinations that use both the cloud cover properties and the solar zenith angle as an input variable. However, for the predictive models developed with only the solar zenith angle as an input, the performance of all the deep learning (CLSTM, CNN, DNN, LSTM) and machine learning (ELM, MARS) models appear to be significantly inferior to those that utilise cloud cover properties and *SZA*. In fact, the *SZA*-based models produce the smallest magnitude of *r* (between 0.796–0.623), and the largest *RMSE / MAE* between 412.77–438.99 / 354.29 – 368.09 μ mol of photons $m^{-2}s^{-1}$ within the testing phase. This contrasts the values *r* (0.894–0.920) and between 210.31–241.26 μ mol of photons $m^{-2}s^{-1}$ for *RMSE* and 150.24–183.11 μ mol of photons $m^{-2}s^{-1}$ for *MAE* for the models that incorporate cloud cover variations. This result indicates the important contributory role played by cloud cover variations in modulating the photosynthetic-active radiation and particularly, in improving the forecasting performance of the hybrid CLSTM and all of the other comparative models.

<Table 3>

In Table 3, we also present several metrics for models developed using cloud cover as well as the *SZA* data where the normalised performance metrics based on the relative percentage error, Nash Sutcliffe coefficient, and the Legates & McCabe's Index is incorporated. It is noteworthy that the inclusion of cloud cover properties is seen to lead to an improved performance of the hybrid CLSTM, and all the other predictive models. That is, we note the smaller error values ranging between 24.92–28.79% (*RRMSE*) and 38.01–56.21% (*RMAE*) for models utilising cloud cover properties, whereas the errors based on *SZA* as the only input variable are relatively larger, between 49.15–51.98% (*RRMSE*) and 128.39–176.72% (*RMAE*). It is therefore deducible that appropriate factoring of the role of cloud cover variations to predict 5-minute *PPFD* can help reduce the forecasted errors very significantly. This deduction also concurs with a much higher value of the Nash-Sutcliffe and the Legate's & McCabe's Index obtained for all models that are trained with cloud cover properties. If the performance of only the hybrid CLSTM model is evaluated against all the comparative models; after factoring the cloud cover properties, we register the values of E_{NS} and *LM* to be 0.846 and 0.679 compared with 0.796–0.829, and 0.607–0.660 for the case of ELM, LSTM, CNN, DNN and MARS models. Again, these metrics ascertain the influence of cloud cover on ground level photosynthetic-active radiation, and the superiority of the newly proposed CLSTM model.

<Figure 10>

Figure 10 is a Taylor diagram that evaluates all predictive models, including those with cloud cover properties and *SZA*-only as inputs. In this figure the most *optimal* model based on the best input combinations are compared to provide a visual framework for the forecasted *PPFD* against a reference (observed *PPFD*) data point. The pertinent statistics in Taylor diagram show the weighted centred pattern correlations and the ratio of the normalized root-mean-square (RMS) difference between the 'tested' data (*i.e.*, CLSTM, CNN, LSTM, DNN, ELM & MARS) and the 'reference' (observed) data. Two important deductions are made: *firstly*, it is clear that all of the *SZA*-based reference models are clustered much further away from the axis representing the observed *PPFD* whose *RMS*-centred difference certainly separates them away from the cloud cover-based models, and *secondly*, the CLSTM model utilising cloud properties (indicated in red) is at a closest location to the observed *PPFD*, and also attains the highest correlation among all tested predictive models. It is also observable that all the cloud cover-based models are within a smaller cluster (and hence, demonstrate comparable performance) whereas those utilising SZA only are more scattered. This suggests that the inclusion of cloud cover is necessary to optimise all the DL and ML models, but among all these models, the CLSTM remains the superior choice to forecast the 5-minute *PPDF* dataset.

<Figure 11>

In Figure 11, we investigate the nature of the predictive error generated by the objective model (*i.e.*, CLSTM) and the counterpart models while also evaluating the role of cloud cover variations using the modelled *PPDF* from the *SZA* only, and the cloud cover-based predictive models. Here, the forecast error $|\mathrm{FE}| = |PPFD_i^{for} - PPFD_i^{obs}|$ is illustrated as a boxplot for both the cloud property-based and the *SZA*-based model. There is a clear consensus that the best model out of the ones designated as M_1–M_{17} utilising cloud features as inputs are able to attain a significantly lower error distribution compared to the reference model M_{18} where *SZA* is the only predictor variable. For all predictive models trained with the *SZA* input data, the maximum error value is manyfold higher, and so is the upper quartile, median and the lower quartile of |*FE*|. This means that when cloud feature is excluded from a predictive model the ability to forecast *PPFD* values is much less, and this can result in a wider distribution of the errors for the *SZA*-based model. A comparison of all models developed using cloud cover properties, including the *SZA*, certainly shows a much smaller lower quartile, upper quartile, maximum and median values of the forecasted error. When all models trained with cloud features are investigated, the boxplots show the smallest value of 5-number summary, with the minimum, maximum, lower quartile, upper quartile and medians occupying smaller magnitudes for the case of CLSTM compared with CNN, LSTM, DNN, MARS and ELM. This is congruent with earlier results (Figs. 8–10) to demonstrate the CLSTM model as being the optimal choice to emulate the near real-time photosynthetic active radiation over a 5-minute scale.

<Figure 12>

To further establish the veracity of the hybrid CLSTM model Figure 12 shows the empirical cumulative distribution function (*ECDF*) of the error encountered in forecasting the photosynthetic-active radiation in the testing phase. The *ECDF* clearly demarcates the important role of cloud cover variations against the standard approach utilising *SZA* as the only input variable. A clear separation point is noted throughout the *ECDF* such that all models trained with cloud cover inputs attain a much smaller forecasted error with a steeper rising curve in contrast to the slower growth in *ECDF* within larger error values. In fact, the cloud-property based models reach an asymptotic state around an $|FE|$ value of 600 μ mol of photons m^{-2}s^{-1} whereas the *SZA*-based models continue to accumulate error values until $|FE|$ values of 900 μ mol of photons m^{-2}s^{-1}. Comparing the *ECDF*s of the hybrid CLSTM model against the other DL and ML models trained with cloud features, this result clearly concurs with Figure 9 where the growth in predictive errors is smaller for the CLSTM compared with the CNN, LSTM, DNN, ELM and MARS models. This establishes the efficacy of the newly developed CLSTM model trained with cloud cover features to generate the most accurate performance in terms of forecasting the 5-minute *PPFD* dataset.

<Figure 13>

We further explore the influence of cloud cover variations on the prescribed objective model (*i.e.*, CLSTM) in Figure 13 where the 5-minute forecasted *PPDF* valued averaged over the entire test dataset is shown *with* and *without* cloud cover input features. Note that these errors, showing both the percentage and absolute error values, are deduced from the forecasted and observed photosynthetic-active radiation measured from 07.00 AM to 05.00 PM. It is obvious that the hybrid CLSTM model utilising the cloud cover-based input features yields the smallest mean error over the whole diurnal cycle. The CLSTM error follow a temporal pattern where the models register relatively larger errors in early morning (~07.00 AM to 09.00 AM) and late afternoon (~04.00 PM to 05.00 PM) compared with the rest of the day. Possible causes for this error is that the CLSTM model did not isolate variability with solar zenith of clear sky aerosol optical thickness and cloud chromic properties associated with forward and backscattering at the cloud edges [127-131] [x7] or aerosol. It is also possible that the CLSTM model is unable to capture enough features to predict the relatively smaller *PPFD* values in the morning and afternoon where the aerosol optical thickness is similar to the cloud scattering. Nonetheless, this analysis clearly outlines the important role of cloud cover conditions in modelling photosynthetic-active radiation and shows an important advancement in photosynthetic-active radiation prediction compared to earlier studies using the traditional (*SZA*) method.

5.0 Further Discussion

The results generated by the proposed CLSTM model have established relationships between photosynthetic-active radiation and cloud cover conditions necessary to model near real-time 5-minute *PPFD* with this objective model exhibiting the best performance against several other competing (*i.e*., deep learning and machine learning-based) approaches. An incremental inclusion of cloud cover features based on time series of segmented cloud properties also captured a different, yet a significant contributory influence, further improving the testing performance of CLSTM model. However, improvements to the CLSTM model can be made with further development and refinement of the cloud segmentation tool itself.

The major contributions have led to significantly improved modelling approaches relative to earlier studies [132-135] where artificial intelligence models have utilised only the solar zenith angle, and failed to consider the effect of cloud cover conditions on photosynthetic-active radiation. Such methods used the more conventional modelling approaches (*i.e*., single hidden layer neuronal architecture) without any deep mining of the predictive features as undertaken by the proposed CLSTM method in this paper. Given that the movement of clouds is highly variable depending on altitude and wind, cloud shape and thickness commonly vary on timescales of much less than 30 minutes, our study has captured such influences on the ground-based photosynthetic active radiation at ~5-minutes. The modelling of photosynthetic radiation at this time interval is also of practical relevance in the monitoring and the supply of enough sunlight for solar energy generation or biofuels exploration, monitoring the healthy growth of plants, monitoring day light integral or available photosynthetic energy for plant functions.

This pilot study has demonstrated how the CLSTM model utilising statistical input features from cloud images can become a sophisticated deep learning system for the future development of solar energy monitoring devices [136]. One such technology that can be particularly useful in the agricultural sector (*i.e*., an automated monitoring and control system for algae photobioreactors) has practical relevance. For specific applications, CLSTM model can be incorporated into a smart environment monitoring system, 24 x 7, by adopting Internet of Things (IoT) and Wireless Sensor Networks, WSN [137] in a monitoring systems to ensure sustained health of crops and particularly considering how cloud conditions can affect their growth. The light available for microalgal photosynthesis remains a function of the surface solar irradiance over day-night cycles with environmental factors such as light, temperature, and nutrient status not only affecting photosynthesis and productivity of algae but also influencing the pattern, pathway and activities of cell metabolism or composition. Therefore, the efficacy of CLSTM model to forecast photosynthetic-active radiation at high temporal resolutions of 5-minutes that also matches a near real-time scale, can be trained on live cloud cover data or other atmospheric conditions. This application of the proposed deep learning system can help in regular prediction of the availability of sunlight in real time including its role in modelling temperature, water salinity, or nutrient status within an algae pond. The CLSTM model can also be employed in biophysical model platforms to improve the robustness of plant-growth models particularly, providing accurate estimations of photosynthetic photon flux density due to the scarcity of their ground-based measurements [138]. As the cost of Total Sky Imagers (TSIs) can be insurmountable for most solar energy or biofuel generation farm locations, geo-stationary satellites such Himawari 8 or 9, operating at roughly 10-minute interval and relatively high spatial resolutions may become good suppliers of sky images to be used as inputs for the CLSTM model to generate predicted PPFD or other components of solar radiation at appropriate temporal resolutions.

Other than agricultural applications, our CLSTM model incorporating cloud conditions also has potential use in public health and energy sectors. In an earlier study, Deo *et al.*, [88] developed a very short-term reactive system for solar ultraviolet (UV) prediction, albeit using a single hidden layer extreme learning machine (ELM) model and without any consideration to cloud cover conditions. Such a UV forecasting system can be a useful avenue for real-time prediction of UV radiation, a component of the solar spectrum known to cause melanoma and eye disease. However, as neither that study, nor any other prior or following study has incorporated the role of cloud cover conditions into a solar UV forecasting system, the proposed CLSTM system built on deep learning technology might be a viable tool to test the role of cloud conditions on UV prediction. One may therefore develop a CLSTM system for short-term (*e.g.*, 5-minute) reactive forecasting of UV index to help in public health risk mitigation. In terms of its application in energy industries, the CLSTM model can become a viable tool for real-time management of solar energy in a photovoltaic system by responding through a cloud image-based forecast system for solar power prediction, and particularly utilising cloud movements, cloud forms or its relative position-based features. Such a sky image-based solar power forecasting system utilising deep data mining can be of great value to the solar energy industry [40].

6.0 Conclusions

The industrial-scale production of solar power, biofuels and agriculture including food and health supplements from micro-algae farming, require reliably predicted solar radiation over short, long, and medium-term periods. This study has established the feasibility of predicting very short-term, 5-minute interval photosynthetic-active radiation using segmented cloud cover properties and solar zenith angle in a sub-tropical region in Toowoomba, Australia. A total of 17 different segmented cloud cover properties based on the mean, standard deviation, differences, and ratios of blue and red pixel values in clouds, including opaque and thin clouds (applied through thresholds on the Total Sky Imager), were acquired as part of the University of Southern Queensland Solar Radiation Monitoring Program running for more than 15 years. Together with the solar zenith angle, the cloud cover properties based on segmented image inputs were applied to develop the hybrid deep learning (*i.e.*, CLSTM) model based on an integration of convolutional neural networks (to map out the cloud and SZA-based input features) and the long short-term memory network (to generate the near real-time forecasts of 5-minute photosynthetic photon flux density, *PPFD*). The CLSTM, verified to be highly superior in predicting 5-minute *PPFD* through 17 different predictor variable (or input) combinations, was benchmarked against three deep learning methods (*i.e.*, LSTM, CNN, DNN) and two machine learning (*i.e.*, ELM & MARS) methods. All these predictive models were evaluated using statistical score metrics and diagnostic plots visualising the degree of agreement between forecasted and observed photosynthetic photon flux density in an independent test dataset where the CLSTM model was applied.

The findings can be enumerated as follows.

(i) Among the objective (CLSTM) and five competing models, the best performance (out of 17 distinct input combinations of segmented cloud properties) was attained by different combinations of cloud features. For example, the best CLSTM model M_8 utilised average of whole sky-blue pixels, standard deviation of blue cloud pixels, *SZA*, standard deviation of the whole sky blue pixels, opaque clouds, averaged whole sky red pixels, standard deviation of red cloud pixels and the average of blue cloud pixels. By contrast, the second-best model (*i.e.*, ELM) used all the 8 inputs required by CLSTM, including two additional inputs (*i.e.*, ratio of whole sky blue to whole sky red average cloud pixels and whole sky red standard deviation) for its optimal model M_{10}. The third-best model, or LSTM required three additional inputs compared with ELM. The CNN model, which was the fourth-best model developed to forecast 5-minute *PPFD* used only 11 input variables, whereas the DNN model relied on only 5 input variables. Despite different numbers of inputs used by the hybridised, deep learning and machine learning models, the performance of CLSTM remained superior.

(ii) In terms of comparing the *SZA*-only models, the CLSTM without cloud registered twice the model error (~50.07%) compared to with cloud ~24.92% in the testing phase. The other metrics for SZA models only were also far less impressive for all models then those where clouds were incorporated. In terms of Taylor diagram comparing the different models to a reference (*i.e.*, observation) point, the non-cloud cover-based models were certainly scattered much further away from this reference point, and their performances were quite disparate relative to a comparable performance for cloud cover-based models (Fig. 10). Likewise, the distribution of forecast error was more widely spread, with significantly larger outliers, upper quartile, or extreme error values for *SZA*-only models (Figs. 11–12). These finding ascertain the important role of considering cloud cover variations to accurately model photosynthetic-active radiation.

Finally, this pilot study highlights the appropriateness of using cloud cover features to develop a deep learning method for very short-term, near real-time forecasting of photosynthetic-active radiation. If cloud segmented image properties from geo-stationary satellites images are available, the need for ground-based inputs that are data expensive for many regional locations can be eliminated. Furthermore, fish-eye lens or adapters used in mobile phones may also be able to supply the relevant images so the developed CLSTM model can be tried with those inputs to make the predictive model more accessible and applicable to all regions where the segmentation software is made available. This newly proposed method can offer major advantages in terms of the model implementation in regions with limited access to data such as agricultural farms. However, the present study only considers cloud properties using local, two-dimensional ground-based sky images so the inclusion of other atmospheric attenuations imposed by water vapour and aerosol should also be considered in the proposed CLSTM model with performance tested in different climatic zones and seasons. The improvement in CLSTM model's practical viability for other regions may also be made through its implementation on hourly, daily, and seasonal scales by sourcing satellite and other remote sensing products. Such testing of the proposed CLSTM predictive model in a wider range of climates, in both remote and regional locations is a necessary step to help in direct harnessing of solar energy, biofuels from microalgae, agricultural crop monitoring and supporting bio-physical sectors where photosynthetic-active radiation needs to be monitored.

Acknowledgements

Data acquired were obtained from University of Southern Queensland Solar Research Laboratory. The authors are grateful to Mr Kai Chen for some insightful discussions.

References

1. McCree, K., *The measurement of photosynthetically active radiation.* Solar energy, 1973. **15**(1): p. 83-87.
2. Proskurina, S., et al., *Global biomass trade for energy— Part 2: Production and trade streams of wood pellets, liquid biofuels, charcoal, industrial roundwood and emerging energy biomass.* Biofuels, Bioproducts and Biorefining, 2019. **13**(2): p. 371-387.
3. Vuppaladadiyam, A.K., et al., *Microalgae cultivation and metabolites production: a comprehensive review.* Biofuels, Bioproducts and Biorefining, 2018. **12**(2): p. 304-324.
4. Ramanna, L., I. Rawat, and F. Bux, *Light enhancement strategies improve microalgal biomass productivity.* Renewable and Sustainable Energy Reviews, 2017. **80**: p. 765-773.
5. Holdmann, C., U. Schmid-Staiger, and T. Hirth, *Outdoor microalgae cultivation at different biomass concentrations — Assessment of different daily and seasonal light scenarios by modeling.* Algal Research, 2019. **38**: p. 101405.
6. Slade, R. and A. Bauen, *Micro-algae cultivation for biofuels: Cost, energy balance, environmental impacts and future prospects.* Biomass and Bioenergy, 2013. **53**: p. 29-38.
7. Chen, C.-Y., et al., *Cultivation, photobioreactor design and harvesting of microalgae for biodiesel production: A critical review.* Bioresource Technology, 2011. **102**(1): p. 71-81.
8. Kumar, M., et al., *Rapid and efficient genetic transformation of the green microalga Chlorella vulgaris.* Journal of Applied Phycology, 2018. **30**(3): p. 1735-1745.
9. Park, S., T.H.T. Nguyen, and E. Jin, *Improving lipid production by strain development in microalgae: Strategies, challenges and perspectives.* Bioresource Technology, 2019. **292**: p. 121953.
10. Zhang, Y., et al., *Genetic Transformation of Tribonema minus, a Eukaryotic Filamentous Oleaginous Yellow-Green Alga.* International Journal of Molecular Sciences, 2020. **21**(6): p. 2106.
11. Pruvost, J., et al., *Microalgae culture in building-integrated photobioreactors: Biomass production modelling and energetic analysis.* Chemical Engineering Journal, 2016. **284**: p. 850-861.
12. Siqueira, S.F., et al., *Mapping the performance of photobioreactors for microalgae cultivation: geographic position and local climate.* Journal of Chemical Technology & Biotechnology, 2020. **95**(9): p. 2411-2420.
13. Grant, R.H. and G.M.J.J.o.A.M. Heisler, *Obscured overcast sky radiance distributions for ultraviolet and photosynthetically active radiation.* 1997. **36**(10): p. 1336-1345.
14. Negi, S., et al., *Impact of nitrogen limitation on biomass, photosynthesis, and lipid accumulation in Chlorella sorokiniana.* Journal of Applied Phycology, 2016. **28**(2): p. 803-812.
15. Patil, S., R. Pandit, and A. Lali, *Responses of algae to high light exposure: prerequisite for species selection for outdoor cultivation.* 2017. **8**: p. 75-83.
16. Hanan, N., et al., *Estimation of absorbed photosynthetically active radiation and vegetation net production efficiency using satellite data.* 1995. **76**(3-4): p. 259-276.
17. Gumma, M.K., et al., *Agricultural cropland extent and areas of South Asia derived using Landsat satellite 30-m time-series big-data using random forest machine learning algorithms on the Google Earth Engine cloud.* GIScience & Remote Sensing, 2020. **57**(3): p. 302-322.
18. Zheng, Y., M. Zhang, and B. Wu. *Using high spatial and temporal resolution data blended from SPOT-5 and MODIS to map biomass of summer maize*. in *2016 Fifth International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*. 2016.

19. Zheng, Y., et al., *Mapping Winter Wheat Biomass and Yield Using Time Series Data Blended from PROBA-V 100- and 300-m S1 Products.* Remote Sensing, 2016. **8**(10): p. 824.

20. Batey, M. and R.J.A.r. Green, *Geometrically effective cloud fraction for solar radiation.* 2000. **55**(2): p. 115-129.

21. Hengl, T., et al., *Global mapping of potential natural vegetation: an assessment of Machine Learning algorithms for estimating land potential.* PeerJ Preprints, 2018. **6**: p. e26811v2.

22. Tang, W., et al., *An efficient algorithm for calculating photosynthetically active radiation with MODIS products.* Remote Sensing of Environment, 2017. **194**: p. 146-154.

23. Rocha, A.V., et al., *Solar position confounds the relationship between ecosystem function and vegetation indices derived from solar and photosynthetically active radiation fluxes.* Agricultural and Forest Meteorology, 2021. **298-299**: p. 108291.

24. Lozano, I.L., et al., *Aerosol radiative effects in photosynthetically active radiation and total irradiance at a Mediterranean site from an 11-year database.* Atmospheric Research, 2021. **255**: p. 105538.

25. Chen, L., et al., *MODIS-derived daily PAR simulation from cloud-free images and its validation.* Solar Energy, 2008. **82**(6): p. 528-534.

26. Grant, R., et al., *Ability to predict daily solar radiation values from interpolated climate records for use in crop simulation models.* 2004. **127**(1-2): p. 65-75.

27. Rao, C.N., *Photosynthetically active components of global solar radiation: measurements and model computations.* Archives for meteorology, geophysics, bioclimatology, Series B, 1984. **34**(4): p. 353-364.

28. Ali, M., et al., *An ensemble-ANFIS based uncertainty assessment model for forecasting multi-scalar standardized precipitation index.* Atmospheric Research, 2018. **207**: p. 155-180.

29. Ali, M., et al., *Multi-stage committee based extreme learning machine model incorporating the influence of climate parameters and seasonality on drought forecasting.* Computers and Electronics in Agriculture, 2018. **152**: p. 149-165.

30. Han, J., et al., *Prediction of Winter Wheat Yield Based on Multi-Source Data and Machine Learning in China.* Remote Sensing, 2020. **12**(2): p. 236.

31. Crane-Droesch, A., *Machine learning methods for crop yield prediction and climate change impact assessment in agriculture.* Environmental Research Letters, 2018. **13**.

32. Cai, Y., et al., *Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches.* Agricultural and Forest Meteorology, 2019. **274**: p. 144-159.

33. Kamir, E., F. Waldner, and Z. Hochman, *Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods.* ISPRS Journal of Photogrammetry and Remote Sensing, 2020. **160**: p. 124-135.

34. Feng, P., et al., *Incorporating machine learning with biophysical model can improve the evaluation of climate extremes impacts on wheat yield in south-eastern Australia.* Agricultural and Forest Meteorology, 2019. **275**: p. 100-113.

35. Wagner, V.s., *Uebertragung strahlungsreleveanter wetterinformation aus punktuellen PAR-sensordaten in groesser versuchsfaechenanlagen mit hifle hemisphaerisher fotos.* Allg. Forst-u. J.-Ztg, 1995. **167(1-2)**: p. 34-40.

36. Ryu, Y., et al., *MODIS-derived global land products of shortwave radiation and diffuse and total photosynthetically active radiation at 5km resolution from 2000.* Remote Sensing of Environment, 2018. **204**: p. 812-825.

37. Gu, L., et al., *Advantages of diffuse radiation for terrestrial ecosystem productivity.* 2002. **107**(D6): p. ACL 2-1-ACL 2-23.

38. Jiang, H., et al., *Surface Diffuse Solar Radiation Determined by Reanalysis and Satellite over East Asia: Evaluation and Comparison.* Remote Sensing, 2020. **12**(9): p. 1387.

39. Deo, R.C., et al., *Adaptive Neuro-Fuzzy Inference System integrated with solar zenith angle for forecasting sub-tropical Photosynthetically Active Radiation.* Food and Energy Security, 2019. **8**(1): p. e00151.

40. Zhen, Z., et al., *Research on a cloud image forecasting approach for solar power forecasting.* 2017. **142**: p. 362-368.

41. Deo, R.C., et al., *Very short-term reactive forecasting of the solar ultraviolet index using an extreme learning machine integrated with the solar zenith angle.* Environmental Research, 2017. **155**: p. 141-166.

42. Igoe, D.P., A.V. Parisi, and N.J. Downs, *Cloud segmentation property extraction from total sky image repositories using Python.* Instrumentation Science & Technology, 2019. **47**(5): p. 522-534.

43. Ghimire, S., et al., *Deep solar radiation forecasting with convolutional neural network and long short-term memory network algorithms.* Applied Energy, 2019. **253**: p. 113541.

44. Al-Musaylh, M.S., R.C. Deo, and Y. Li, *Electrical Energy Demand Forecasting Model Development and Evaluation with Maximum Overlap Discrete Wavelet Transform-Online Sequential Extreme Learning Machines Algorithms.* Energies, 2020. **13**(9): p. 2307.

45. Al-Musaylh, M.S., et al., *Short-term electricity demand forecasting with MARS, SVR and ARIMA models using aggregated demand data in Queensland, Australia.* Advanced Engineering Informatics, 2018. **35**: p. 1-16.

46. Wang, K., X. Qi, and H. Liu, *A comparison of day-ahead photovoltaic power forecasting models based on deep learning neural network.* Applied Energy, 2019. **251**: p. 113315.

47. Al-Musaylh, M.S., R.C. Deo, and Y. Li. *Particle Swarm Optimized–Support Vector Regression Hybrid Model for Daily Horizon Electricity Demand Forecasting Using Climate Dataset*. in *E3S Web of Conferences*. 2018. EDP Sciences.

48. Chen, J., et al., *Wind speed forecasting using nonlinear-learning ensemble of deep learning time series prediction and extremal optimization.* Energy Conversion and Management, 2018. **165**: p. 681-695.

49. Wang, J., et al. *Dimensional sentiment analysis using a regional CNN-LSTM model*. in *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*. 2016.

50. Sainath, T.N., et al. *Convolutional, long short-term memory, fully connected deep neural networks*. in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. 2015. IEEE.

51. Ullah, A., et al., *Action recognition in video sequences using deep bi-directional LSTM with CNN features.* IEEE access, 2017. **6**: p. 1155-1166.

52. Oh, S.L., et al., *Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats.* Computers in biology and medicine, 2018. **102**: p. 278-287.

53. Zhao, R., et al., *Learning to monitor machine health with convolutional bi-directional LSTM networks.* Sensors, 2017. **17**(2): p. 273.

54. Kim, T.-Y. and S.-B. Cho, *Predicting residential energy consumption using CNN-LSTM neural networks.* Energy, 2019. **182**: p. 72-81.

55. Ullah, F.U.M., et al., *Short-term prediction of residential power energy consumption via CNN and multi-layer bi-directional LSTM networks.* IEEE Access, 2019. **8**: p. 123369-123380.

56. Wang, F., et al., *Wavelet decomposition and convolutional LSTM networks based improved deep learning model for solar irradiance forecasting.* applied sciences, 2018. **8**(8): p. 1286.

57. Gao, B., et al., *Hourly forecasting of solar irradiance based on CEEMDAN and multi-strategy CNN-LSTM neural networks.* Renewable Energy, 2020. **162**: p. 1665-1683.

58. Lee, W., et al., *Forecasting solar power using long-short term memory and convolutional neural networks.* IEEE Access, 2018. **6**: p. 73068-73080.

59. Jaseena, K.U. and B.C. Kovoor, *Decomposition-based hybrid wind speed forecasting model using deep bidirectional LSTM networks.* Energy Conversion and Management, 2021. **234**: p. 113944.

60. Hong, Y.-Y. and T.R.A. Satriani, *Day-ahead spatiotemporal wind speed forecasting using robust design-based deep learning neural network.* Energy, 2020. **209**: p. 118441.

61. Meka, R., A. Alaeddini, and K. Bhaganagar, *A robust deep learning framework for short-term wind power forecast of a full-scale wind farm using atmospheric variables.* Energy, 2021. **221**: p. 119759.

62. Vidal, A. and W. Kristjanpoller, *Gold volatility prediction using a CNN-LSTM approach.* Expert Systems with Applications, 2020. **157**: p. 113481.

63. Yadav, A., C.K. Jha, and A. Sharan, *Optimizing LSTM for time series prediction in Indian stock market.* Procedia Computer Science, 2020. **167**: p. 2091-2100.

64. Kuo, C.-C.J., *Understanding convolutional neural networks with a mathematical model.* Journal of Visual Communication and Image Representation, 2016. **41**: p. 406-413.

65. Chimmula, V.K.R. and L. Zhang, *Time series forecasting of COVID-19 transmission in Canada using LSTM networks.* Chaos, Solitons & Fractals, 2020. **135**: p. 109864.

66. Song, X., et al., *Time-series well performance prediction based on Long Short-Term Memory (LSTM) neural network model.* Journal of Petroleum Science and Engineering, 2020. **186**: p. 106682.

67. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* nature, 2015. **521**(7553): p. 436-444.

68. Li, T., M. Hua, and X. Wu, *A hybrid CNN-LSTM model for forecasting particulate matter (PM2. 5).* IEEE Access, 2020. **8**: p. 26933-26940.

69. Xie, H., L. Zhang, and C.P. Lim, *Evolving CNN-LSTM Models for Time Series Prediction Using Enhanced Grey Wolf Optimizer.* IEEE Access, 2020. **8**: p. 161519-161541.

70. Ma, L. and S. Tian, *A Hybrid CNN-LSTM Model for Aircraft 4D Trajectory Prediction.* IEEE Access, 2020. **8**: p. 134668-134680.

71. Barzegar, R., M.T. Aalami, and J. Adamowski, *Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model.* Stochastic Environmental Research and Risk Assessment, 2020: p. 1-19.

72. Zang, H., et al., *Short-term global horizontal irradiance forecasting based on a hybrid CNN-LSTM model with spatiotemporal correlations.* Renewable Energy, 2020. **160**: p. 26-41.

73. Bengio, Y., P. Simard, and P. Frasconi, *Learning long-term dependencies with gradient descent is difficult.* IEEE transactions on neural networks, 1994. **5**(2): p. 157-166.

74. Hochreiter, S. and J. Schmidhuber, *Long short-term memory.* Neural computation, 1997. **9**(8): p. 1735-1780.

75. Graves, A., *Generating sequences with recurrent neural networks.* arXiv preprint arXiv:1308.0850, 2013.

76. Wu, Q. and H. Lin, *Daily urban air quality index forecasting based on variational mode decomposition, sample entropy and LSTM neural network.* Sustainable Cities and Society, 2019. **50**: p. 101657.

77. Sabburg, J.M., *Quantification of cloud around the sun and its correlation with global UV measurement*. 2000, Queensland University of Technology.

78. Gill, D., T. Ming, and W. Ouyang, *Improving the Lake Erie HAB Tracker: A Forecasting & Decision Support Tool for Harmful Algal Blooms*. 2017.

79. Johnson, D., et al. *A New Quantum Sensor for Measuring Photosynthetically Active Radiation*. in *AGU Fall Meeting Abstracts*. 2015.

80. Ghonima, M., et al., *A method for cloud detection and opacity classification based on ground based sky imagery.* Atmospheric Measurement Techniques, 2012. **5**(11): p. 2881-2892.

81. Dev, S., et al., *Rough-set-based color channel selection.* IEEE Geoscience and remote sensing letters, 2016. **14**(1): p. 52-56.

82. Sabburg, J. and J. Wong, *Evaluation of a Ground-Based Sky Camera System for Use inSurface Irradiance Measurement.* Journal of Atmospheric and Oceanic Technology, 1999. **16**(6): p. 752-759.

83. Li, Q., W. Lu, and J. Yang, *A hybrid thresholding algorithm for cloud detection on ground-based color images.* Journal of atmospheric and oceanic technology, 2011. **28**(10): p. 1286-1296.

84. Liu, M., J. Zhang, and X. Xia, *Evaluation of multiple surface irradiance-based clear sky detection methods at Xianghe—A heavy polluted site on the North China Plain: 香河站太阳辐射识别晴空方法的评估.* Atmospheric and Oceanic Science Letters, 2021. **14**(2): p. 100016.

85. A. Jebar, M.A., et al., *Influence of clouds on OMI satellite total daily UVA exposure over a 12-year period at a southern hemisphere site.* International Journal of Remote Sensing, 2020. **41**(1): p. 272-283.

86. Sabburg, J. and C.N. Long, *Improved sky imaging for studies of enhanced UV irradiance.* Atmospheric Chemistry and Physics, 2004. **4**(11/12): p. 2543-2552.

87. Parisi, A.V., J. Sabburg, and M.G. Kimlin, *Scattered and filtered solar UV measurements*. Vol. 17. 2004: Springer Science & Business Media.

88. Deo, R.C., et al., *Very short-term reactive forecasting of the solar ultraviolet index using an extreme learning machine integrated with the solar zenith angle.* Environmental research, 2017. **155**: p. 141-166.

89. Long, C.N., et al., *Retrieving cloud characteristics from ground-based daytime color all-sky images.* Journal of Atmospheric and Oceanic Technology, 2006. **23**(5): p. 633-652.

90. Slater, D., C. Long, and T. Tooman. *Total sky imager/whole sky imager cloud fraction comparison*. in *Eleventh ARM Science Team Meeting Proceedings, Atlanta, Georgia*. 2001.

91. Van Der Walt, S., S.C. Colbert, and G. Varoquaux, *The NumPy array: a structure for efficient numerical computation.* Computing in science & engineering, 2011. **13**(2): p. 22-30.

92. van der Walt, S., *Schö nberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, Gouillart E, Yu T, scikitimage contributors. 2014. scikit-image: image processing in python.* PeerJ. **2**: p. e453.

93. Konasani, V.R. and S. Kadre, *Machine Learning and Deep Learning Using Python and TensorFlow*. 2021, McGraw-Hill Education.

94. Moler, C., *Matlab incorporates LAPACK.* Cleve's Corner, MATLAB News&Notes, 2000.

95. Deo, R.C., X. Wen, and Q. Feng, *A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset.* Applied Energy, 2016. **168**: p. 568–593.

96. Michalsky, J.J., *The astronomical almanac's algorithm for approximate solar position (1950–2050).* Solar energy, 1988. **40**(3): p. 227-235.

97. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* Journal of machine learning research, 2011. **12**(Oct): p. 2825-2830.

98. Chollet, F., *Keras: The python deep learning library.* Astrophysics Source Code Library, 2018.

99. Ketkar, N., *Introduction to keras*, in *Deep Learning with Python*. 2017, Springer. p. 97-111.

100. Nwankpa, C., et al., *Activation functions: Comparison of trends in practice and research for deep learning.* arXiv preprint arXiv:1811.03378, 2018.

101. Hohman, F., et al., *S ummit: Scaling deep learning interpretability by visualizing activation and attribution summarizations.* IEEE transactions on visualization and computer graphics, 2019. **26**(1): p. 1096-1106.

102. Agarap, A.F., *Deep learning using rectified linear units (relu).* arXiv preprint arXiv:1803.08375, 2018.

103. Garbin, C., X. Zhu, and O. Marques, *Dropout vs. batch normalization: an empirical study of their impact to deep learning.* Multimedia Tools and Applications, 2020: p. 1-39.

104. Cai, S., et al., *Effective and efficient dropout for deep convolutional neural networks.* arXiv preprint arXiv:1904.03392, 2019.

105. Zhang, Q., et al., *An adaptive dropout deep computation model for industrial IoT big data learning with crowdsourcing to cloud computing.* IEEE Transactions on Industrial Informatics, 2018. **15**(4): p. 2330-2337.

106. Lambert, J., O. Sener, and S. Savarese. *Deep learning under privileged information using heteroscedastic dropout*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

107. Sato, M., et al., *Application of deep learning to the classification of images from colposcopy.* Oncology letters, 2018. **15**(3): p. 3518-3523.

108. Antczak, K., *On regularization properties of artificial datasets for deep learning.* arXiv preprint arXiv:1908.07005, 2019.

109. Ayinde, B.O. and J.M. Zurada, *Deep learning of constrained autoencoders for enhanced understanding of data.* IEEE transactions on neural networks and learning systems, 2017. **29**(9): p. 3969-3979.

110. Byrd, J. and Z. Lipton. *What is the effect of importance weighting in deep learning?* in *International Conference on Machine Learning*. 2019. PMLR.

111. Jaiswal, S., A. Mehta, and G. Nandi. *Investigation on the Effect of L1 an L2 Regularization on Image Features Extracted Using Restricted Boltzmann Machine*. in *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*. 2018. IEEE.

112. Chollet, F., *Keras (2015)*. 2017.

113. Rice, L., E. Wong, and Z. Kolter. *Overfitting in adversarially robust deep learning*. in *International Conference on Machine Learning*. 2020. PMLR.

114. Li, M., M. Soltanolkotabi, and S. Oymak. *Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks*. in *International Conference on Artificial Intelligence and Statistics*. 2020. PMLR.

115. Zhang, Y.-D., et al., *Voxelwise detection of cerebral microbleed in CADASIL patients by leaky rectified linear unit and early stopping.* Multimedia Tools and Applications, 2018. **77**(17): p. 21825-21845.

116. Dodge, J., et al., *Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping.* arXiv preprint arXiv:2002.06305, 2020.

117. Mahsereci, M., et al., *Early stopping without a validation set.* arXiv preprint arXiv:1703.09580, 2017.

118. Zhang, X., et al., *Template-oriented synthesis of monodispersed SnS2@SnO2 hetero-nanoflowers for Cr(VI) photoreduction.* Applied Catalysis B: Environmental, 2016. **192**: p. 17-25.

119. Swietojanski, P., A. Ghoshal, and S. Renals, *Convolutional Neural Networks for Distant Speech Recognition.* IEEE Signal Processing Letters, 2014. **21**(9): p. 1120-1124.

120. Jekabsons, G., *Adaptive Regression Splines toolbox for Matlab/Octave.* Version, 2013. **1**: p. 72.

121. Kooperberg, C. and D.B. Clarkson, *Hazard regression with interval-censored data.* Biometrics, 1997: p. 1485-1494.

122. Zareipour, H., K. Bhattacharya, and C. Canizares. *Forecasting the hourly Ontario energy price by multivariate adaptive regression splines*. in *2006 IEEE Power Engineering Society General Meeting*. 2006. IEEE.

123. ASCE, T.C., *Criteria for evaluation of watershed models.* Journal of Irrigation and Drainage Engineering, 1993. **119**(3): p. 429-442.

124. Ghimire, S., et al., *Self-adaptive differential evolutionary extreme learning machines for long-term solar radiation prediction with remotely-sensed MODIS satellite and Reanalysis atmospheric products in solar-rich cities.* Remote Sensing of Environment, 2018. **212**: p. 176-198.

125. Ghimire, S., et al., *Wavelet-based 3-phase hybrid SVR model trained with satellite-derived predictors, particle swarm optimization and maximum overlap discrete wavelet transform for solar radiation prediction.* Renewable and Sustainable Energy Reviews, 2019. **113**: p. 109247.

126. Ghimire, S., et al., *Global solar radiation prediction by ANN integrated with European Centre for medium range weather forecast fields in solar rich cities of Queensland Australia.* Journal of Cleaner Production, 2019. **216**: p. 288-310.

127. Robinson, P.J.J.o.A.M. and Climatology, *Measurements of downward scattered solar radiation from isolated cumulus clouds.* 1977. **16**(6): p. 620-625.

128. Segal, M. and J. Davis, *The impact of deep cumulus reflection on the ground-level global irradiance.* Journal of Applied Meteorology, 1992. **31**(2): p. 217-222.

129. Aida, M., *Scattering of solar radiation as a function of cloud dimensions and orientation.* Journal of Quantitative Spectroscopy, 1977. **17**(3): p. 303-310.

130. Liou, K.-N., *On the absorption, reflection and transmission of solar radiation in cloudy atmospheres.* Journal of Atmospheric sciences, 1976. **33**(5): p. 798-805.

131. González, J. and J. Calbó, *Modelled and measured ratio of PAR to global radiation under cloudless skies.* Journal of Agricultural Forest Meteorology, 2002. **110**(4): p. 319-325.

132. Deo, R.C., et al., *Adaptive Neuro-Fuzzy Inference System integrated with solar zenith angle for forecasting sub-tropical Photosynthetically Active Radiation.* Food and Energy Security, 2019. **8**(1): p. e00151.

133. Lopez, G., et al., *Estimation of hourly global photosynthetically active radiation using artificial neural network models.* Agricultural and forest Meteorology, 2001. **107**(4): p. 279-291.

134. Pankaew, P., et al. *Estimating photosynthetically active radiation using an artificial neural network*. in *2014 International Conference and Utility Exhibition on Green Energy for Sustainable Development (ICUE)*. 2014. IEEE.

135. Yu, X. and X. Guo, *Hourly photosynthetically active radiation estimation in Midwestern United States from artificial neural networks and conventional regressions models.* International journal of biometeorology, 2016. **60**(8): p. 1247-1259.

136. Wang, L., et al., *Modeling and comparison of hourly photosynthetically active radiation in different ecosystems.* Renewable and Sustainable Energy Reviews, 2016. **56**: p. 436-453.

137. Ullo, S.L. and G.J.S. Sinha, *Advances in smart environment monitoring systems using iot and sensors.* 2020. **20**(11): p. 3113.

138. García-Rodríguez, A., et al., *Photosynthetic Active Radiation, Solar Irradiance and the CIE Standard Sky Classification.* Applied Sciences, 2020. **10**(22): p. 8007.

Figures

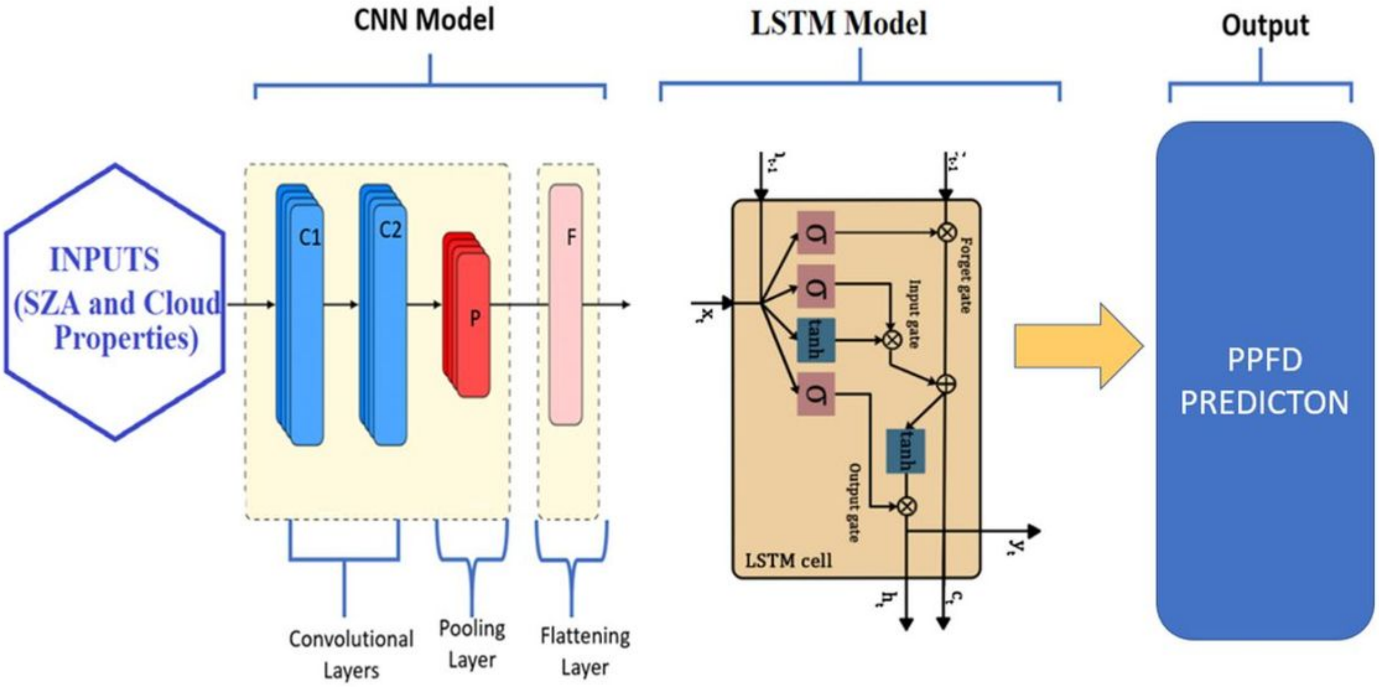


Figure 1

Schematic illustration of Convolutional Neural Network-Long Short-Term Memory Network (CLSTM) predictive framework. CNN used for feature extraction from solar zenith angle (SZA) and cloud chromatic properties from Total Sky Imager (TSI) and LSTM is used for time sequential modelling of the photosynthetic-active radiation (represented as photosynthetic photon flux density, PPFD).

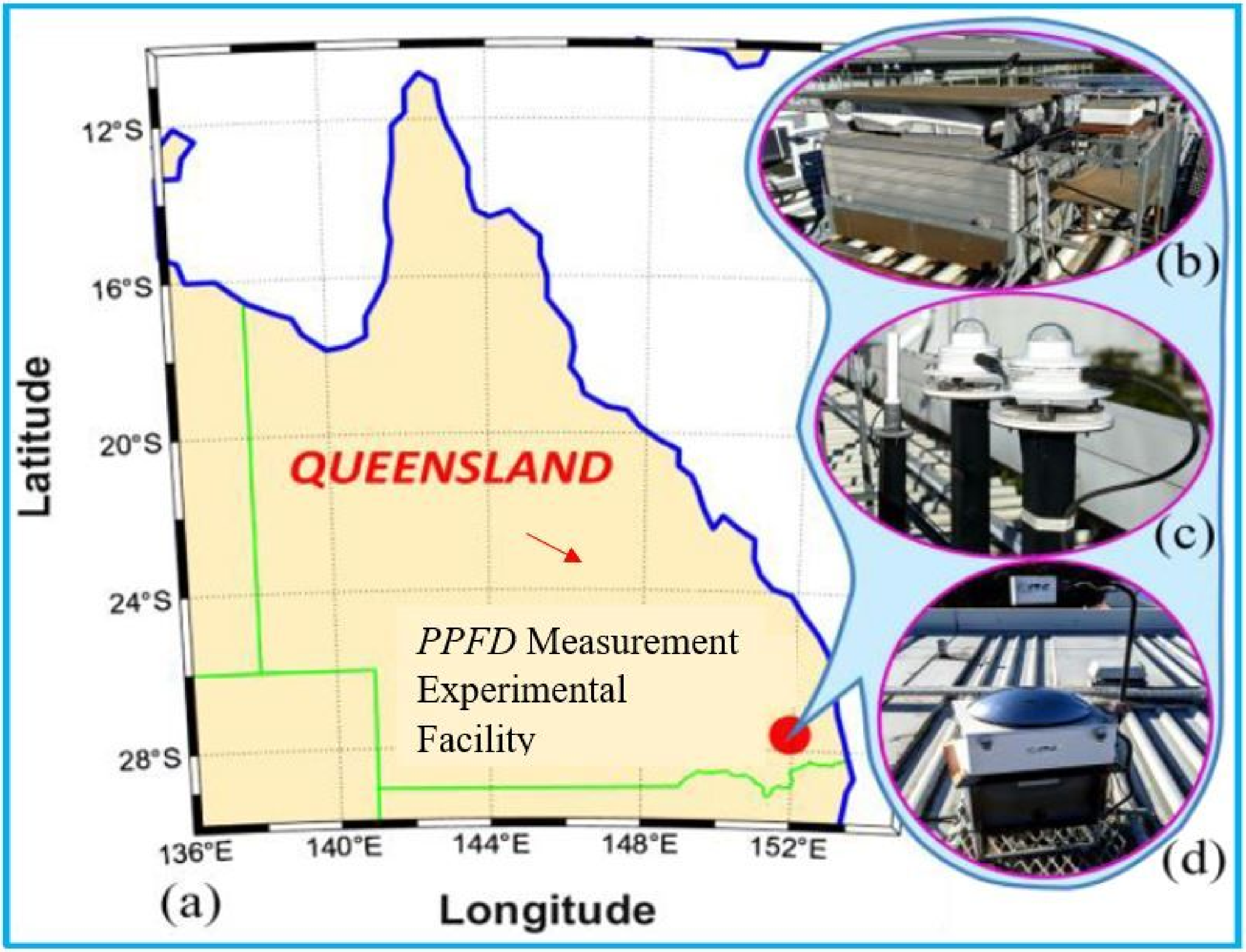
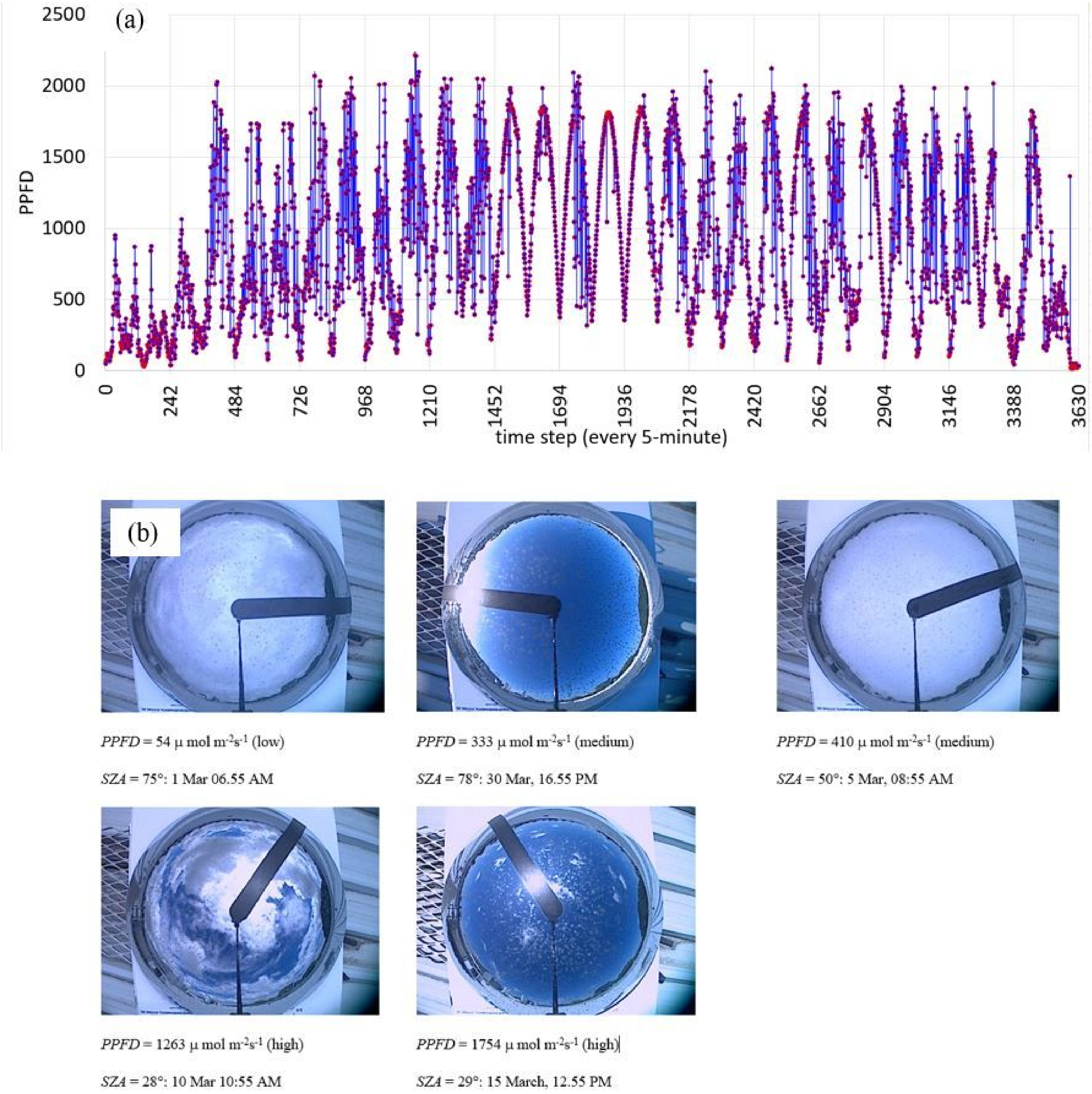


Figure 2

(a) Geographic location of the measurement facility in Queensland, Australia where CLSTM model is implemented. (b) Roof-top mounted Bentham DTM300 Spectroradiometer for 5-minute PPFD (μ mol of photons m-2 s-1) measurement. (c) Co-located 501 broadband UVR Biometer. (d) Synchronous Total Sky Imager, TSI440 set-up to capture sky images and record solar zenith angle (SZA). Note that the LI-COR is connected to CR100 Campbell data logger at University of Southern Queensland Solar Research Laboratory.



PPFD = 54 μ mol m-2s-1 (low)

SZA = 75°: 1 Mar 06.55 AM

PPFD = 333 μ mol m-2s-1 (medium)

SZA = 78°: 30 Mar, 16.55 PM

PPFD = 410 μ mol m-2s-1 (medium)

SZA = 50°: 5 Mar, 08:55 AM

PPFD = 1263 μ mol m-2s-1 (high)

SZA = 28°: 10 Mar 10:55 AM

PPFD = 1754 μ mol m-2s-1 (high)

SZA = 29°: 15 March, 12.55 PM

Figure 3

(a) Right: Temporal variations in photosynthetic photon flux density (PPFD, μ mol of photons m-2 s-1) over a 30-day period (01–31 Mar 2013) measured at every 5-minute intervals 07.00 AM to 05.00 PM. Note that the stochastic variations in PPDF occur in response to the subtle or rapid pertubations in cloud cover conditions that are not captured by a clear sky model. (b) Bottom: Sample images obtained by Total Sky Imager (TSI) capturing cloud cover conditions associated with simultaneously measured PPFD, solar zenith angle (SZA) and the time of the day.

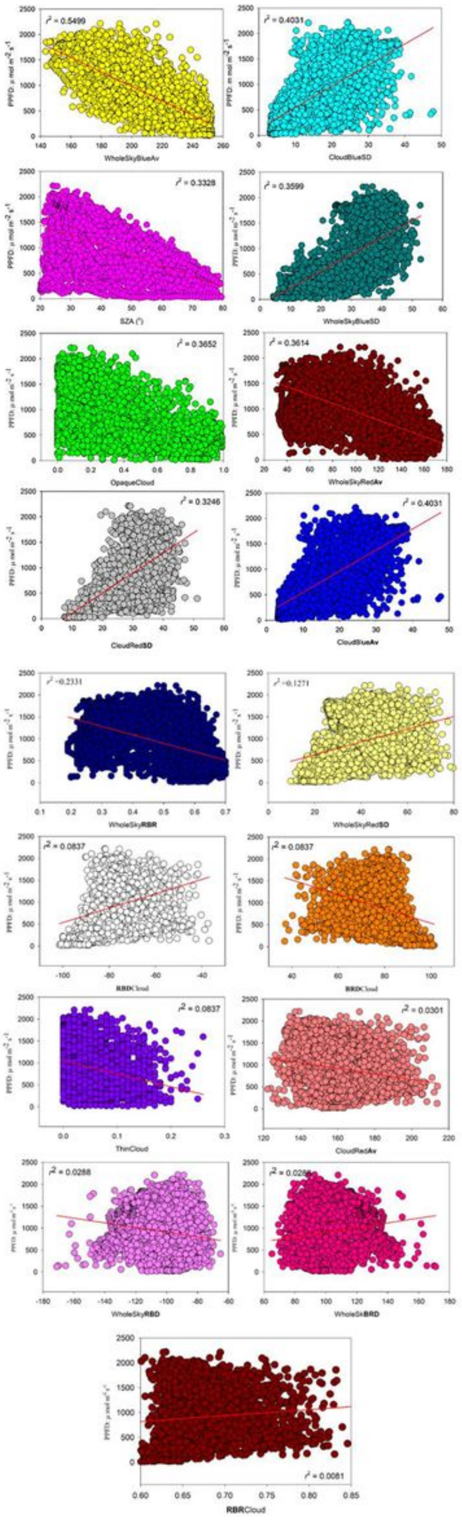


Figure 4

Scatterplot-based correlation analysis of the 5-minute PPFD (i.e., the objective variable) in respect to the 17 cloud-image derived predictor variables used in training the proposed CSLTM model. Least square regression lines with the coefficient of determination (r2) is included for each sub-panel with the definition of each cloud-image derived predictor variable as per in Table 1.

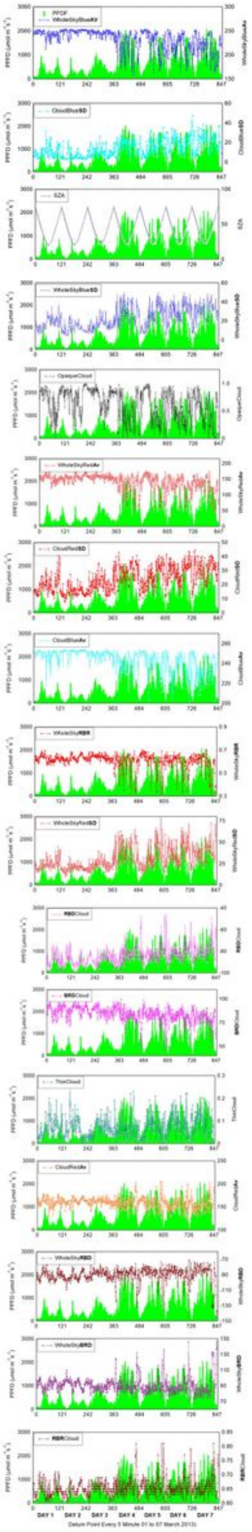


Figure 5

Comparison of the 5-minute PPFD (left axis) plotted for the first 7 days within the CLSTM model's training phase in respect to the 17 cloud-image derived predictor variables. Definition of each predictor (right axis) is as per Table 1.

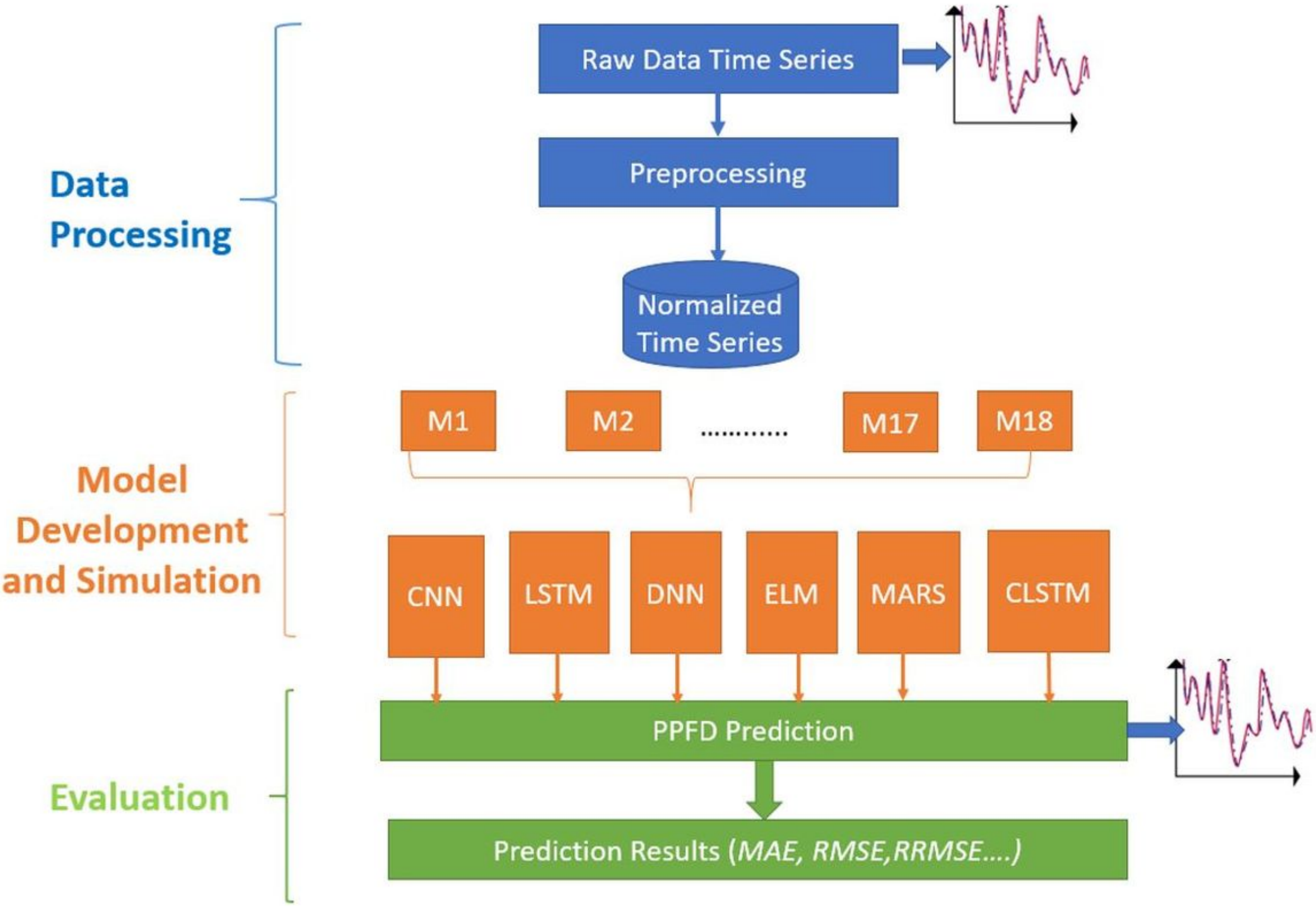


Figure 6

Schematic diagram of the relevant steps in designing the CLSTM predictive model.

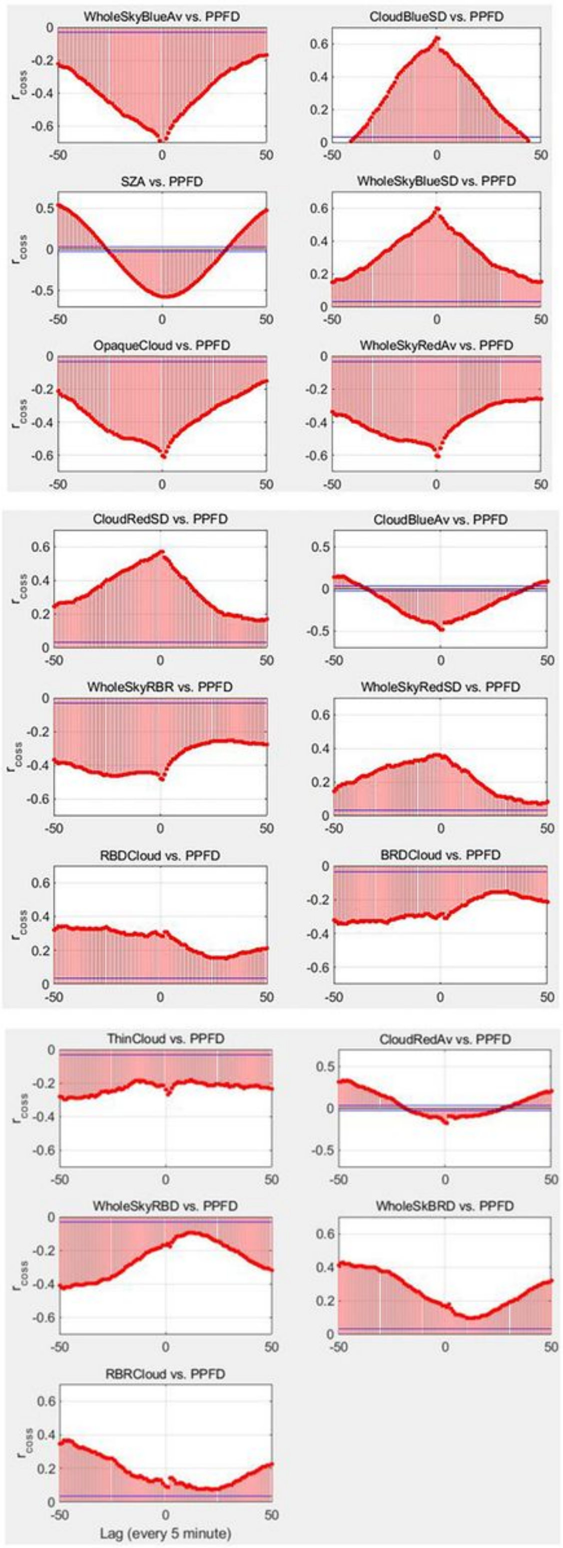


Figure 7

Correlograms plotted to identify the degree of covariance between PPFD (i.e., the objective variable) and the 17 different cloud-image derived predictor variables within the CLSTM model's training phase The y-axis shows cross-correlation coefficient, rcross with blue line representing the level at the 95% confidence interval.

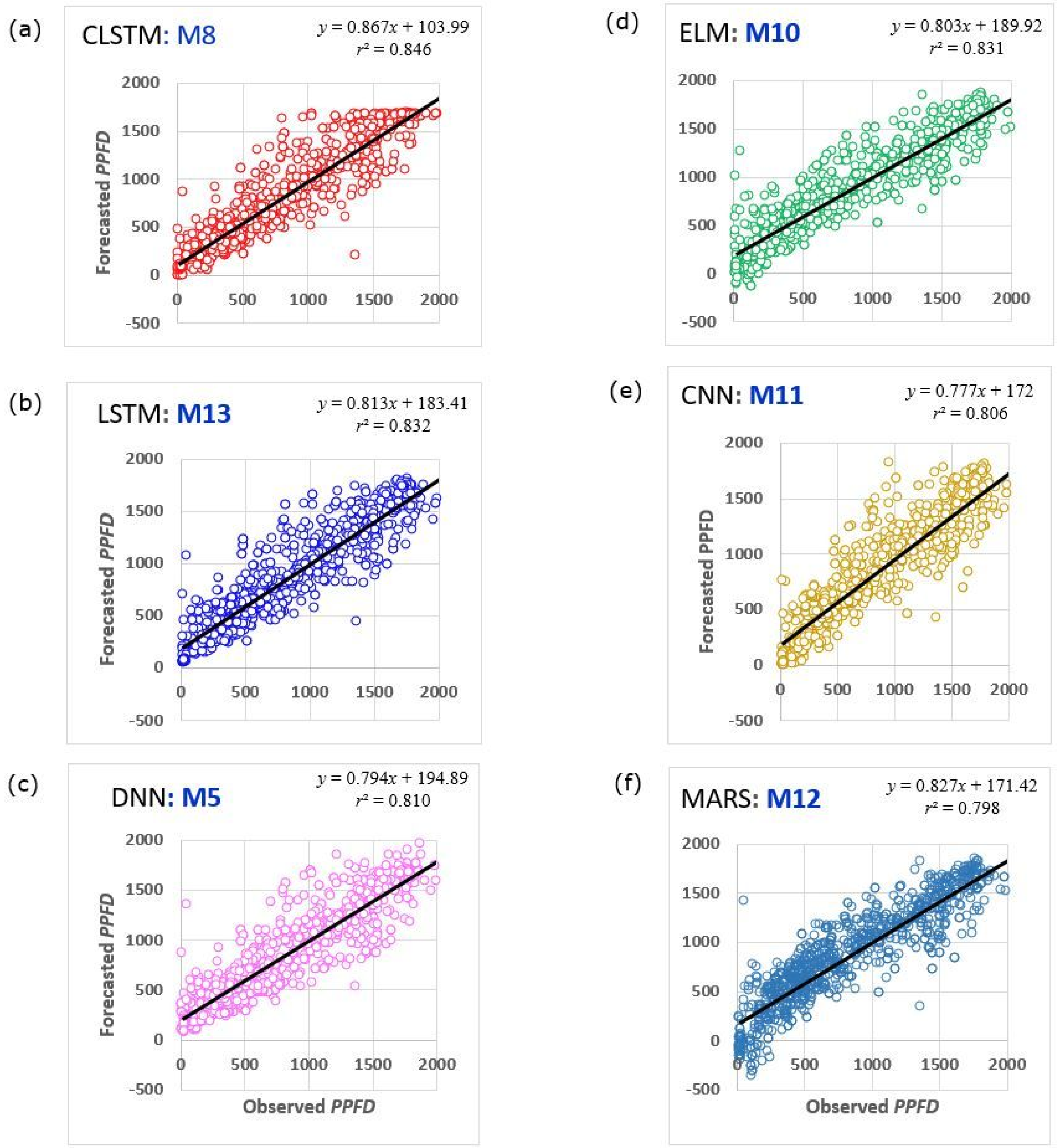


Figure 8

Scatterplots of forecasted against observed PPFD values (μ mol of photons m-2s-1) emulated by the CLSTM model in the testing phase, compared with benchmark models. Only the optimal results (out of all designated models, M1 to M17) for each predictive algorithm based on best input combinations utilising cloud chromatic statistics and SZA as predictors, as per Table 2, are shown.

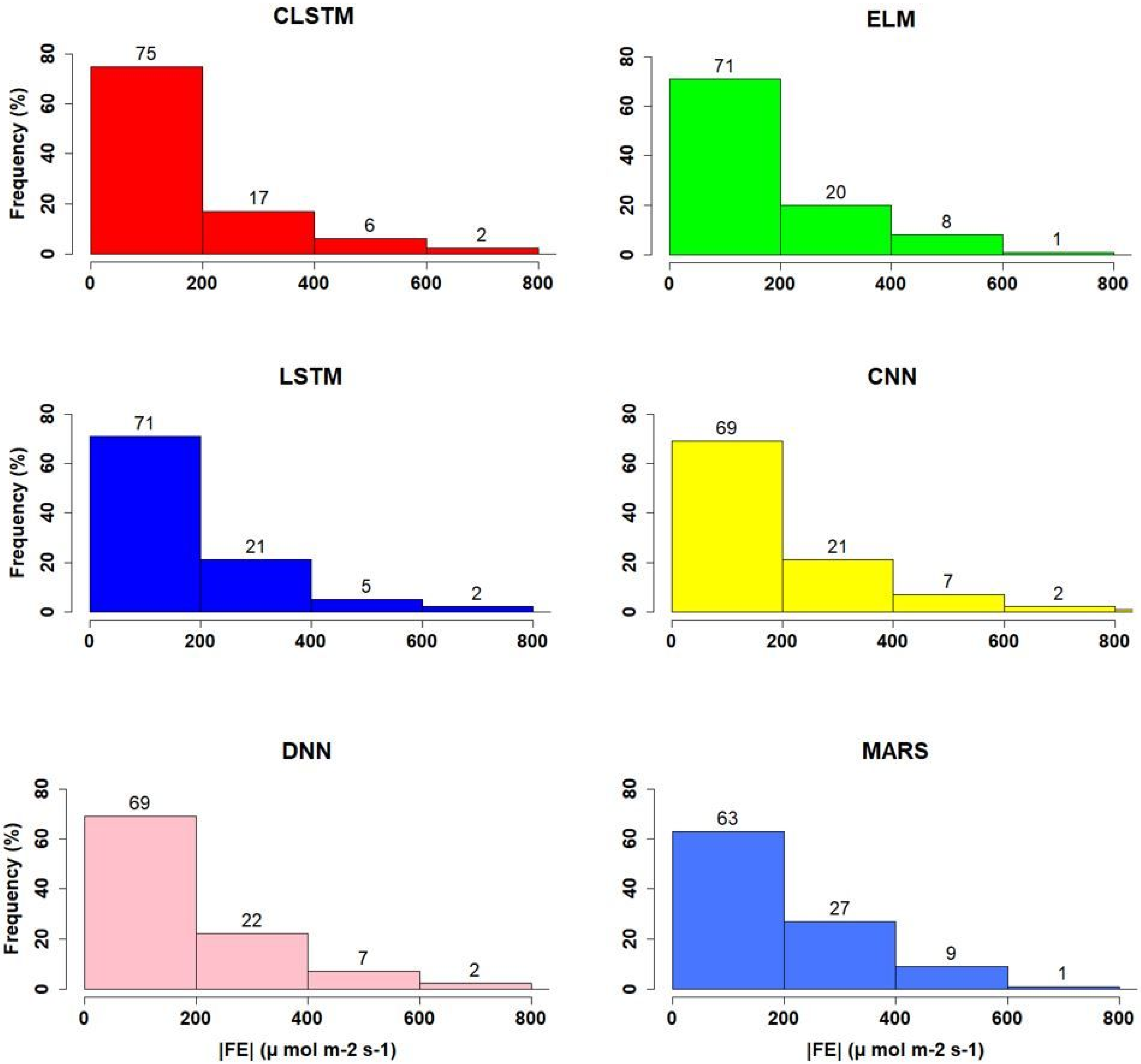


Figure 9

The percentage frequency of the forecasted error generated by the CLSTM model against the deep learning (i.e., LSTM, CNN, DNN) and machine learning (ELM, MARS)-based models developed using best input combinations utilising cloud chromatic statistics and SZA as the predictors, in accordance with Table 2.

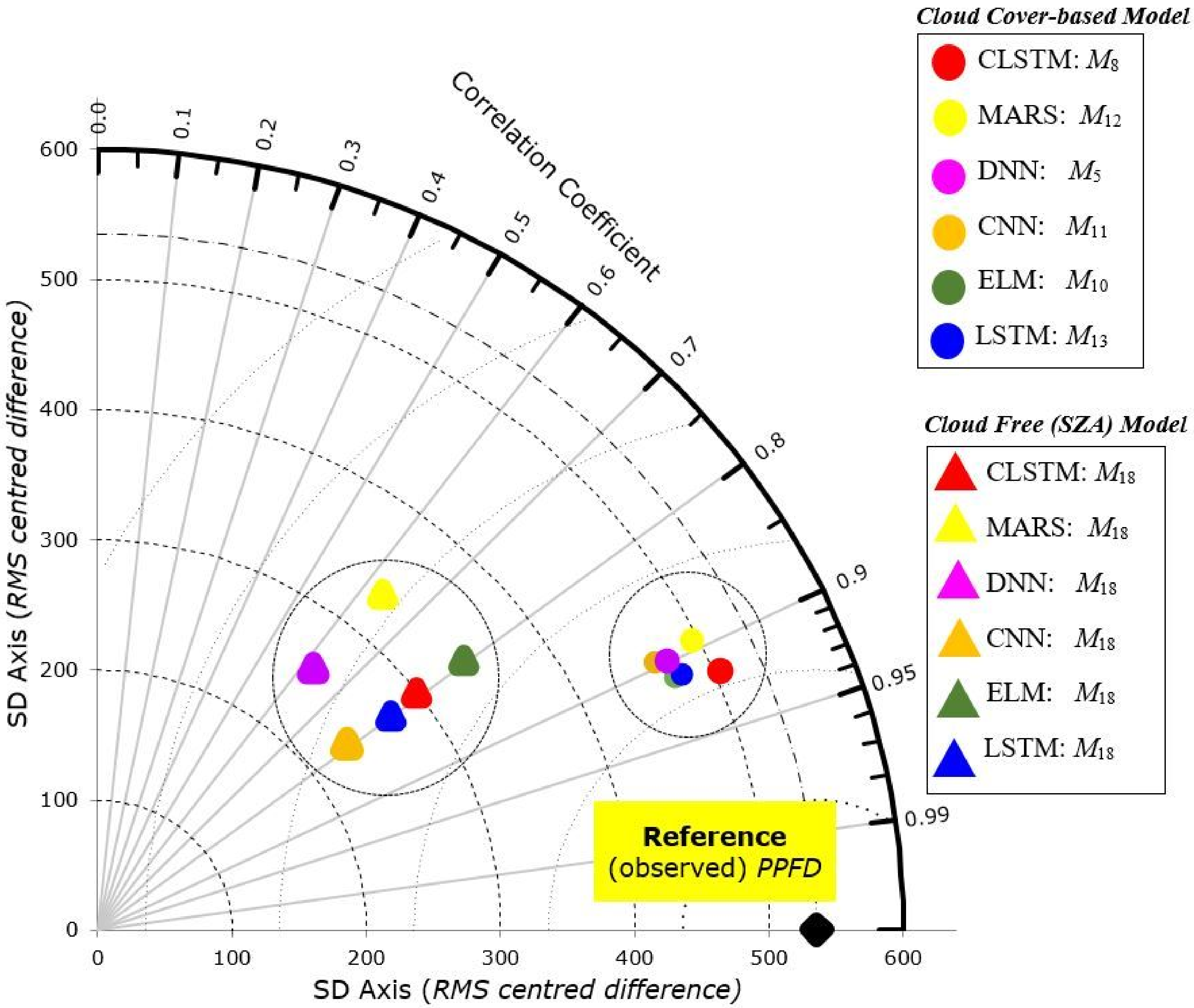


Figure 10

Taylor diagram with a concise statistical summary of how well the simulations from the CLSTM predictive model match with the other models in terms of their correlations between observed and forecasted PPFD, root-mean-square difference and the ratio of the variance in testing phase. Only the most optimal model with cloud cover properties (i.e., M¬8, M13, M12, M5, M11 and M10) and without cloud properties (i.e., M18 trained with SZA as input variable) are shown.

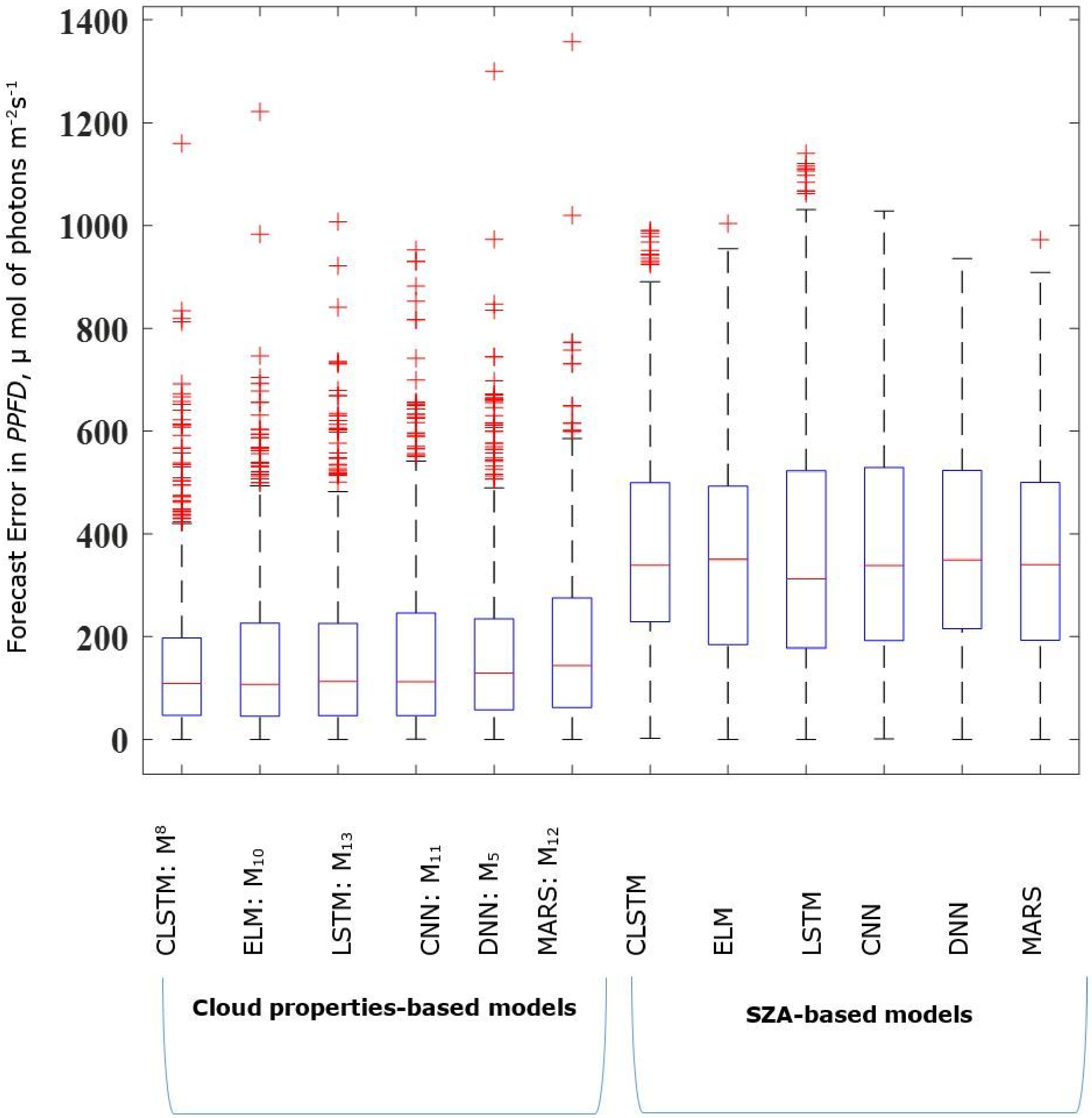


Figure 11

Boxplot of the absolute forecasted error in PPFD: |FE| = |PPFDifor - PPFDiobs| within the testing phase using the cloud cover-based and the SZA only reference models. Figure legend should also indicate what the line, box, whiskers and points represent.

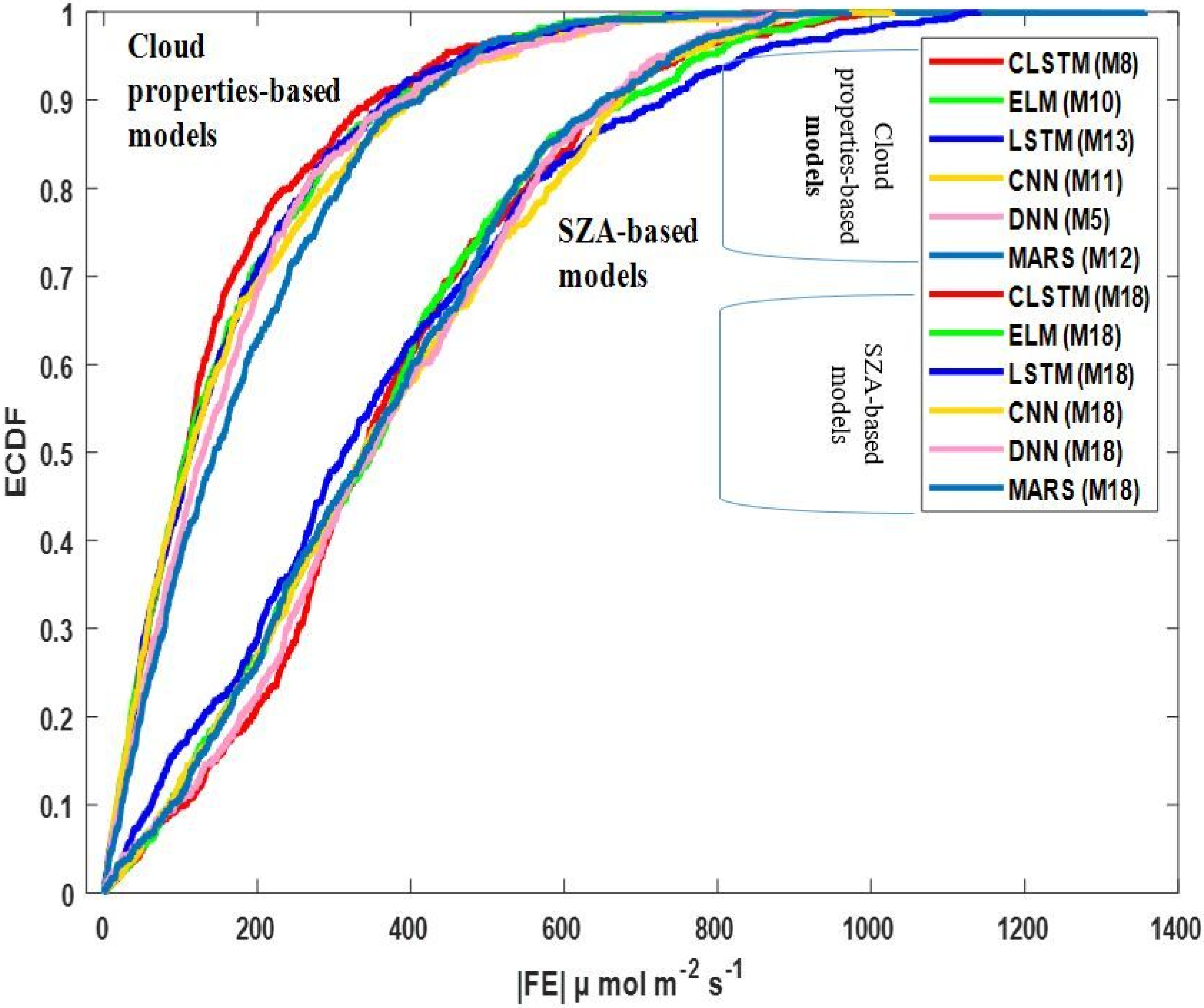


Figure 12

Empirical cumulative distribution function (ECDF) of the PPFD forecasting error |FE| in the testing phase.

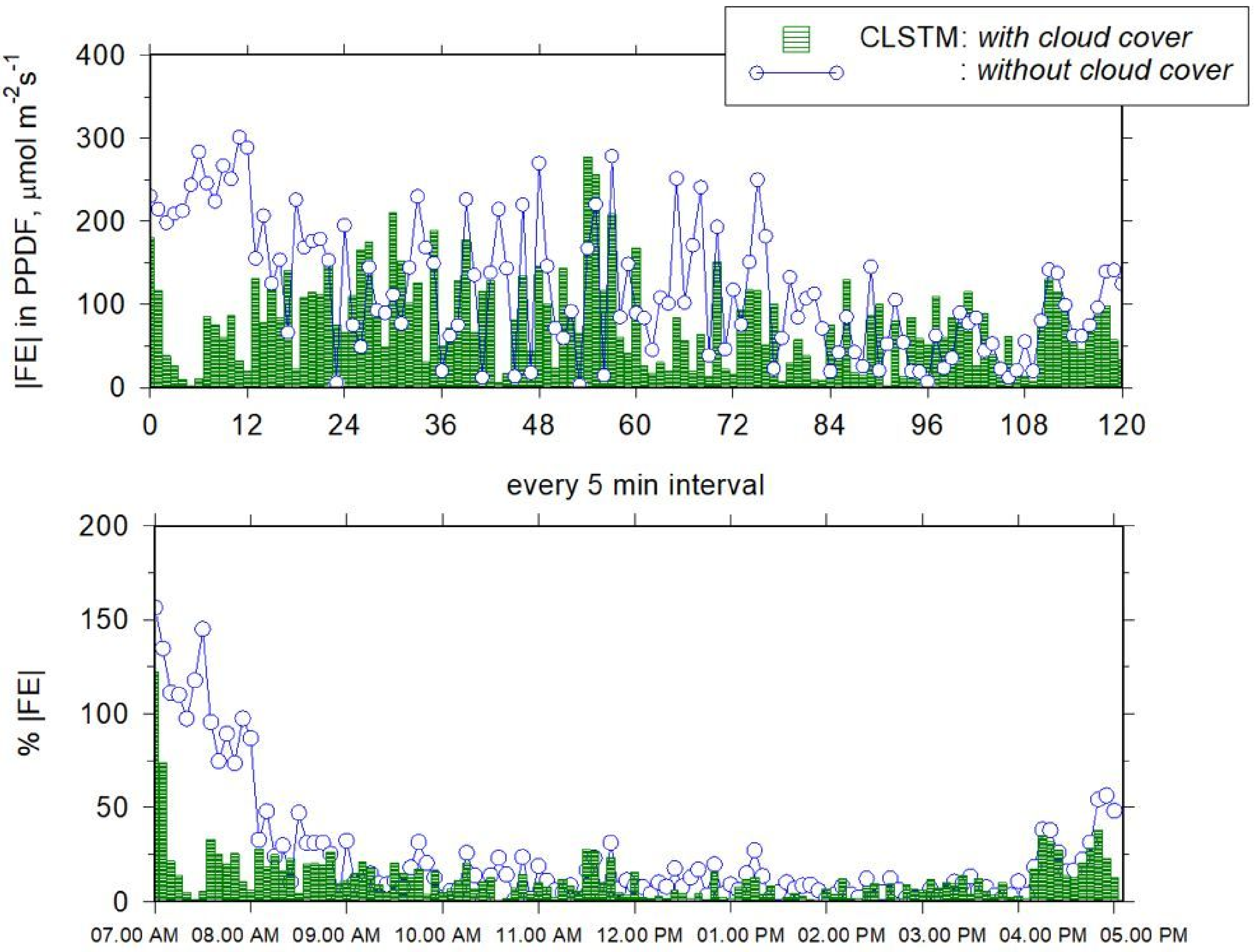


Figure 13

The effect of cloud cover properties used as inputs for the CLSTM model with 5-minute forecasted PPFD averaged over the entire testing dataset from 07.00 AM to 05.00 PM.