Stochastic Environmental Research and Risk Assessment
https://doi.org/10.1007/s00477-022-02188-0

ORIGINAL PAPER

Forecasting solar photosynthetic photon flux density under cloud cover 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]

Accepted: 1 February 2022

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-min) *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 $RRMSE = 24.92\%$ $MAPE = 38.01\%$, and Nash Sutcliffe's coefficient $E_{NS} = 0.85$, and Legate and 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 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}) (McCree 1973) 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

Extended author information available on the last page of the article

Published online: 06 April 2022

worldwide energy production. Currently, green biofuels account for 11% of the world's total energy supply (Proskurina et al. 2019) coming from primary plant and vegetable oil crops, secondary lignocellulosic by-products (Ramanna et al. 2017; Vuppaladadiyam et al. 2018), 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 (Chen et al. 2011; Slade and Bauen 2013; Holdmann et al. 2019). 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 (Kumar et al. 2018; Park et al. 2019; Zhang et al. 2020). 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 (Ramanna et al. 2017). 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 (Pruvost et al. 2016). 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 (Siqueira et al. 2020). 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% (Grant and Heisler 1997; Holdmann et al. 2019). Broken cloud can bring about short-term cloud enhancement of solar radiation (up to $\sim$ 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* (Patil et al. 2017).

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 (Hanan et al. 1995) and as a result can be used to determine the best locations for establishing farms, greenhouses or other high value agricultural hubs (Zheng et al. 2016a, b; Gumma et al. 2020; Siqueira et al. 2020). Satellite remote sensing methods inherently must approximate the geometric absorption, scattering and transmission of clouds from relatively low resolution single-direction reflectance (Batey and Green 2000). The most important environmental predictors to determine the global *PPFD* on the earth's surface are the annual precipitation, monthly cloud fraction, bio-climate layer information and month (Chen et al. 2008; Tang et al. 2017; Hengl et al. 2018; Lozano 2021; Rocha 2021).

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 (Ali et al. 2018a, b), China (Han et al. 2020), USA (Crane-Droesch 2018) and Australia (Cai et al. 2019; Feng et al. 2019; Kamir et al. 2020). 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 (Wagner 1995). 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 (Ryu et al. 2018).

Biomass productivity is not only dependent on total *PPFD* but also the diffuse fraction of *PPFD* (Gu et al. 2002). Methods for retrospective *PPFD* estimation employ a mixture of remote satellite products, global reanalysis of climate information (Jiang et al. 2020) and local surface instrumentation (Deo et al. 2019) to model both direct and diffuse photosynthetic-active radiation and output biomass for a range of ecological and agricultural applications have also been developed (Hengl et al. 2018).

In respect to solar energy, monitoring or integration into electricity grids, intermittencies in power production are

highly driven by cloud variations (Zhen et al. 2017a). However, the ability to develop reliable models to predict short-term (e.g., 5–10 min) 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. (2017) 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. (2019) was based on an adaptive neuro-fuzzy inference system to predict *PPFD* over 5-min 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 min 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-min 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-min intervals through a custom-built cloud segmentation algorithm (Igoe et al. 2019) 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 (Ghimire 2019a), 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/validation (i.e., 01-March-2013 to 24-March-2013) and testing (25-March-2013 to 31-March-2013) 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. This site was also used to develop a wavelet CLSTM hybrid model for multi-step prediction of cloud-affected solar UV radiation (Prasad et al. 2022).

2 Theoretical overview

The theoretical details of deep learning (i.e., CNN, LSTM, DNN) and conventional machine learning (ELM and MARS) methods are described elsewhere (Al-Musaylh et al. 2018a, b, 2020; Chen et al. 2018; Ghimire 2019a; Wang et al. 2019). The CLSTM model, constructed by integrating CNN and LSTM, had been used elsewhere in natural language processing where emotions were analysed with text inputs (Wang et al. 2016), in speech processing where voice search tasks were performed using CLDNN combining CNN, LSTM and DNN (Sainath et al. 2015), in video processing with CNN and Bi-directional LSTM models built to recognize human actions in video sequences (Ullah et al. 2017), in the medical area where the CNN-LSTM method was developed to detect arrhythmias in electrocardiograms (Oh et al. 2018) and in industrial areas where a convolutional bi-directional LSTM model was designed to predict tool wearing (Zhao et al. 2017). Other studies with CLSTM are evident, for example, time series application for prediction of residential energy consumption (Kim and Cho 2019; Ullah et al. 2019), solar radiation prediction (Lee et al. 2018; Wang, et al. 2018; Ghimire 2019a; Gao et al. 2020) and wind speed prediction (Hong and Satriani 2020; Jaseena and Kovoor 2021; Meka et al. 2021) as well as stock market applications in the prediction of share prices (Vidal and Kristjanpoller 2020; Yadav et al. 2020). In the solar radiation forecasting area, the study of Ghimire et al. (2019a) 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.

Following earlier implementations (Ghimire 2019a), 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-min intervals. This objective model is depicted by a simplified schematic architecture in Fig. 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 (Kuo 2016) 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 (Chimmula and Zhang 2020; Song 2020). First introduced for object recognition in image processing (LeCun et al. 2015), the CNN model has a prominent structure composed of many convolution layers, pooling layers and one or more fully connected layer (Vidal and Kristjanpoller 2020). The primary building block applies a convolution filter (i.e., a kernel function) for input data to generate a feature mapping scheme (Li et al. 2020a). Using different filters, many sets of convolutions are performed in order to create different feature maps (Xie et al. 2020). 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 (Ma and Tian 2020). 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 (Barzegar et al. 2020).

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 (Zang et al. 2020). The RNN is generally considered to be somewhat incompetent in describing long-term dependences due to the gradient vanishing phenomenon (Bengio et al. 1994). Because of this, LSTM was developed by Hochreiter and Schmidhuber in 1997 (Hochreiter and Schmidhuber 1997) and enhanced by Graves in 2013 (Graves 2013). 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. (Wu and Lin 2019). In general, an LSTM block is made of the sigmoid (σ) and hyperbolic tangent (*tanh*) layers, and two operations including pointwise summation ($\oplus$) and multiplication ($\otimes$) operations, as shown schematically in Fig. 1. Mathematically, these processes can be defined by Eqs. 1–6 (Ghimire 2019a).

Input gate i_t:

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

Forget gate f_t:

$$f_t = \sigma\left(w_f x_t + R_f h_{t-1} + b_f\right) \tag{2}$$

Output gate y_t:

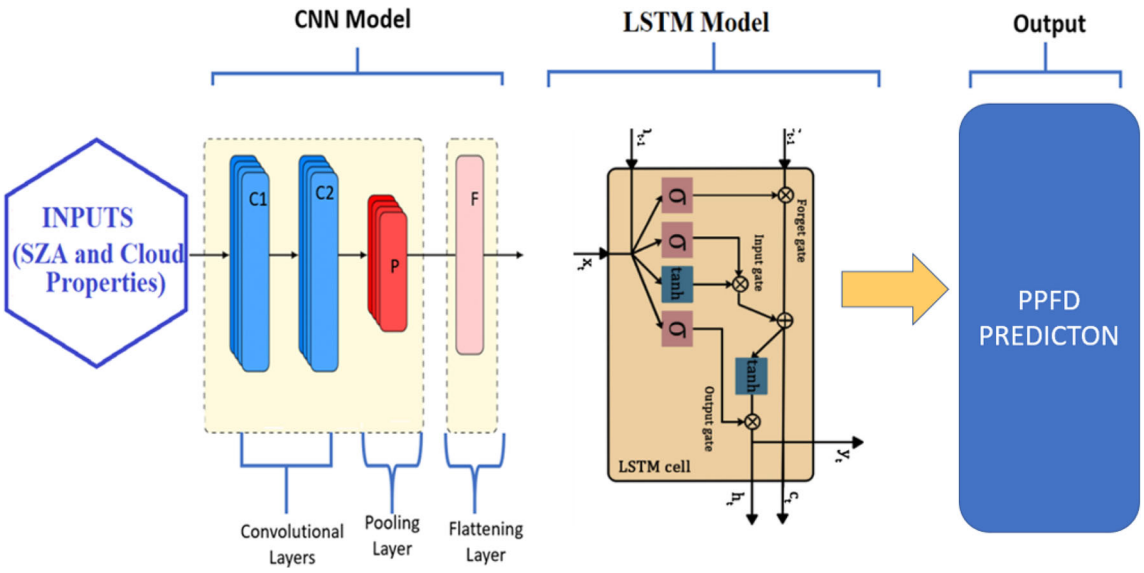


Fig. 1 Schematic illustration of convolutional neural network-long short-term memory network (CLSTM) predictive framework. CNN used for feature extraction from solar zenith angle (*SZA*) and cloud chromatic properties from total sky imager (*TSI*) and LSTM is used for time sequential modelling of the photosynthetic-active radiation (represented as photosynthetic photon flux density, *PPFD*)

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

Cell c_t:

$$c_t = f_t c_{t-1} + i_t \overline{c_t} \quad (4)$$

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

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

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

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

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

b_i, bf, 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 Materials and methods

3.1 Experimental apparatus and data acquisition system

The measurements were conducted using an experimental apparatus that has been reported elsewhere (e.g., Prasad et al. 2022). 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. Figure 2a 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 (Sabburg 2000) that also makes it an ideal site to implement the CLSTM model for real-time forecasting of photosynthetic-active radiation.

To build the proposed CLSTM predictive model, high-quality, yet cloud-influenced measurements of *PPFD* were acquired over the austral summer solstice period (01–31 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-min intervals over a 24-h period. Employed in several other research works (Johnson et al. 2015; Gill et al. 2017; Deo et al. 2019), 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.

Figure 3a shows the temporal patterns in measured *PPFD* time series sampled at 5-min 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). Figure 3b 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 (Igoe et al. 2019) 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 (Igoe et al. 2019) 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

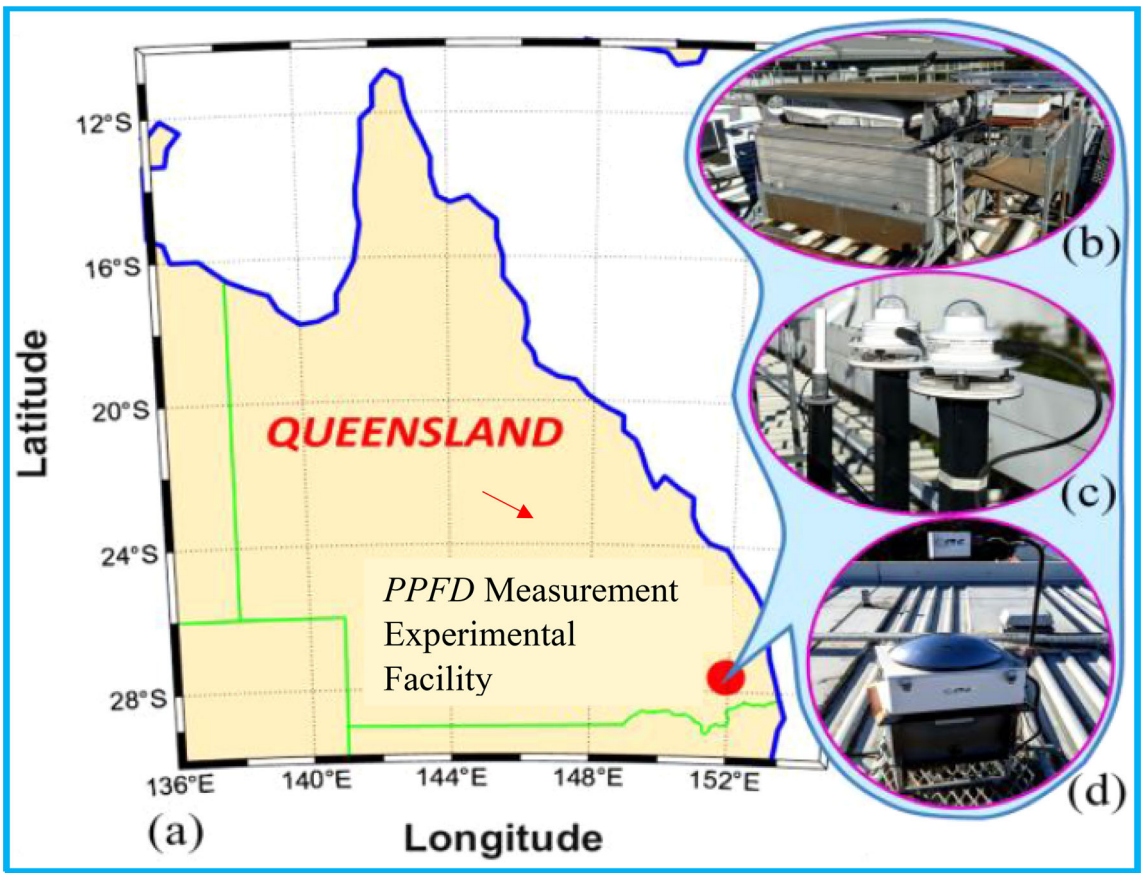


Fig. 2 **a** Geographic location of the measurement facility in Queensland, Australia where CLSTM model is implemented. **b** Roof-top mounted LI-COR sensor connected to the Campbell data logger for 5-min *PPFD* (μ mol of photons m^{-2} s^{-1}) measurement. **c** Co-located 501 broadband UVR Biometer. **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. The experimental set-up are also reported elsewhere (see Prasad et al. 2022)

paper (Deo et al. 2019) based on the commonly used red-blue ratio (Ghonima et al. 2012) 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 (Dev et al. 2016). 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 (Sabburg and Wong 1999) with the latter cloud type presenting a difficulty in cloud segmentation especially when aerosols are present (Li et al. 2011), 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-min interval cloud images considering over 200,000 images collected at a 480 × 320 spatial resolution. These whole-sky images have been captured using *TSI440* (Sabburg and Long 2004; Jebar et al. 2020; Liu et al. 2021) used in previous research (e.g., Sabburg and Wong 1999; Parisi et al. 2004; Deo et al. 2017). The *TSI440* instrument consists of a reflective dome with a camera suspended above it (Slater et al. 2001; Long et al. 2006) 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* (van der Walt et al. 2011) the OpenCV (van der Walt et al. 2014) libraries in Python. This is converted from OpenCV's blue-green–red (*BGR*) to red–green–blue (*RGB*) format for further image processing.

To account for any errors propagated in the proposed CLSTM model we note that by using a set threshold, as explained in the paper to be 0.56, and based on a test sample of data used in an earlier work (Igoe et al. 2019), the following was found about the % difference between TSI Analyser determined values and the TSI values (i.e., from observations): (i) A Pearson correlation = 0.93, (ii) about 85% of calculated values were within 10% of TSI observed values, (iii) the mean difference was 0.09% with standard

deviation difference of 9.29%, and (iv) furthermore, the median difference was 0.74% and the interquartile range was between 2 and 3% either side of the median. The earlier work (Igoe et al. 2019) also stated limitations that must be considered in using this segmentation method and using these, one can determine the impact on errors that will likely be propagated in the hybrid model.

Table 1 summarises the data for cloud chromatic properties derived from segmented images including the descriptive statistics (i.e., mean, standard deviation, difference, and 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 $r_{cross} \sim -0.747$, followed by the standard deviation of the blue cloud pixel ($r_{cross} \sim 0.640$). 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., Deo et al. 2019). 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.

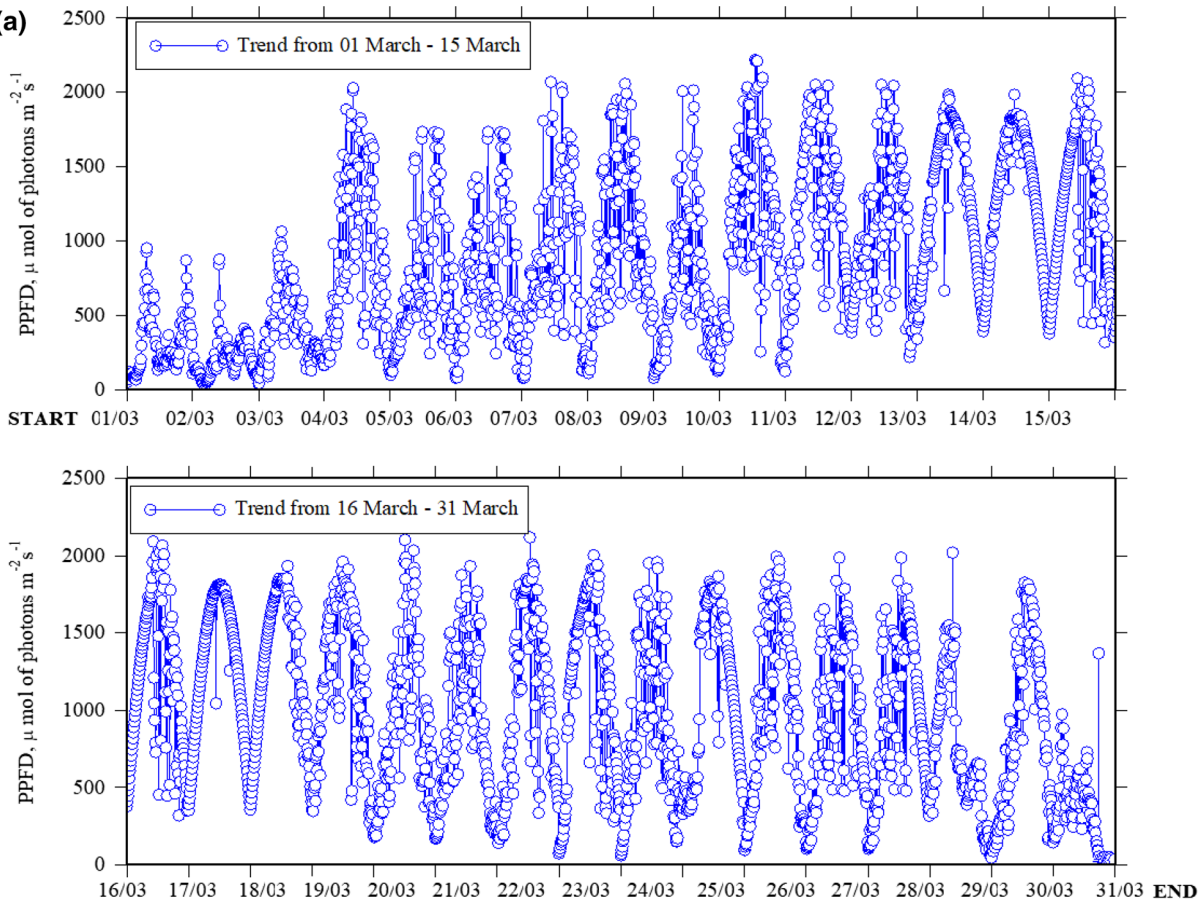
To corroborate the findings in Table 1 we now inspect visually the covariance in cloud chromatic properties against measured photosynthetic-active radiation. Figure 4 displays a scatterplot of the cloud cover statistics as well as *SZA* data that are regressed against the measured *PPFD* in the model training phase (i.e., 01-March-2013 to 17-March-2013). 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., Deo et al. 2019).

A comparison of the *PPFD* data series within the first 7 days (01-March-2013 to 07-March-2013) of model training data is made against cloud-image derived predictor series in Fig. 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 363–847 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.

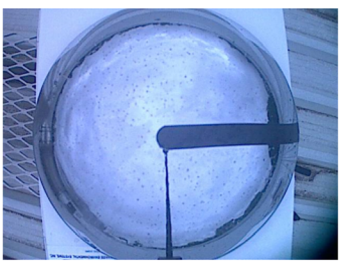
3.3 Predictive model design

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 (Konasani and Kadre 2021) and the MATLAB-based (Moler 2000) scripts were implemented on Intel *i*7 computer with 3.40 GHz processor running on 32 GB 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%, 01-March 2013 to 17-March-2013), validation (23.3%, 18-March-2013 to 24-March-2013), and testing (23.3%, 25-March-2013 to 31-March-2013) subsets.

(a)

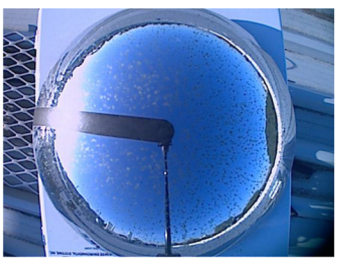


(b)



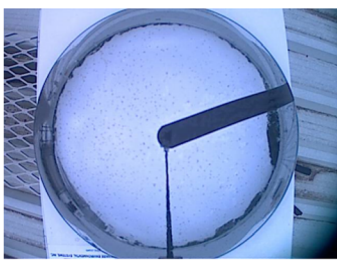
PPFD = 54 μ mol $m^{-2}s^{-1}$ (low)

SZA = 75°: 1 Mar 06.55 AM



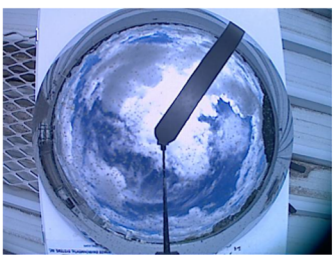
PPFD = 333 μ mol $m^{-2}s^{-1}$ (medium)

SZA = 78°: 30 Mar, 16.55 PM



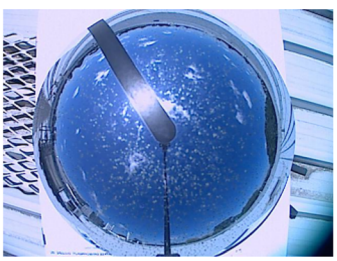
PPFD = 410 μ mol $m^{-2}s^{-1}$ (medium)

SZA = 50°: 5 Mar, 08:55 AM



PPFD = 1263 μ mol $m^{-2}s^{-1}$ (high)

SZA = 28°: 10 Mar 10:55 AM



PPFD = 1754 μ mol $m^{-2}s^{-1}$ (high)

SZA = 29°: 15 March, 12.55 PM

◀**Fig. 3** **a** 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-min 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** 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

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., Deo et al. 2016) 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 (Michalsky 1988)]. 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 (Deo et al. 2017, 2019) where *SZA* was the only predictor used to forecast *PPFD* and solar UV index ignoring cloud variations. This study advances the standard approaches (Deo et al. 2017, 2019) 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 (Deo et al. 2016) 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 (Deo et al. 2019), solar UV index (Deo et al. 2017) and global solar models (Deo et al. 2016), 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.

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-min) 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 min is used:

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

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-min 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:

$$X_N = \frac{X - \breve{X}}{\hat{X} - \breve{X}} \tag{10}$$

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-min 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 (Ghimire 2019a) to generate a 4-layered objective model (denoted as hybrid CLSTM). For model development parameters, see Appendix (Tables 5, 6, 7). 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.

Table 1 Description of predictors and target variable (photosynthetic active radiation measured in terms of photosynthetic photon flux density, PPFD, μmol m^{-2} s^{-1}) used to construct the deep learning hybrid CLSTM predictive model

Dataset and annotated short form		Description	Symbol	Descriptive statistics from time series				Cross correlation statistic	
				Mean	Standard deviation	Minimum	Maximum	r_{cross}	Rank (order of the CLSTM model input)
Predictor (input) variable	Average of the whole sky-blue [WholeSkyBlueAv]	Average of the blue pixel of the whole sky image that excludes black shadow band and the arm suspending the measuring camera Note: The blue wavelength within visible spectrum is scattered to a greater extent. The PPFD is expected to be larger with more blue sky	B_{av}	204.974	28.692	143.460	252.830	− 0.749	1
	Standard deviation of cloud—Blue [CloudBlueSD]	Standard deviation of the blue pixel values for just the cloud mass part within any image	BC_{sd}	18.770	8.393	3.055	47.715	0.640	2
	Solar zenith angle [*SZA*]	Angular position of solar disc used as a default measure of modelling the PPFD (Deo et al. 2019)	*SZA* (°)	46.519	15.929	20.030	79.550	− 0.631	3
	Standard deviation of the whole sky-blue [WholeSkyBlueSD]	The standard deviation of the blue pixels of the whole sky image	B_{sd}	30.745	9.927	4.350	52.750	0.599	4
	Opaque cloud [OpaqueCloud]	The proportion of the sky covered in thick clouds 0 = no thick cloud 1 = all sky covered in thick cloud	*OC*	0.364	0.323	0.000	0.990	− 0.581	5
	Average of the whole sky—Red [WholeSkyRedAv]	The average of the red pixels of the whole sky image but not including the black shadow band and the arm suspending the measuring camera The red wavelengths of visible are scattered more by cloud than clear sky and so are a good indicator of cloud	R_{av}	104.455	40.346	31.220	174.580	− 0.580	6
	Standard deviation of cloud—Red [CloudRedSD]	The standard deviation of the red pixel values for just the red clouds in the image	RC_{sd}	30.324	7.921	7.508	51.157	0.567	7
	Average of cloud—Blue [CloudBlueAv]	The average of blue pixel values for just the pixels classified as a blue cloud	BC_{av}	239.048	10.957	191.868	253.178	− 0.486	8
	Whole sky Red & Blue pixel ratio [WholeSkyRBR]	Average red pixel values divided by average blue pixel values, R_{Av}/B_{Av} for each image As clouds can enhance R/B ratio (Ghonima et al. 2012; Chauvin et al. 2015; Parisi et al. 2016; Igoe et al. 2018), this could be a good indicator of the amount of clouds in any image provided as the average RBR for each image	R_{av}/B_{av}	0.494	0.138	0.190	0.700	− 0.454	9
	Standard deviation of whole sky—Red [WholeSkyRedSD]	Standard deviation of red pixel of whole sky image	R_{sd}	38.851	13.399	9.670	79.690	0.350	10

Table 1 continued

Dataset and annotated short form		Description	Symbol	Descriptive statistics from time series				Cross correlation statistic	
				Mean	Standard deviation	Minimum	Maximum	r_{cross}	Rank (order of the CLSTM model input)
	Red & Blue cloud pixel difference [RBDCloud]	Difference of average red and blue pixel values (i.e., R_{av}—B_{av} for just the clouds in an image)	RBC_{diff}	− 77.101	9.671	− 101.394	− 36.899	0.280	11
	Blue & Red cloud pixel difference [BRDCloud]	Difference of average blue and average red pixel value difference (i.e., B_{av}—R_{av} for just the clouds in the image)	BRC_{diff}	77.101	9.671	36.899	101.394	− 0.280	12
	Thin cloud [ThinCloud]	The proportion of sky covered in thin cloud (0 = no thin cloud, 1 = sky covered in thin cloud) Software distinguishes between opaque and thin clouds using a threshold applied to red and blue pixel values of a sky image. Thin cloud could have an influence on PAR (e.g., if it is cirrus)	TC	0.045	0.045	0.000	0.260	− 0.250	13
	Average of cloud—Red [CloudRedAv]	Calculates the average of red pixel values for just the pixels classified as a cloud	RC_{av}	161.947	14.888	122.355	213.913	− 0.170	14
	Whole sky Red & Blue pixel difference [WholeSkyRBD]	Difference, R_{Av}—B_{Av} of average red and average blue pixel values for each image As clouds can enhance the RBD, this is a good indicator of the amount of clouds within any image, provided the average RBD for each sky image	RB_{diff}	− 100.519	17.533	− 171.090	− 65.130	− 0.107	15
	Whole sky Blue & Red pixel difference [WholeSkBRD]	Difference, B_{Av}—R_{Av} of average blue and average red pixel values for each image As clouds enhance the BRD, this is a good indicator of the amount of clouds within any image, provided as the average BRD for each sky image	BR_{diff}	100.519	17.533	65.130	171.090	0.107	16
	Red & Blue cloud pixel ratio [RBRCloud]	Average of red pixels/average of blue pixels value (i.e., R_{av}/B_{av} for just the clouds in the image)	RC_{av}/BC_{av}	0.677	0.043	0.584	0.853	0.093	17
Target variable	Photosynthetic active radiation (as a PPFD value)	Solar energy absorbed to produce organic compound required to sustain green biomass in terms of photosynthetic photon flux density (PPFD) This represents visible light photons 400 nm $\leq \lambda \leq$ 700 nm of solar spectrum incident upon a unit area per unit of time, absorbed by a plant	$PPFD$ (μ mol m^{-2} s^{-1})	2.732	1.647	0.042	6.651	1	Not applicable

Inputs were derived from segmented cloud images captured at 5-min interval for each measured PPFD with each time-dependent image presented as descriptive statistics based on blue, red, thin and the opaque cloud chromatic features. The mean, standard deviation, difference, or ratios are considered to determine the cross-correlation coefficient, r_{cross} against PPFD data series

Table 2 Input combinations used to construct CLSTM model with the data partitioned into training, validation and testing subsets

Desingated model	Input combinations (using cloud chromatic properties as per Table 1)	Data period (dd-mm-yyyy)	Data points/period			
			Total	Training (50%)	Validation (25%)	Testing (25%)
M1	$PAR = f\{B_{av}\}$	01–03-2013 to 31–03-2013	3630 for 30 days	1936 for 16 days	847 for 7 days	847 for 7 days
M2	$PAR = f\{B_{av}, BC_{sd}\}$					
M3	$PAR = f\{B_{av}, BC_{sd}, SZA\}$					
M4	$PAR = f\{B_{av}, BC_{sd}, SZA, B_{sd}\}$					
M5	$PAR = f\{B_{av}, BC_{sd}, SZA, B_{sd}, OC\}$					
M6	$PAR = f\{B_{av}, BC_{sd}, SZA, B_{sd}, OC, R_{av}\}$					
M7	$PAR = f\{B_{av}, BC_{sd}, SZA, B_{sd}, OC, R_{av}, RC_{sd}\}$					
M8	$PAR = f\{B_{av}, BC_{sd}, SZA, B_{sd}, OC, R_{av}, RC_{sd}, BC_{av}\}$					
M9	$PAR = f\{B_{av}, BC_{sd}, SZA, B_{sd}, OC, R_{av}, RC_{sd}, BC_{av}, R_{av}/B_{av}\}$					
M10	$PAR = f\left\{B_{av}, BC_{sd}, SZA, B_{sd}, OC, R_{av}, RC_{sd}, BC_{av}, \frac{R_{av}}{B_{av}}, R_{sd}\right\}$					
M11	$PAR = f\left\{B_{av}, BC_{sd}, SZA, B_{sd}, OC, R_{av}, RC_{sd}, BC_{av}, \frac{R_{av}}{B_{av}}, R_{sd}, RBC_{diff}\right\}$					
M12	$PAR = f\left\{B_{av}, BC_{sd}, SZA, B_{sd}, OC, R_{av}, RC_{sd}, BC_{av}, \frac{R_{av}}{B_{av}}, R_{sd}, RBC_{diff}, BRC_{diff}\right\}$					
M13	$PAR = f\left\{B_{av}, BC_{sd}, SZA, B_{sd}, OC, R_{av}, RC_{sd}, BC_{av}, \frac{R_{av}}{B_{av}}, R_{sd}, RBC_{diff}, BRC_{diff}, TC\right\}$					
M14	$PAR = f\left\{B_{av}, BC_{sd}, SZA, B_{sd}, OC, R_{av}, RC_{sd}, BC_{av}, \frac{R_{av}}{B_{av}}, R_{sd}, RBC_{diff}, BRC_{diff}, TC, RC_{av}\right\}$					
M15	$PAR = f\left\{B_{av}, BC_{sd}, SZA, B_{sd}, OC, R_{av}, RC_{sd}, BC_{av}, \frac{R_{av}}{B_{av}}, R_{sd}, RBC_{diff}, BRC_{diff}, TC, RC_{av}, RBD_{diff}\right\}$					
M16	$PAR = f\left\{B_{av}, BC_{sd}, SZA, B_{sd}, OC, R_{av}, RC_{sd}, BC_{av}, \frac{R_{av}}{B_{av}}, R_{sd}, RBC_{diff}, BRC_{diff}, TC, RC_{av}, RBD_{diff}, RBD_{diff}\right\}$					
M17	$PAR = f\left\{B_{av}, BC_{sd}, SZA, B_{sd}, OC, R_{av}, RC_{sd}, BC_{av}, \frac{R_{av}}{B_{av}}, R_{sd}, RBC_{diff}, BRC_{diff}, TC, RC_{av}, RBD_{diff}, RBD_{diff}, \frac{RC_{av}}{BC_{av}}\right\}$					
M18	$PAR = f\{SZA\}$					
	REFERENCE MODEL without cloud cover properties					

Note that the model M18 is a reference model built without cloud cover conditions: $PAR = f\{B_{AV}\}$. This denotes the respective model designed with just the B_{av} time-series as a predictor variable'

3.3.1 Common hyperparameters for deep learning (DL) models

Open-source DL Python libraries, Scikit-Learn (Pedregosa et al. 2011) and Keras (Ketkar 2017; Chollet 2018) 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 (Nwankpa et al. 2018; Hohman et al. 2019) so we have used the rectified linear unit (*ReLU*) (Agarap 2018).
- *Dropout*: This is considered as a potential regularization to minimize overfitting issues in order to improve the training performance (Garbin et al. 2020). The dropout aims to select a fraction of the neurons (defined as a real hyperparameter over the range 0–1) at each model iteration and prevent them from retraining (Lambert et al. 2018; Zhang et al. 2018a; Cai et al. 2019). 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 (Ayinde and Zurada 2017; Sato et al. 2018; Antczak 2019). 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 (Jaiswal et al. 2018; Byrd and Lipton 2019).
- *Early stopping*: The issue of overfitting can be further addressed by introducing an early stopping (ES) phase in Kera (Chollet 2017, 2018) so that the mode is set to a minimum while the patience is set to 30 (Byrd and Lipton 2019; Li et al. 2020b; Rice et al. 2020). 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 (Dodge et al. 2002; Mahsereci et al. 2017; Zhang et al. 2018b).

3.3.2 CNN hyperparameters and hybrid CNN-LSTM model development

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

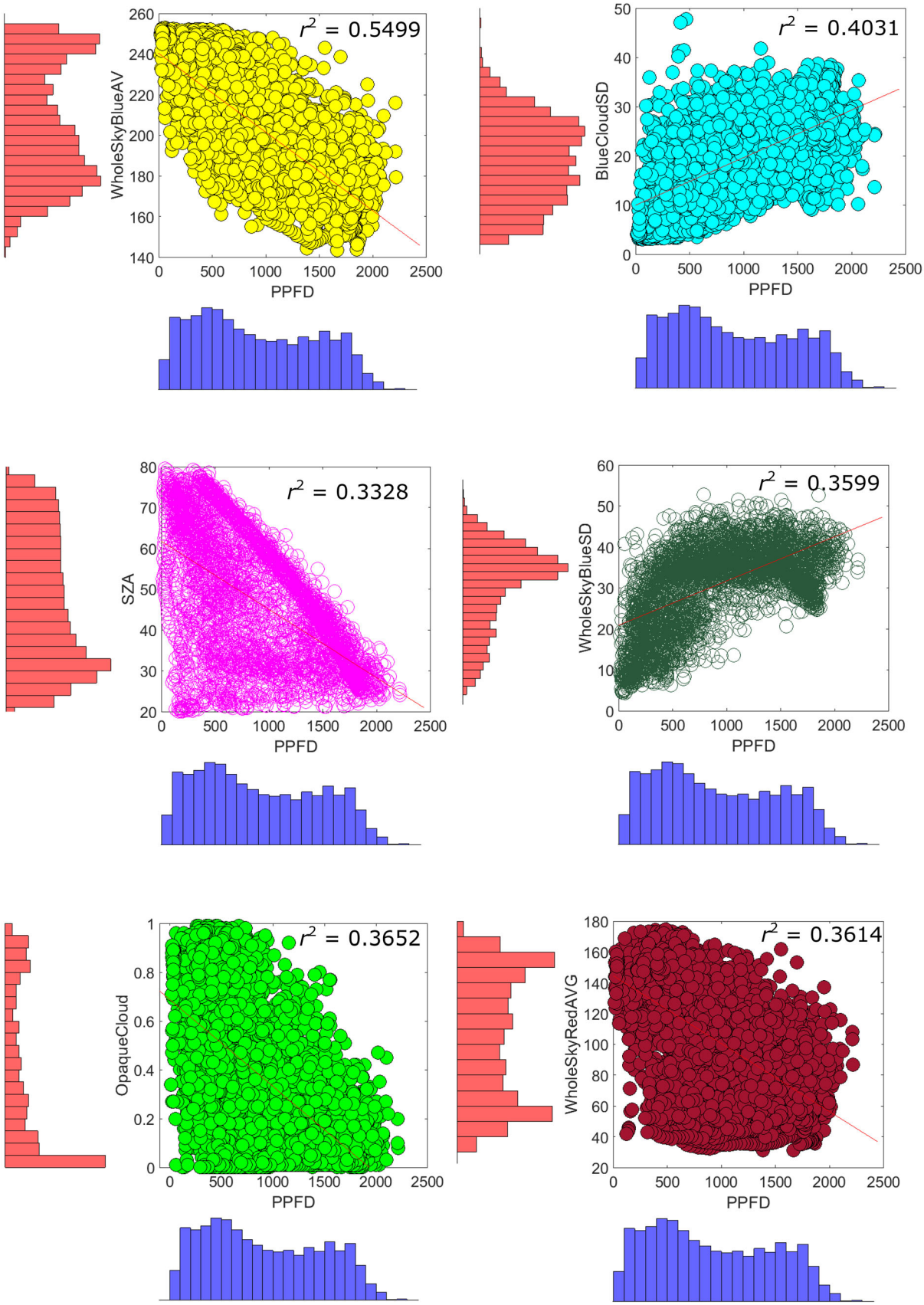
- *Filter size*: The size of the convolution operation filter was optimised.
- *Number of convolutions*: The number of convolutional layers in each CNN was optimised.
- *Padding*: This study has utilized the same padding in order to ensure that the input feature map and output feature map dimensions were identical (Zhang et al. 2016).
- *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 (Swietojanski et al. 2014). 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 (Deo et al. 2017). 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 3), 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 (Deo et al. 2017) with an optimal model achieved using logarithmic sigmoid equation indicated in Table 3. 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) (Jekabsons 2013) is adopted. Out of the two basis functions (i.e., cubic and linear) within its piecewise equation, the cubic form is adopted (Kooperberg and Clarkson 1997) given its capacity to handle multiple


r^2 = 0.5499
WholeSkyBlueAV
PPFD
r^2 = 0.4031
BlueCloudSD
PPFD
r^2 = 0.3328
SZA
PPFD
r^2 = 0.3599
WholeSkyBlueSD
PPFD
r^2 = 0.3652
OpaqueCloud
PPFD
r^2 = 0.3614
WholeSkyRedAVG
PPFD

◀**Fig. 4** Scatterplot-based correlation analysis with their respective histograms of each variable distribution showing the 5-min *PPFD* (i.e., the objective variable) in respect to the 16 cloud-image derived predictor variables used in training the proposed CSLTM model. Least square regression lines with the coefficient of determination (r^2) is included for each sub-panel with the definition of each cloud-image derived predictor variable as per in Table 1

predictors. The generalized recursive partitioning regression (RPR) is also employed as an adaptive algorithm for function approximation (Zareipour et al. 2006) 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 also lists the optimal MARS model equation. For greater details about ELM and MARS, readers can consult earlier References (Deo et al. 2017).

To further validate the hybrid CLSTM model, we adopt the skill score metric (*RMSE*ss) utilizing a persistence model ($RMSE_{\mathrm{P}}$) in respect to the measured PPFD data as a reference comparison. The *RMSE*ss is defined as follows:

$$RMSE_{ss} = 1 - \frac{RMSE_{CLSTM}}{RMSE_{persis\,\tan ce}} \tag{11}$$

where $RMSE_{\mathrm{CLSTM}}$ is the refers the error obtained by the objective model and $RMSE_{\mathrm{persistence}}$ is the error of the persistence model where immediate antecedent (past) value is used to estimate the current PPFD value. To interpret metric, we consider that an $RMSE_{\mathrm{CLSTM}}$ close to 0 will indicate that the performance of the hybrid model is similar to that of the persistence model. By contrast, if this metric is a positive value, our method is likely to outperform the persistence model (which is the baseline) whereas if the RMSEss attains a negative value, then the persistence model is likely to be better than the proposed hybrid model. Table 4 shows that the $RMSE_{\mathrm{SS}}$ for the hybrid CLSTM attains a positive skill score (showing an edge over the persistence model) whereas the CLSTM model without cloud inputs (but using solar zenith angle) is quite poor as it attains negative skill score metric.

3.5 Predictive model performance evaluation

The study adopts the model performance metrics recommended by American Society for Civil Engineers (ASCETC 1993) to evaluate the hybrid CLSTM (and all the other benchmark) models. By appraising the degree of agreement between $PPFD_{\mathrm{for}}$ and $PPFD_{\mathrm{obs}}$ the computed metrics include correlation coefficient (r), mean absolute error (*MAE*, mmol $\mathrm{m}^{-2}\,\mathrm{s}^{-1}$), root mean square error (*RMSE*, mmol $\mathrm{m}^{-2}\,\mathrm{s}^{-1}$), including the relative % magnitudes of *RMSE* and *MAE*, Legate and McCabe's (*LM*) and the Nash Sutcliffe's coefficient (E_{NS}). Mathematically, these are as follows (Ghimire et al. 2018, 2019; Ghimire 2019a, b):

$$r = \left(\frac{\sum_{i=1}^{N} \left(PPFD_{for,i} - \overline{\underset{obs,i}{PPFD}} \right) \left(PPFD_{for,i} - \overline{\underset{obs,i}{PPFD}} \right)}{\sqrt{\sum_{i=1}^{N} \left(PPFD_{for,i} - \overline{\underset{obs,i}{PPFD}} \right)^2} \sqrt{\sum_{i=1}^{N} \left(PPFD_{for,i} - \overline{\underset{obs,i}{PPFD}} \right)^2}} \right) \tag{12}$$

$$MAE = \frac{1}{N} \sum_{i=1}^{N} \left| \left(PPFD_{for,i} - PPFD_{obs,i} \right) \right| \tag{13}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(PPFD_{for,i} - PPFD_{obs,i} \right)^2} \tag{14}$$

$$MAPE = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{\left(PPFD_{for,i} - PPFD_{obs,i} \right)}{PPFD_{obs,i}} \right| \times 100 \tag{15}$$

$$RRMSE = \frac{\sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(PPFD_{for,i} - PPFD_{obs,i} \right)^2}}{\frac{1}{N} \sum_{i=1}^{N} \left(PPFD_{obs,i} \right)} \times 100 \tag{16}$$

$$LM = 1 - \left[\frac{\sum_{i=1}^{N} \left| PPFD_{obs,i} - PPFD_{for,i} \right|}{\sum_{i=1}^{N} \left| PPFD_{obs,i} - \overline{\underset{obs,i}{PPFD}} \right|} \right], \; 0 \le LM \le 1 \tag{17}$$

$$E_{NS} = 1 - \left[\frac{\sum_{i=1}^{N} \left(PPFD_{obs,i} - PPFD_{for,i} \right)^2}{\sum_{i=1}^{N} \left(PPFD_{obs,i} - \overline{\underset{obs,i}{PPFD}} \right)^2} \right], \; -\infty \le E_{NS} \le 1 \tag{18}$$

where $PPFD_{obs}$ and $PPFD_{for}$ are the observed and forecasted i^{th} value in test period, $\overline{\underset{obs}{PPFD}}$ and $\overline{\underset{\mathrm{for}}{PPFD}}$ 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-min forecasts models.

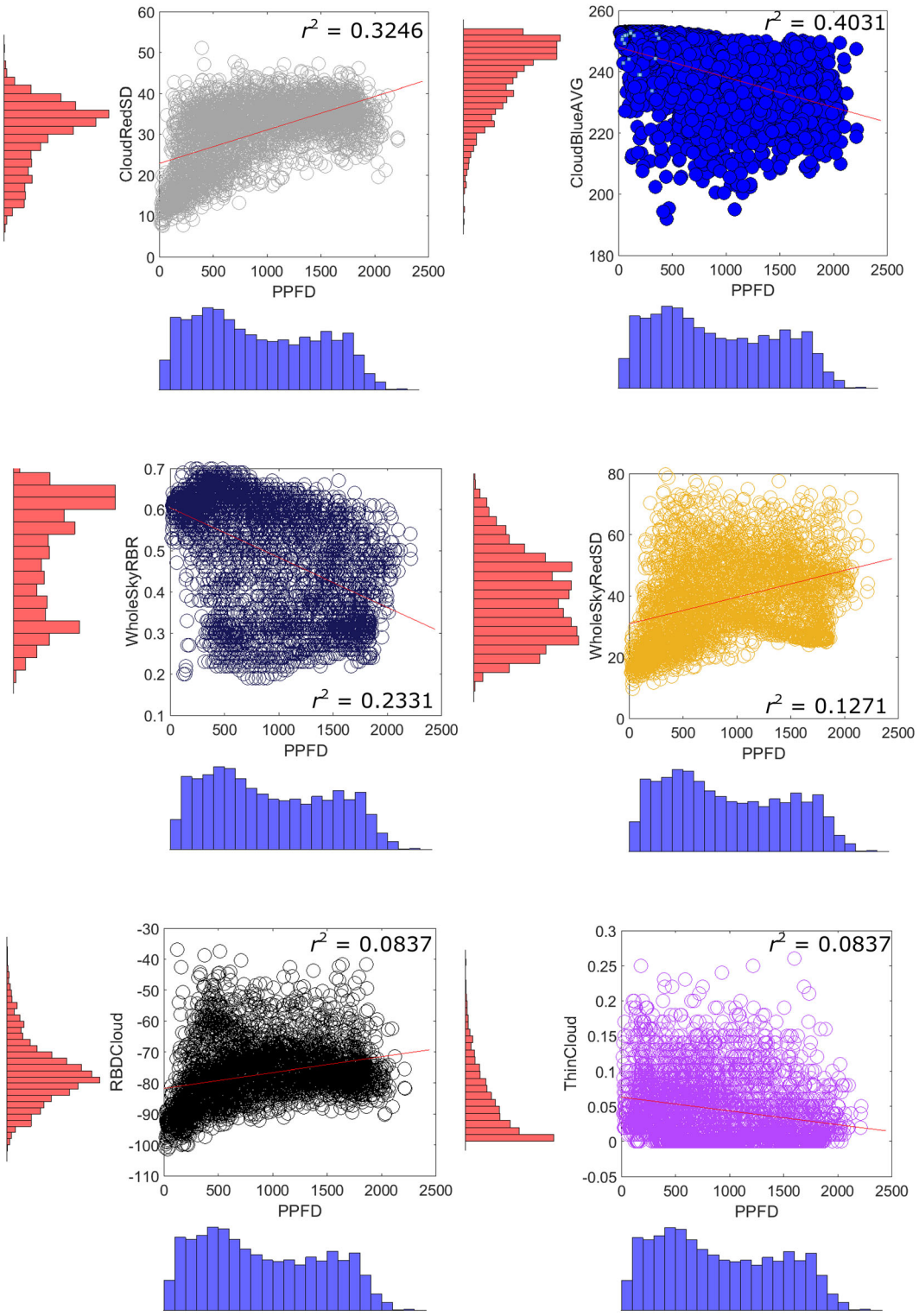


Fig. 4 continued

4 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-min 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-min *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 ($\sim$ 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

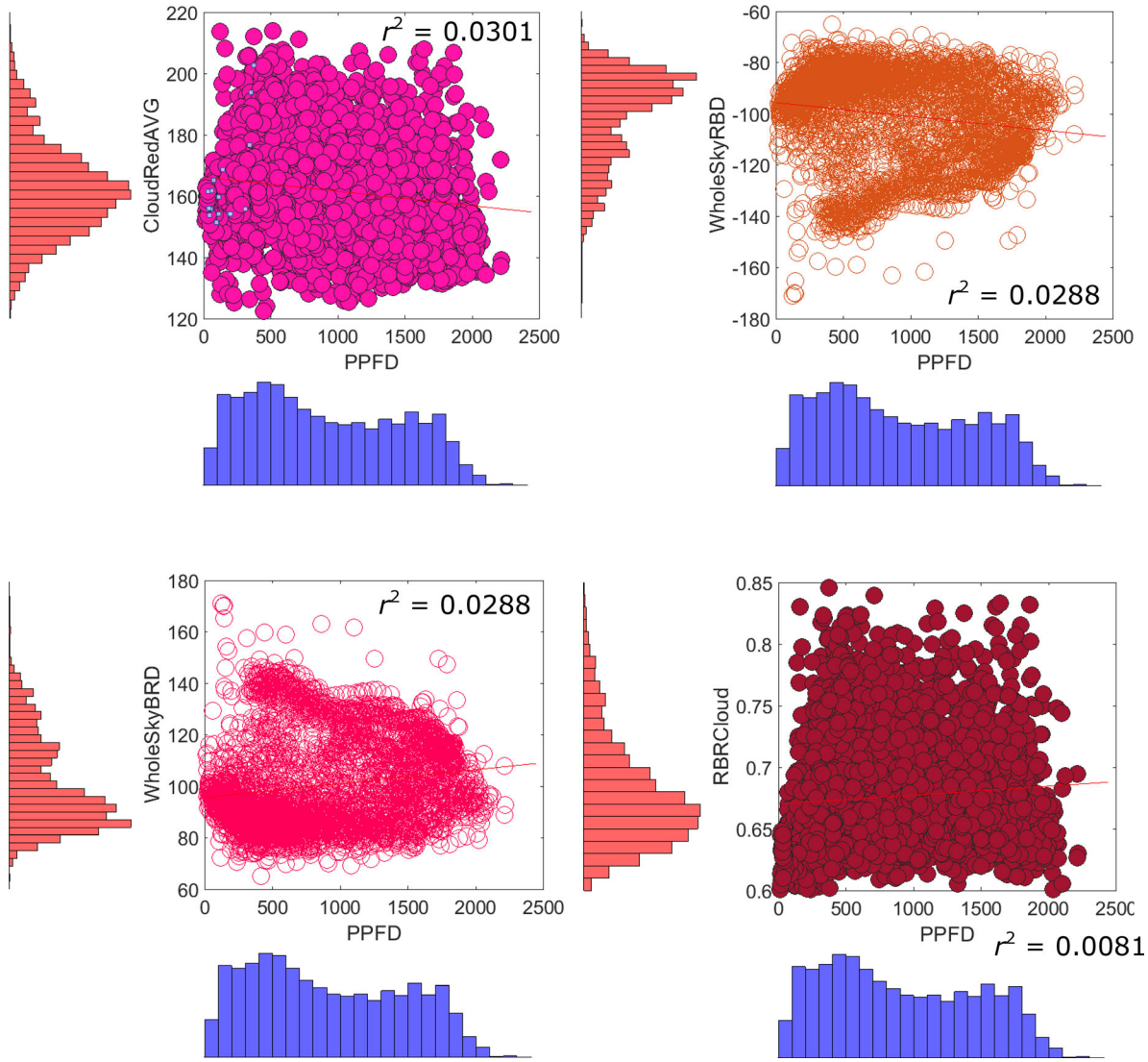


Fig. 4 continued

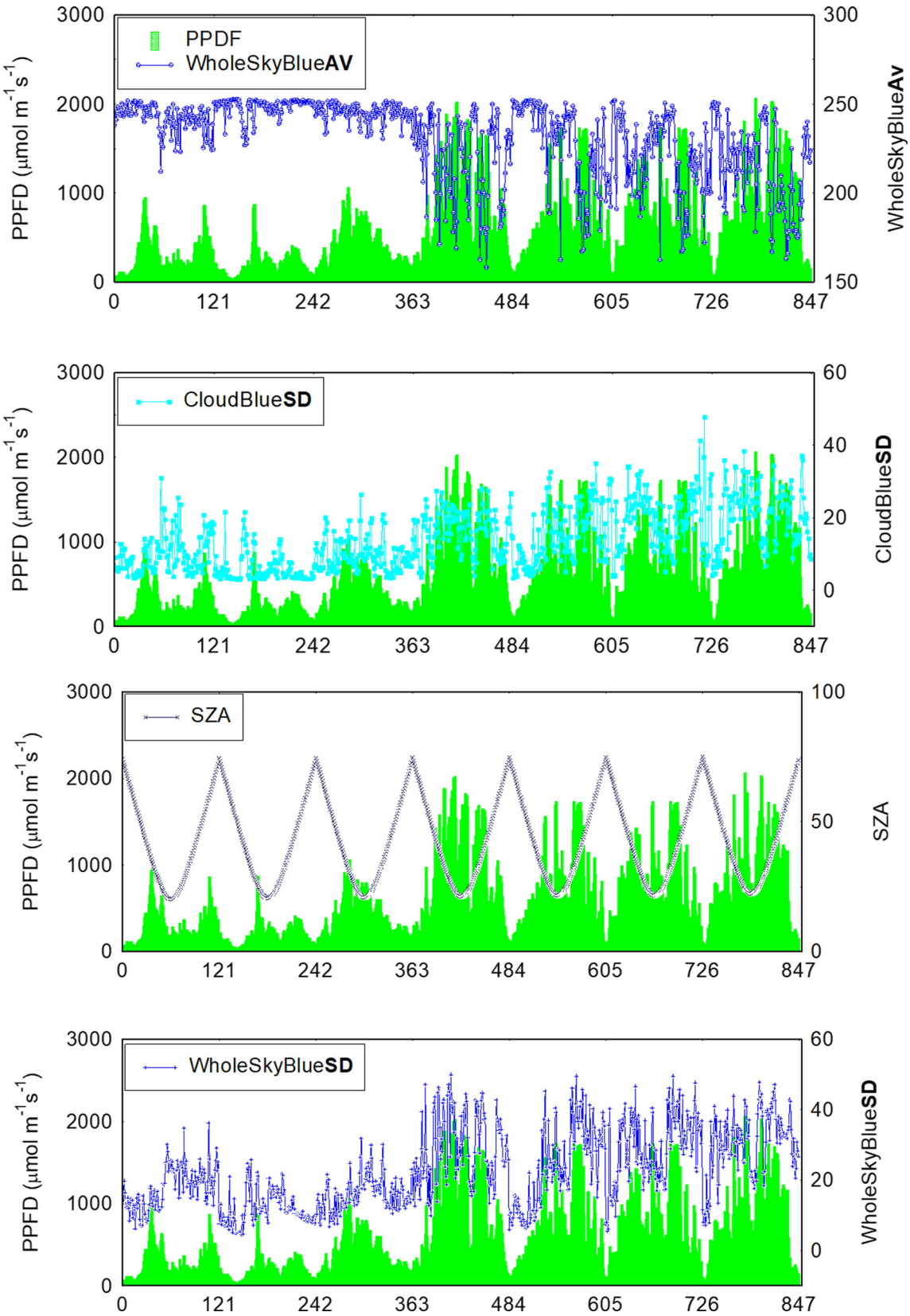


Fig. 5 Comparison of the 5-min *PPFD* (left axis) plotted for the first 7 days within the CLSTM model's training phase in respect to the 17 cloud-image derived predictor variables. Definition of each predictor (right axis) is as per Table 1

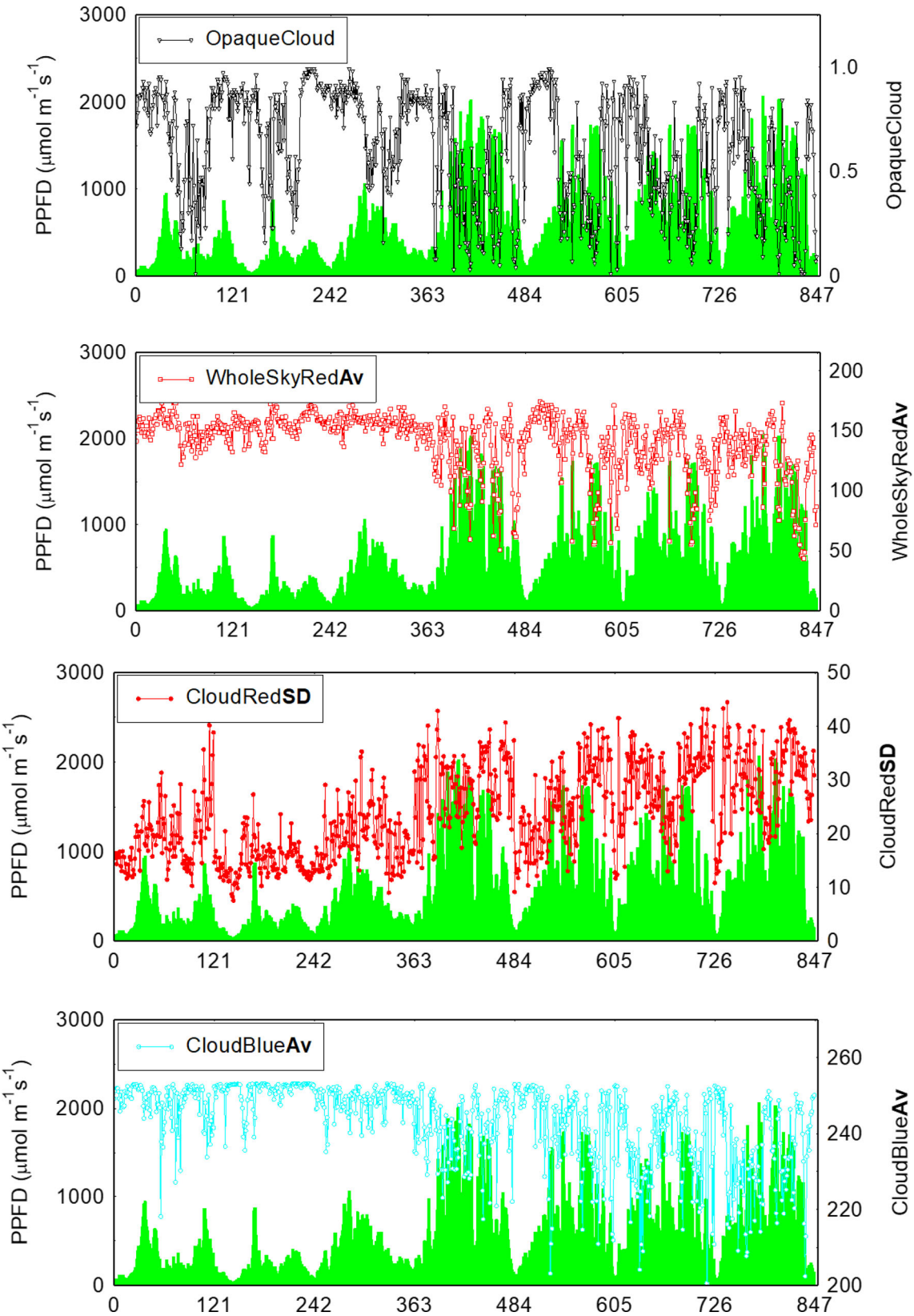


Fig. 5 continued

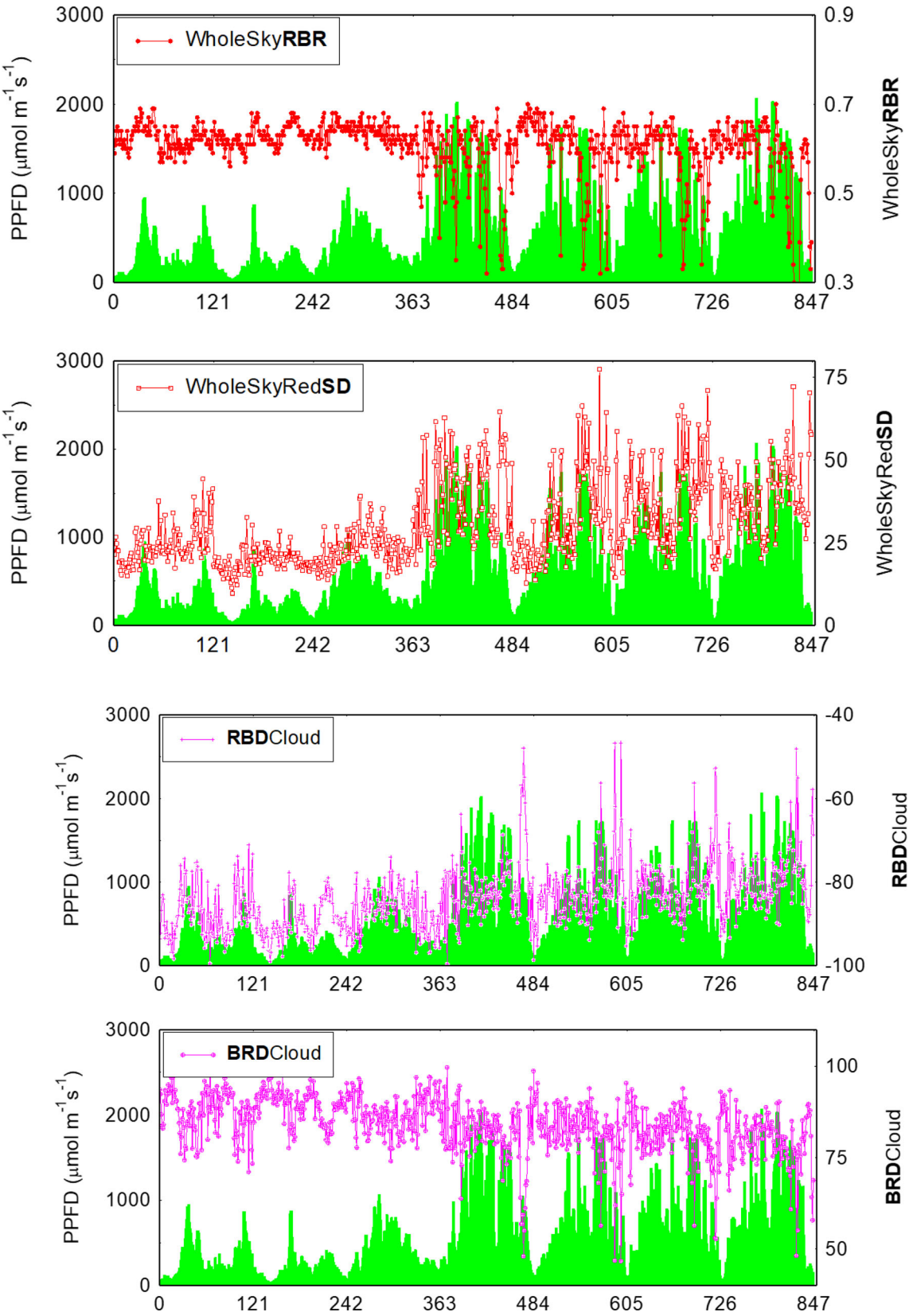


Fig. 5 continued

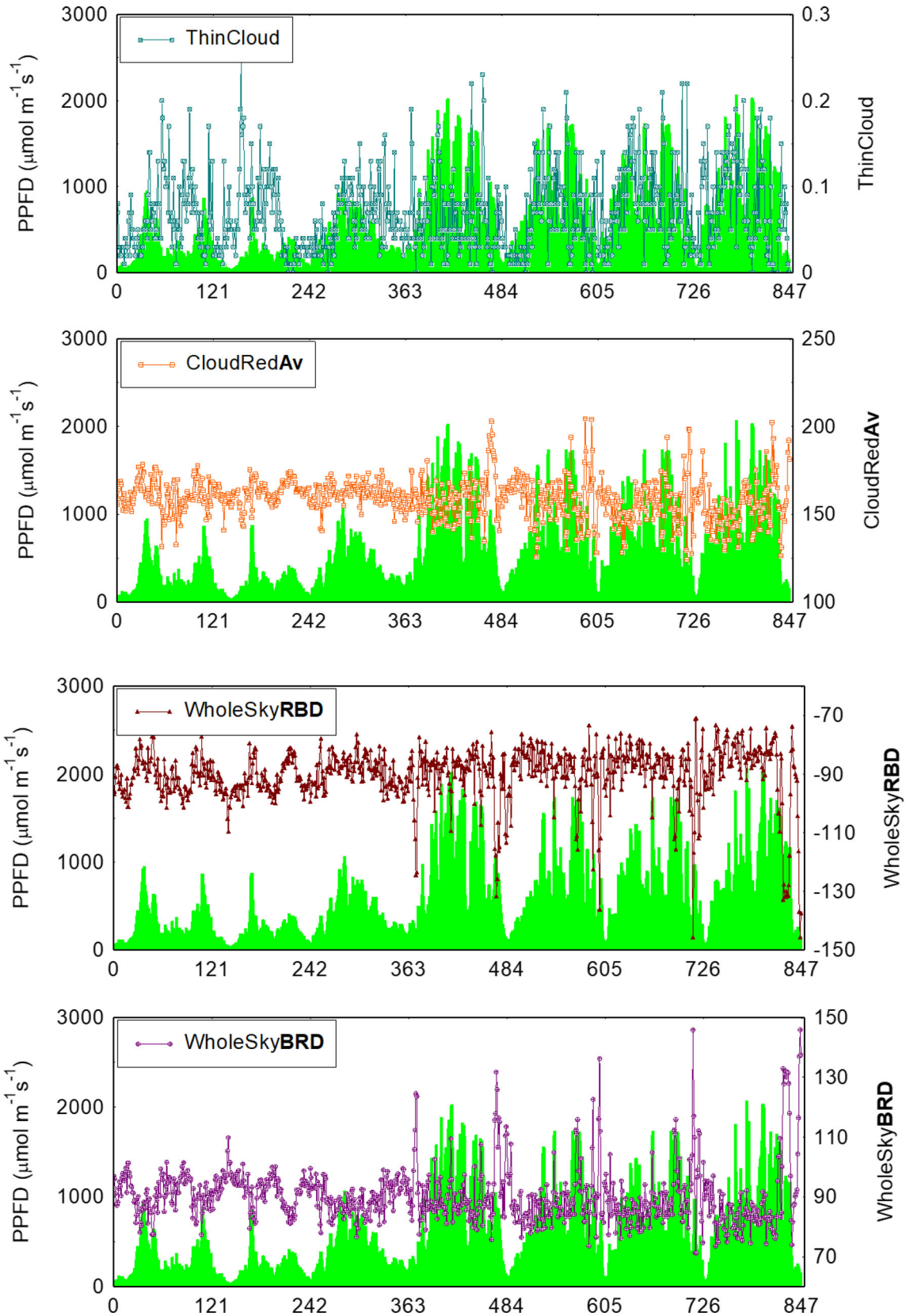


Fig. 5 continued

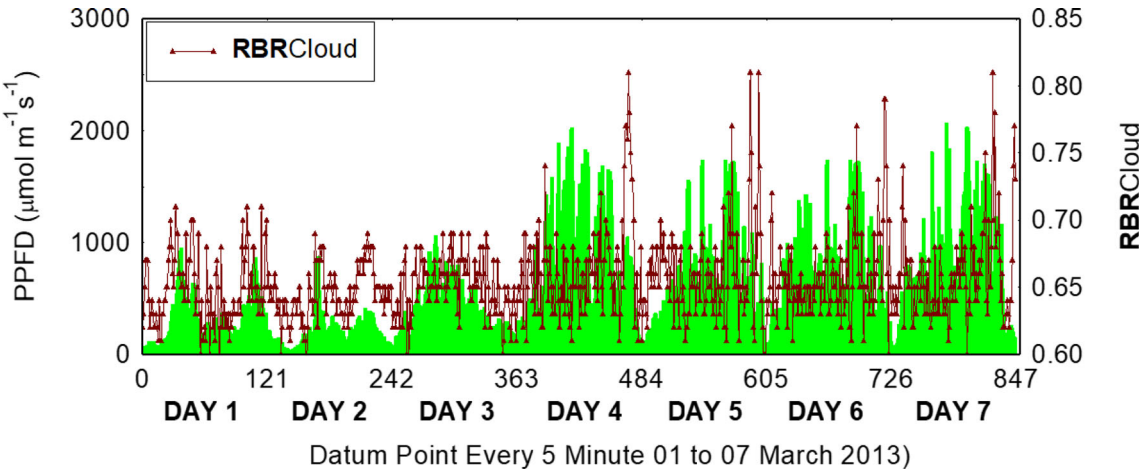


Fig. 5 continued

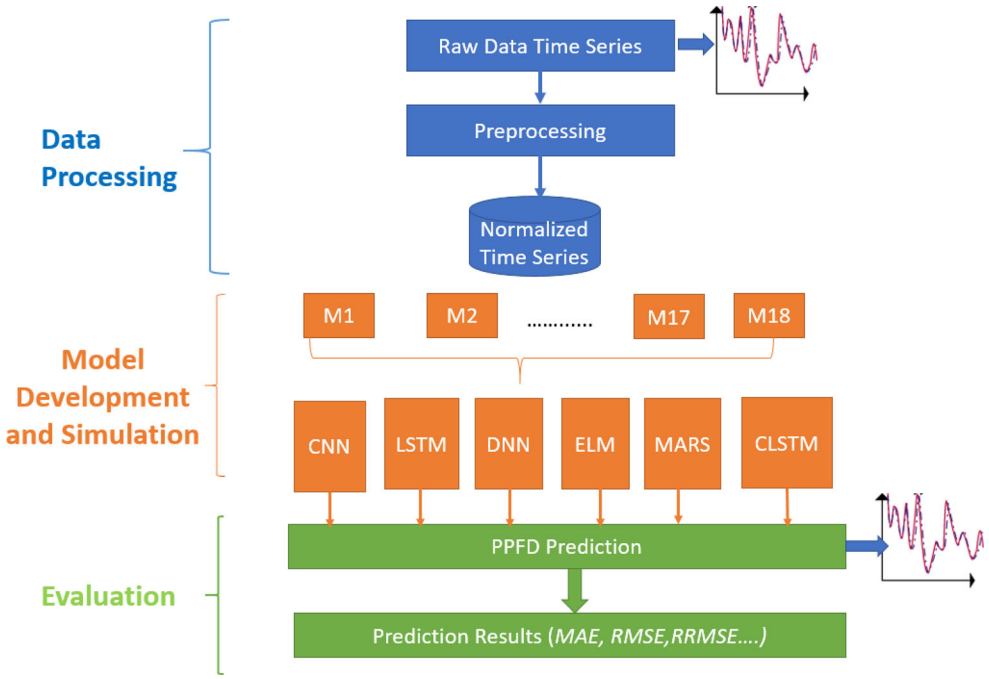


Fig. 6 Schematic diagram of the relevant steps in designing the CLSTM predictive model

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 $\frac{R_{av}}{B_{av}}, R_{sd}, RBC_{diff}, BRC_{diff}$ andTC 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

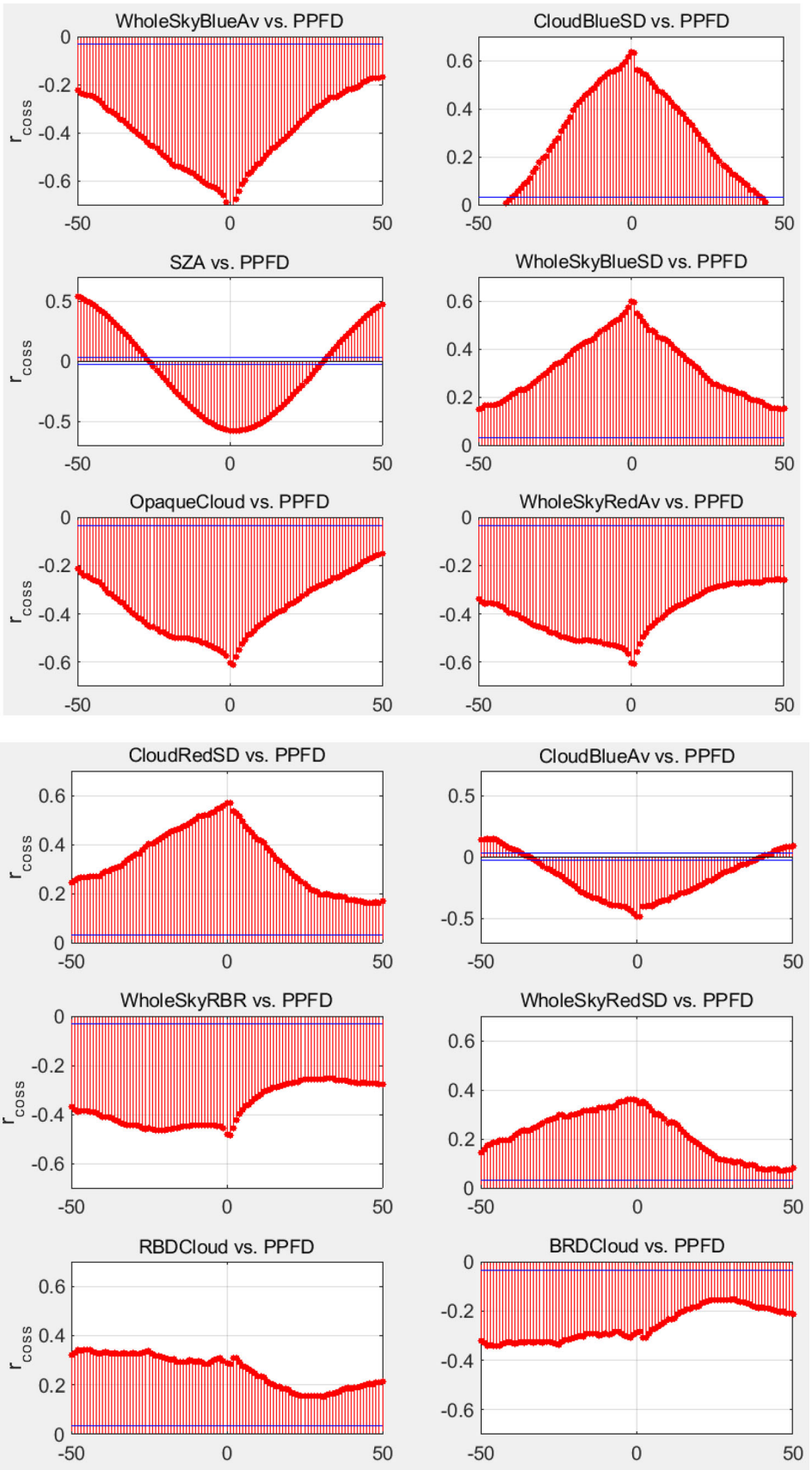
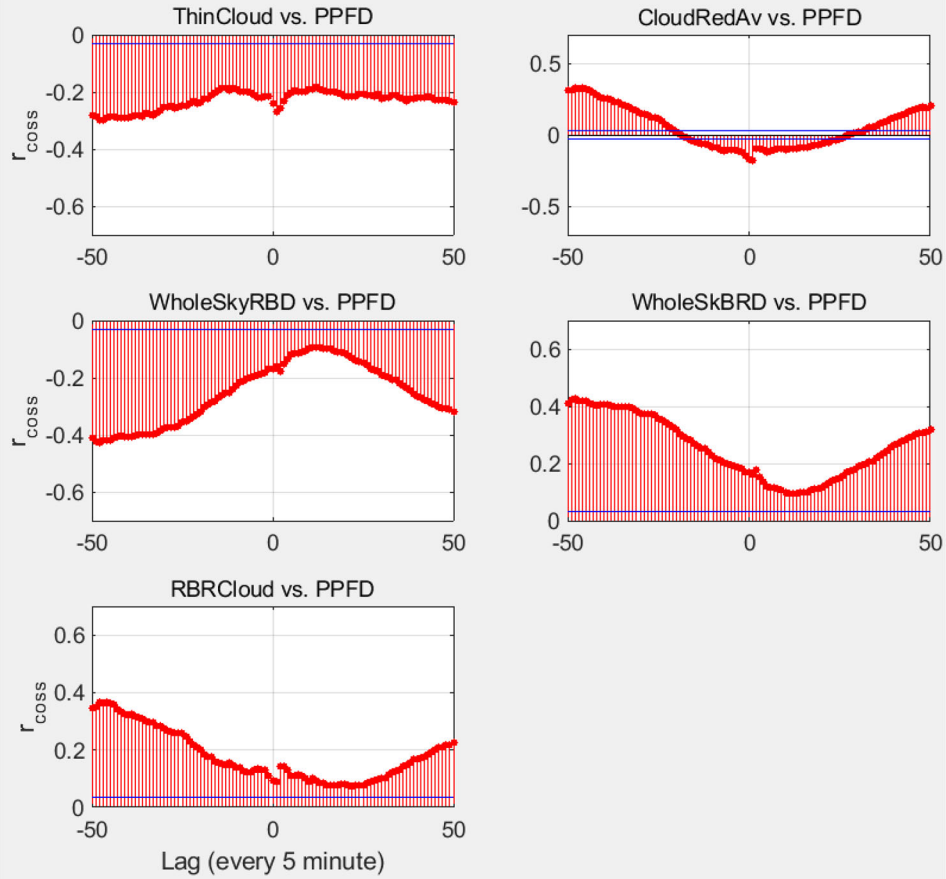


Fig. 7 Correlograms plotted to identify the degree of covariance between *PPFD* (i.e., the objective variable) and the 17 different cloud-image derived predictor variables within the CLSTM model's training phase. The *y*-axis shows cross-correlation coefficient, r_{cross} with blue line representing the level at the 95% confidence interval

Fig. 7 continued



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-min *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.

In congruence with previous results shown in Fig. 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 Fig. 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.

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-min *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 (Deo et al. 2019). 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 and 0.623), and the largest *RMSE / MAE* between 412.77 and 438.99/354.29 and 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 and 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.

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 and 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-min *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 and 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 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-min *PPDF* dataset.

In Fig. 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 many fold 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 CLM compared with CNN, LSTM, DNN, MARS and ELM. This is congruent with earlier results (Figs. 8, 9, 10) to demonstrate the CLSTM model as being the optimal choice to emulate the near real-time photosynthetic active radiation over a 5-min scale.

To further establish the veracity of the hybrid CLSTM model Fig. 12 shows the empirical cumulative distribution function (*ECDF*) of the error encountered in forecasting

Table 3 Statistical performance of models in testing phase utilising correlation coefficient (r), Nash–Sutcliffe efficiency (ENS), root-mean square error (RMSE), mean absolute error (MAE), including relative RMSE and MAE, and Legates and McCabes Index (LM)

Predictive model		Model number	r	*RMSE* (μmol m^{-2} s^{-1})	*MAE* (μmol m^{-2} s^{-1})	*RRMSE* (%)	*MAPE* (%)	E_{NS}	*LM*
Cloud properties-based models									
Objective model	CLSTM	M8	0.920	210.308	150.241	24.918	38.009	0.846	0.678
Benchmark models	ELM	M10	0.912	220.830	158.475	26.296	58.536	0.829	0.660
	LSTM	M13	0.912	221.358	159.942	26.227	45.719	0.829	0.657
	CNN	M11	0.898	236.714	168.593	28.046	38.365	0.804	0.639
	DNN	M5	0.900	233.498	171.089	27.805	61.498	0.809	0.632
	MARS	M12	0.894	241.255	183.109	28.729	56.207	0.796	0.607
SZA-based models									
Objective Model	CLSTM	M18	0.796	420.762	364.866	50.071	150.075	0.380	0.216
Benchmark models	ELM		0.795	438.990	360.877	51.984	176.717	0.369	0.227
	LSTM		0.794	428.180	366.961	50.704	156.829	0.360	0.214
	CNN		0.792	424.102	368.095	50.221	153.104	0.377	0.212
	DNN		0.623	418.370	355.870	49.819	134.277	0.386	0.236
	MARS		0.634	412.771	354.294	49.153	128.395	0.402	0.239

The test scenarios include: (i) the best model trained with cloud chromatic properties deduced from M1 to M17, and (ii) the baseline model using *SZA* as an input only (i.e., M18)

Deo et al. (2019)

Table 4 Skill score metric, $RMSE_{SS}$ for hybrid CLSTM with, versus hybrid CLSTM without cloud cover inputs based on only the *SZA* as an input variable

Model name	Skill score metric, $RMSE_{SS}$
CLSTM with cloud cover features [M8]	0.230
CLSTM without cloud features [M18]	− 0.540

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 Fig. 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-min *PPFD* dataset.

We further explore the influence of cloud cover variations on the prescribed objective model (i.e., CLSTM) in Fig. 13 where the 5-min forecasted *PPDF* valued averaged over the entire test dataset is shown *with* and *without* cloud 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, over the test period of 25-March-2013 to 31-March-2013. It is obvious that the hybrid CLSTM model utilising cloud cover-based input features yields the smallest mean error over the whole diurnal cycle, but this occurs with some degree of discrepancy. The CLSTM error follow a temporal pattern where the models register relatively larger percentage errors (see second panel), which occurs 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 or aerosol (Liou 1976; Aida 1977; Robinson 1977; Segal and Davis 1992; González and Calbó 2002). It is also possible that the CLSTM model is unable to capture enough features to predict the relatively smaller *PPFD* values in the

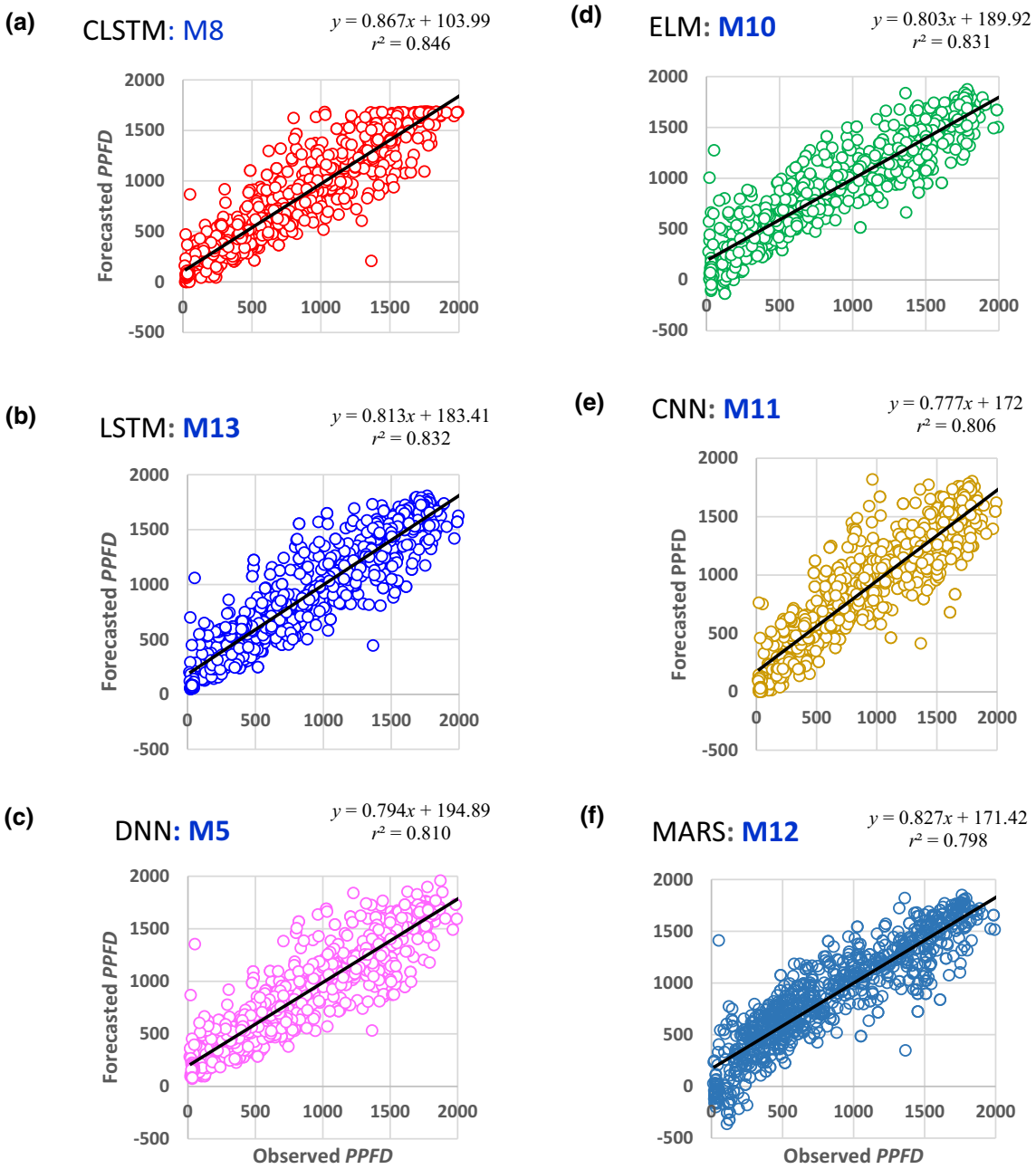


Fig. 8 Scatterplots of forecasted against observed *PPFD* values (μ mol of photons m^{-2} s^{-1}) emulated by the CLSTM model in the testing phase, compared with benchmark models. Only the optimal results (out of all designated models, M_1 to M_{17}) for each predictive algorithm based on best input combinations utilising cloud chromatic statistics and *SZA* as predictors, as per Table 2, are shown

morning and afternoon where the aerosol optical thickness is similar to the cloud scattering. In terms of the discrepancy in how the CLSTM model with, and the CLSTM model without cloud features performs, we note that there are instances where CLSTM without cloud cover performs better than with cloud cover (e.g., most of the timestamps from 84 to 108) corresponding to ∼ 2–4 pm. While the exact cause of this is not clear yet, it is possible that our segmentation process of sky images does not capture all of the dynamic cloud features that has a strong spatial and temporal signature, particularly if clouds are enhanced much later in the day for a location. Chen and Houze (Chen and Houze 1997) reports that the maximum occurrence of cold clouds in the afternoon follows a diurnal solar heating of the ocean surface and atmospheric boundary layer, whereas Theeuwes et al. (2019) provided observational evidence of a systematic enhancement of cloud cover in the afternoon and evening. They showed that initially, the day

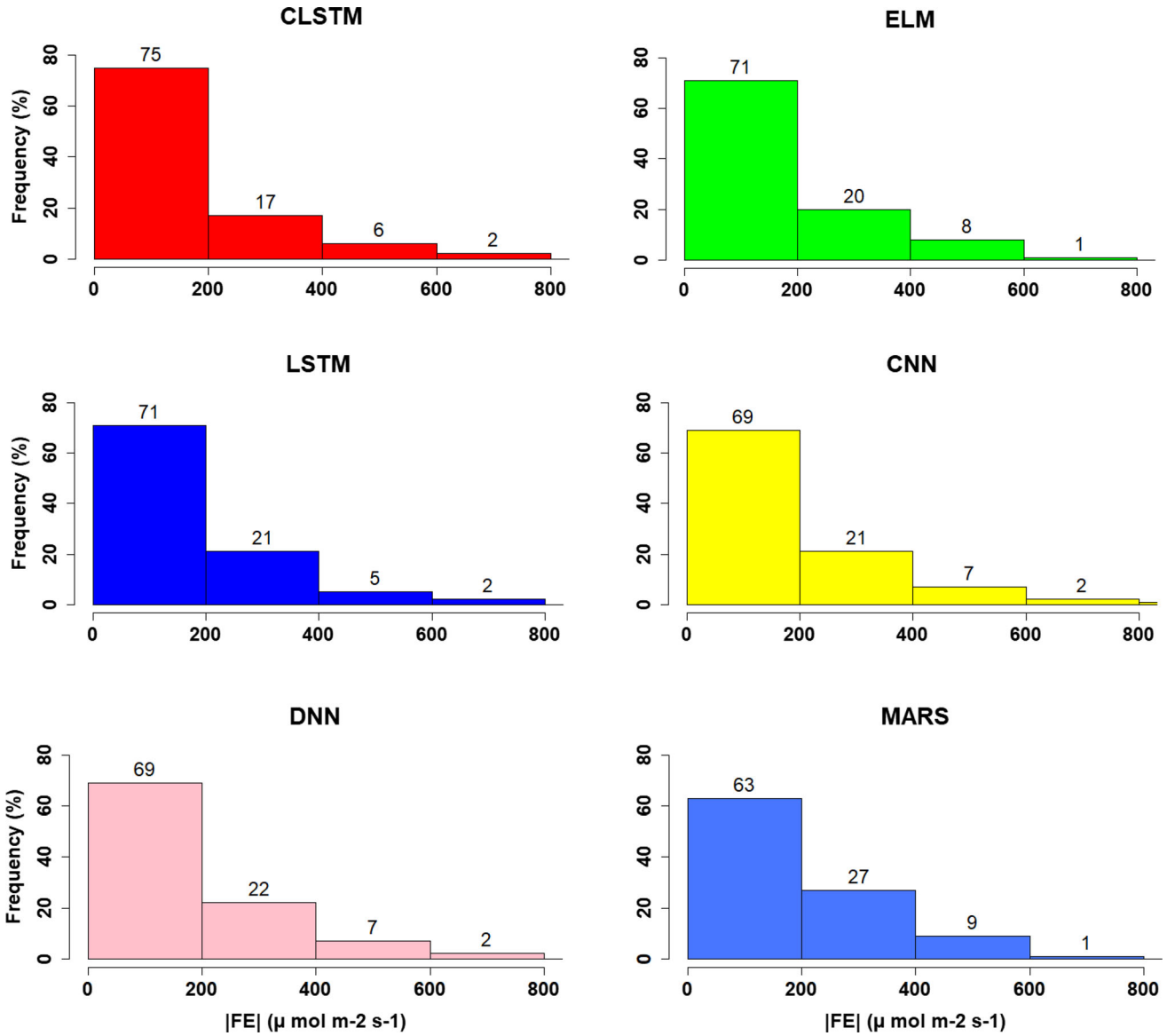


Fig. 9 The percentage frequency of the forecasted error generated by the CLSTM model against the deep learning (i.e., LSTM, CNN, DNN) and machine learning (ELM, MARS)-based models developed using best input combinations utilising cloud chromatic statistics and *SZA* as the predictors, in accordance with Table 2

is generally clear with a north-westerly flow so the cumulus clouds form during the morning and remain in afternoon, but the cloud-base height is higher during the day. Our cloud cover segmentation aimed to derive the overall R-B-G statistics, but it is unable to precisely consider cloud dynamics. When these cloud cover statistics (Tables 1, 2) are used in CLSTM model which are largely pixel-based representations rather than cloud height and topography, the model is perhaps unable to predict PPFD accurately, as the inputs could be redundant relative to the dynamic features of clouds not captured by our method. To address this issue, future studies on more accurate segmentation of sky images to derive features related to aerosols, water vapour, ozone, as well as considering cloud height, topography, optical depth etc. should be conducted where the proposed technique could be improved to better capture dynamical nature of cloud movements throughout the day. Nonetheless, the present analysis provides sufficient evidence of the important role of cloud cover conditions in modelling solar radiation and shows an important advancement in photosynthetic-active radiation prediction compared to earlier studies using the traditional (*SZA*) method.

5 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-min *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 (Lopez et al. 2001; Pankaew et al. 2014; Yu and Guo 2016; Deo et al. 2018) 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 min, our study has captured such influences on the ground-based photosynthetic active radiation at $\sim$ 5-min.

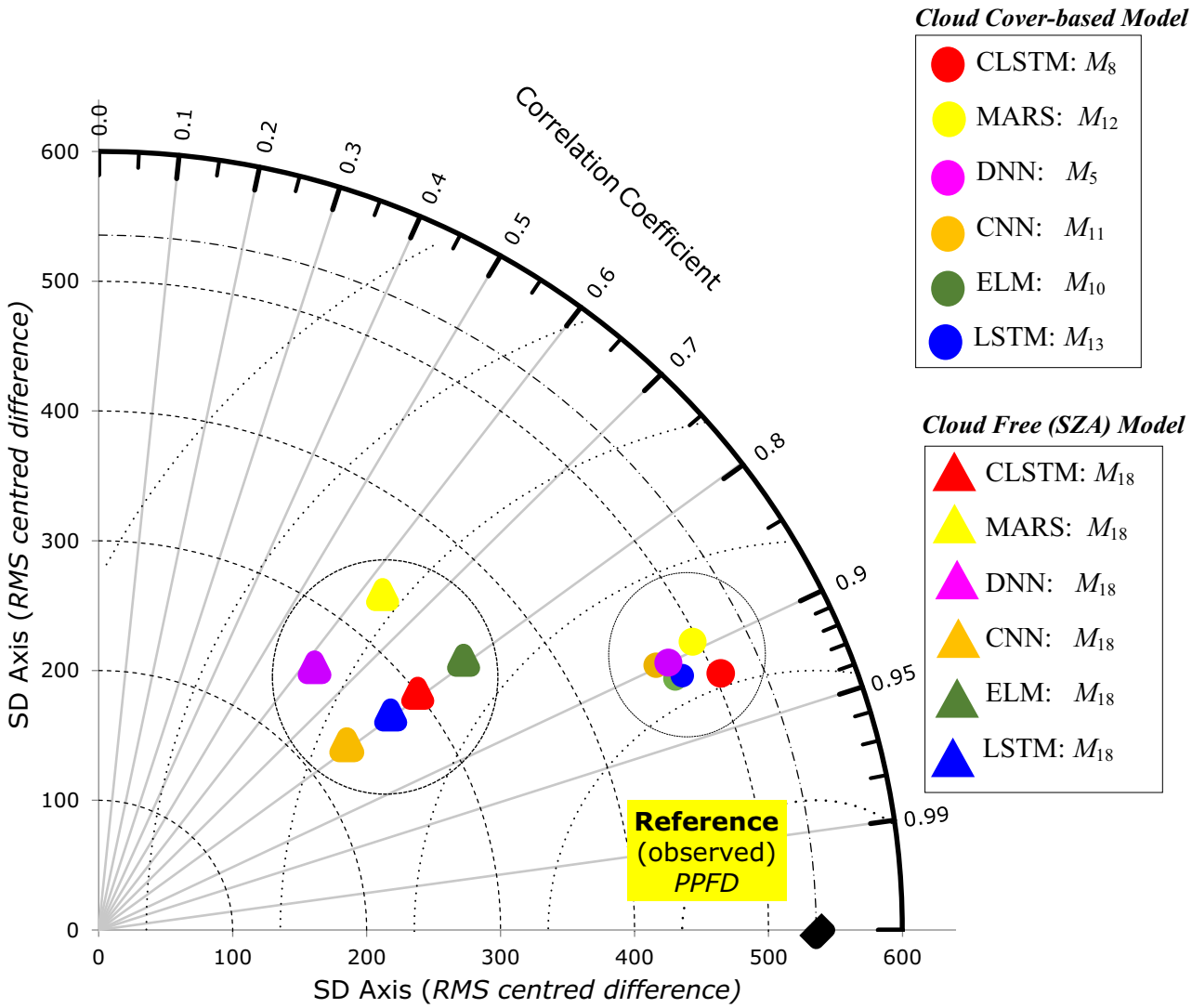


Fig. 10 Taylor diagram with a concise statistical summary of how well the simulations from the CLSTM predictive model match with the other models in terms of their correlations between observed and forecasted *PPFD*, root-mean-square difference and the ratio of the variance in testing phase. Only the most optimal model with cloud cover properties (i.e., M_8, M_{13}, M_{12}, M_5, M_{11} and M_{10}) and without cloud properties (i.e., M_{18} trained with *SZA* as input variable) are shown

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 (Wang et al. 2016). 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 × 7, by adopting internet of things (IoT) and wireless sensor networks, WSN (Ullo and Sinha 2020) 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-min 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 (García-Rodríguez et al. 2020). As the cost of total sky imagers (TSIs) can be insurmountable for most solar energy or biofuel generation farm locations, geostationary satellites such as Himawari 8 or 9, operating at roughly 10-min 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

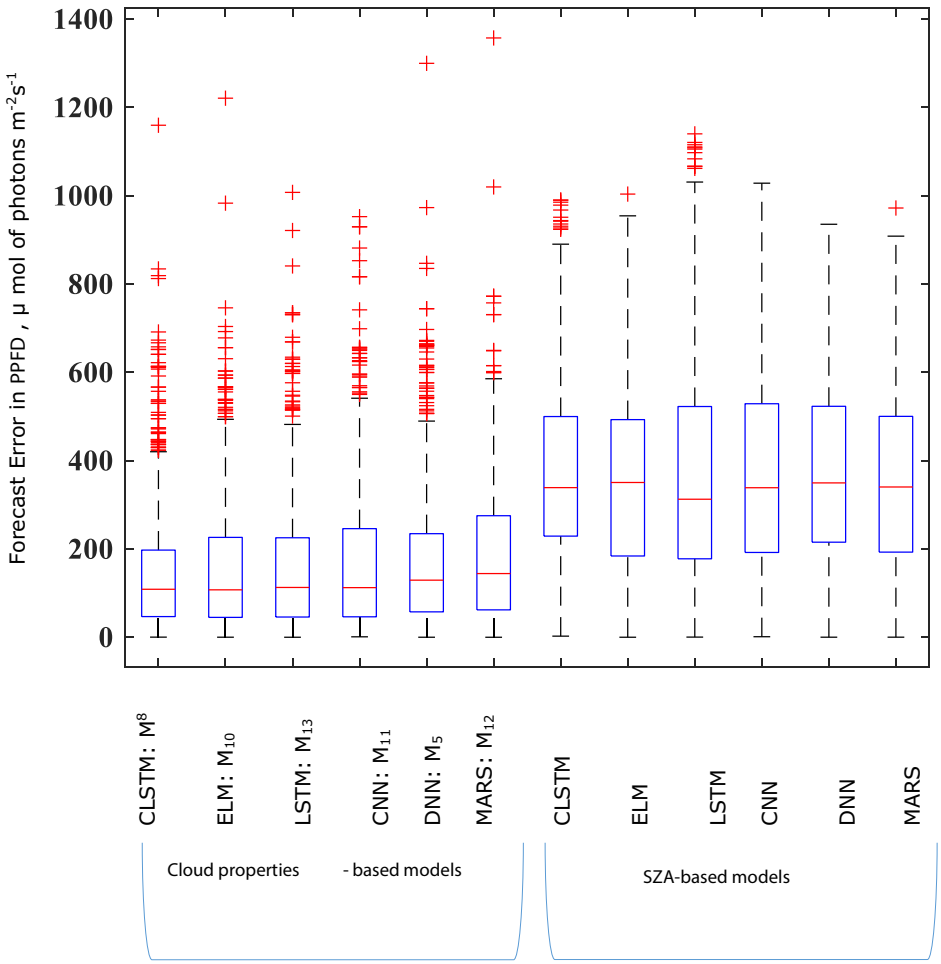


Fig. 11 Boxplot of the absolute forecasted error in *PPFD*: $|\mathrm{FE}| = |PPFD_i^{for} - PPFD_i^{obs}|$ within the testing phase using the cloud cover-based and the *SZA* only reference models. Figure legend should also indicate what the line, box, whiskers and points represent

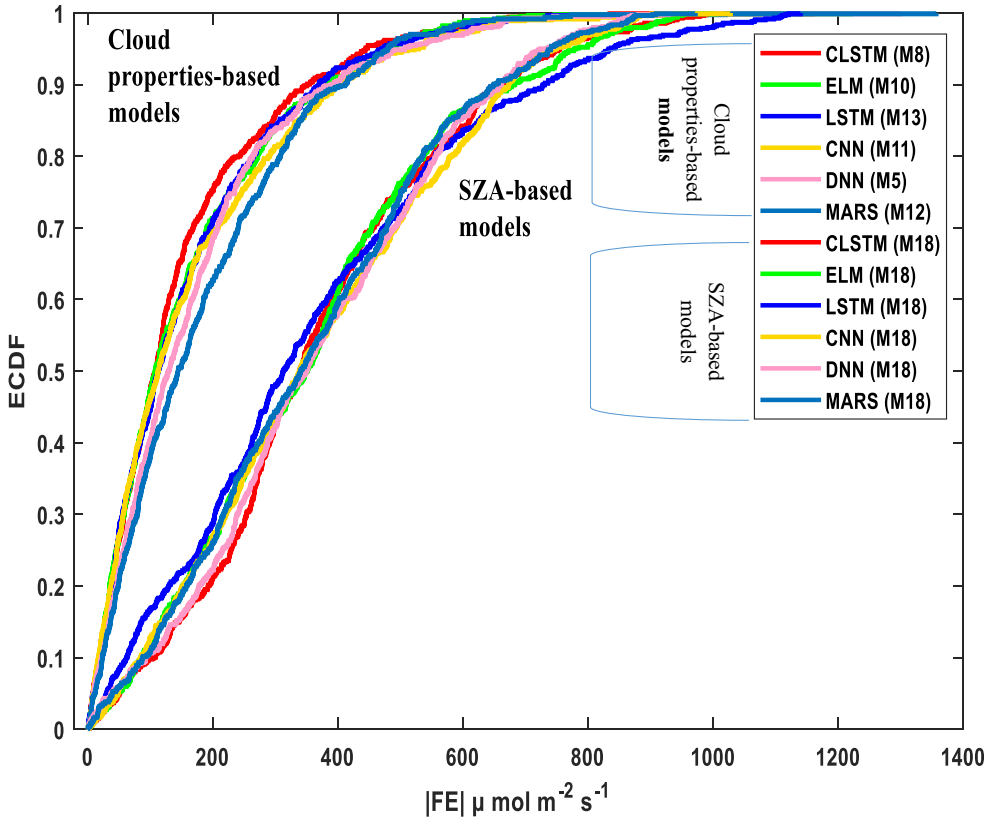


Fig. 12 Empirical cumulative distribution function (*ECDF*) of the *PPFD* forecasting error |FE| in the testing phase

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. (2017) 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-min) 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 (Zhen et al. 2017b).

6 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-min 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-min photosynthetic photon flux

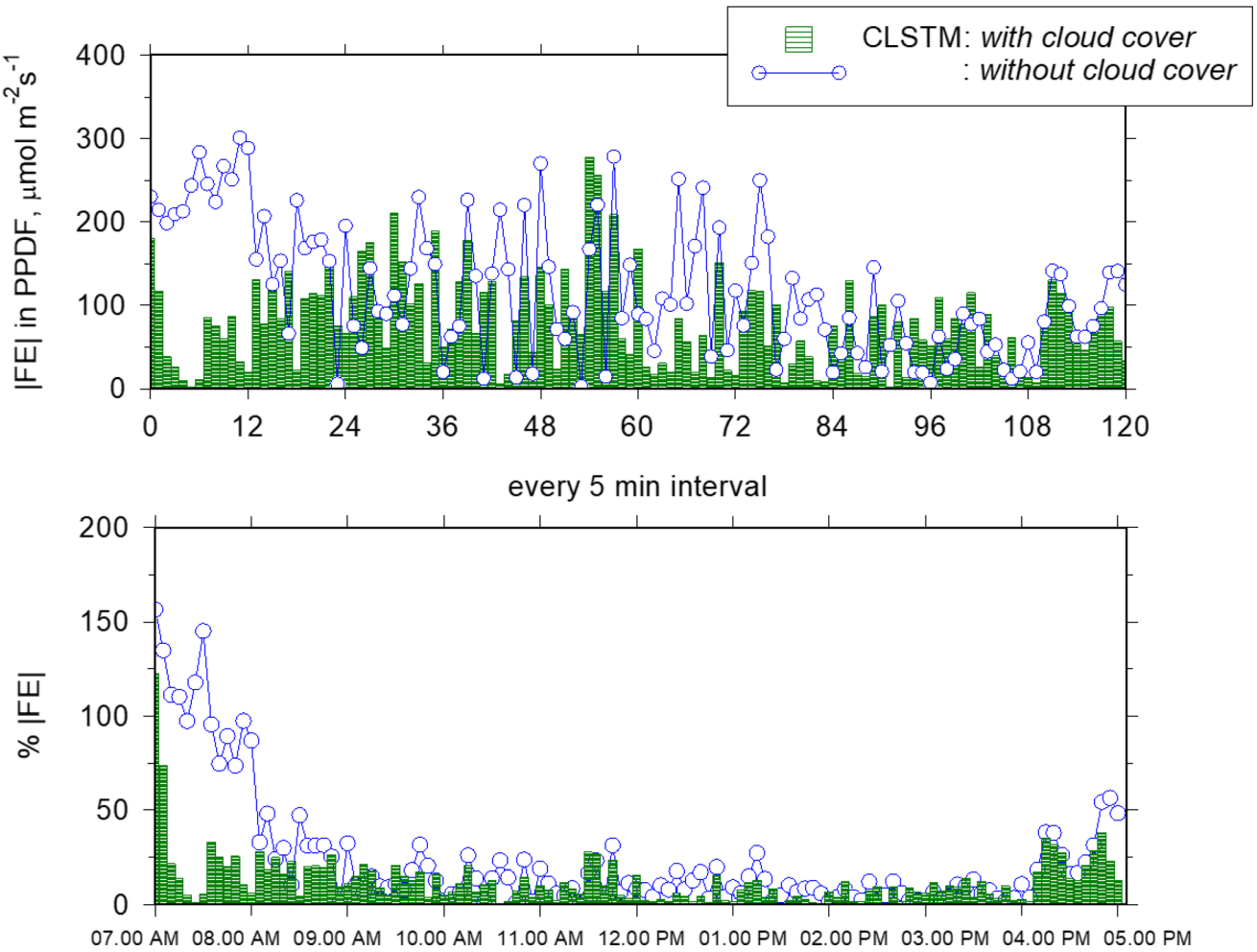


Fig. 13 The effect of cloud cover properties used as inputs for the CLSTM model with 5-min forecasted *PPFD* averaged over the entire testing dataset from 07.00 AM to 05.00 PM

density, *PPFD*). The CLSTM, verified to be highly superior in predicting 5-min *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-min *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 ($\sim$ 50.07%) compared to with cloud $\sim$

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. Improvements in CLSTM model's practical viability for other regions globally may also be made through its implementation on hourly, daily, and seasonal scales by sourcing real-time satellite and other remote sensing products. One such data product is the Himawari 8 & 9 satellite data that offers 10-min scan of the sky. Imagery from the advanced Himawari imager (AHI) instrument has finer spatial resolution (0.5–2 km, compared to 1–4 km for MTSAT) and precision (12–14-bit images, vs 10-bit for MTSAT). This satellite offers a higher temporal resolution with images recorded more frequently with one 'full disk' scan of the observable area every ten minutes (compared to hourly from MTSAT satellite). Our group at University of Southern Queensland is currently investigating the use of such data for real-time simulation of solar energy in Australia with results expected to be reported in future. Such testing of the proposed CLSTM predictive model, at high temporal resolutions, in a wider range of climates or seasons, in both remote and regional locations is a necessary step to help in direct harnessing of solar energy, biofuels, agricultural monitoring and supporting bio-physical sectors where global solar radiation, direct normal irradiance, global horizontal irradiance or photosynthetic-active radiation needs to be monitored for production of solar energy.

Appendix

See Tables 5, 6 and 7.

Table 5 The parameter search space for the hybrid deep learning-based (i.e., CLSTM) model architecture, including CNN, LSTM and DNN models

Designated model		Model hyperparameters	Search space for grid search for hyper-parameter optimization
Objective model	CLSTM	Filter1	[10, 20,50,100]
		Filter 2	[40,50,60,70,80]
		Filter 3	[20,10,30,5]
		LSTM cell units	[40,50,60,100]
		Epochs	[300,400,700]
		Activation function	[ReLU]
		Optimizer	[Adam]
		Batch Size	[1,5,10,20,50]
Benchmark models	CNN	Filter1	[20,50, 60,100]
		Filter 2	[40,50,60,70]
		Filter 3	[20,10,30,5]
		Activation function	[ReLU]
		Optimizer	[Adam]
		Epochs	[300,400,700]
		Batch Size	[1,5,10,20,50]
	LSTM	LSTM cell 1	[50, 60,100]
		LSTM cell 2	[40,50,60,70]
		LSTM cell 3	[20,10,30,5]
		Epochs	[300,400,700]
		Activation function	[ReLU]
		Optimizer	[Adam]
		Drop rate	[0.1,0.2]
		Batch Size	[1,5,10,20,50]
	DNN	Hidden neuron 1	[100,200,300,400,50]
		Hidden neuron 2	[20,30,40,50,60,70]
		Hidden neuron 3	[10,20,30,40,50]
		Hidden neuron 4	[5,6,7,8,12,15,18]
		Activation function	[ReLU]
		Optimizer	[Adam]
		Epochs	[100,200,500]
		Batch Size	[1,5,10,20,50]

ReLU and Adam stands for rectified linear units and adaptive moment estimation, respectively

Table 6 The optimal architecture of the deep learning CNN, LSTM and DNN models where the hyperparameters were obtained through a grid search procedure over all parameters specified in Table 5

Architecture of deep learning										
Designated model	Layer 1 (L1)	L1 activation function	Dropout percentage	Layer 2 (L2)	L2 activation function	Layer 3 (L3)	Layer 4 (L4)	L4 activation function	Batch size	Epochs
LSTM	50	ReLU	0.1	40	ReLU	20		ReLU	5	300
DNN	100	ReLU	0.1	60	ReLU	30	10	ReLU	10	100

	Convolution layers 1 (C1)	Convolution layers 2 (C2)	Convolutional layers 3 (C3)	Activation function	Pooling size	Padding	LSTM layer (L1)	L1 activation function	Dropout rate	batch size	Epochs
CLSTM	100	60	30	ReLU	2	Same	25	ReLU	0.1	10	400
CNN	60	40	5	ReLU	2	Same				5	700

Architecture of backpropagation (BP) algorithm for deep learning				
BP optimizers for deep learning model	Alpha, α	Epsilon, ε	Beta, β_1	Beta, β_2
Adaptive moment estimation, (Adam)	0.001	0.0000001	0.990	0.990

where

α = Learning rate, the proportion that weights are updated

ε = Is a very small number to prevent any division by zero in the model implementation

β_1 = The exponential decay rate for the 1st moment estimates

Table 7 The optimal architecture of the ELM and MARS models

Designated model	Dign parameters	
ELM	Number of layers	3
	Input neurons	Maximum value of 17 (i.e., for *SZA*, and cloud chromatic property-based inputs, as per Table 1)
	Inputs (i.e., predictor variables)	Cloud image statistical properties (associated with 5-min PPFD) and SZA as per Table 2
		$B_{av}, BC_{sd}, SZA, B_{sd}, OC, R_{av}, RC_{sd}, BC_{av}, \frac{R_{av}}{B_{av}}, R_{sd}, RBC_{diff}, BRC_{diff}, TC, RC_{av}, RBD_{diff}, RBD_{diff}, \frac{RC_{av}}{BC_{av}}$
	Hidden neurons	10, 20, …, 1000
	Output neurons	1 (*PPFD*) which is the measured photosynthetic radiation in μmol m^{-1} s^{-1}
	Activation functions	*Sigmoid, sine, hard limit, triangular basis, radial basis, tangent sigmoid, logarithmic sigmoid was the* optimal function
	ELM architecture	10–23–1 (input-hidden-output) determined iteratively by trial and error [*Model M10* (optimal)]
MARS	Number of basis functions	18
	Split threshold	0.05
	Spline type	'Cubic'
	Generalized cross validation	1.7751
	Optimal model equation	$y = -1.01e+03 + 0.0988 \times BF1 - 0.36 \times BF2 + 0.653 \times BF3 + 2.83 \times BF4 - 0.009 \times BF5 + 0.552 \times BF6 - 0.0768 \times BF7 - 8.47 \times BF8 - 1.64 \times BF9 + 1.25 \times BF10 + 2.32 \times BF11 + 0.000749 \times BF12 + 0.0102 \times BF13 - 0.13 \times BF14 - 946 \times BF15 + 943 \times BF16 + 946 \times BF17'$ [*Model M11*(optimal)]

Acknowledgements Data acquired were obtained from the University of Southern Queensland Solar Research Laboratory. The authors are grateful to Mr Kai Chen for some insightful discussions and Professors Rajendra Archarya and Prabal Barua for reading the original drafts of this paper.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

References

Agarap AF (2018) Deep learning using rectified linear units (relu). Accessed from https://arxiv.org/abs/1803.08375

Aida M (1977) Scattering of solar radiation as a function of cloud dimensions and orientation. J Quant Spectrosc 17(3):303–310

Ali M et al (2018a) An ensemble-ANFIS based uncertainty assessment model for forecasting multi-scalar standardized precipitation index. Atmos Res 207:155–180

Ali M et al (2018b) Multi-stage committee based extreme learning machine model incorporating the influence of climate parameters and seasonality on drought forecasting. Comput Electron Agric 152:149–165

Al-Musaylh MS et al (2018a) Short-term electricity demand forecasting with MARS, SVR and ARIMA models using aggregated demand data in Queensland, Australia. Adv Eng Inform 35:1–16

Al-Musaylh MS, Deo RC, and Li Y (2018b) Particle swarm optimized–support vector regression hybrid model for daily horizon electricity demand forecasting using climate dataset. In: E3S Web of Conferences. EDP Sciences

Al-Musaylh MS, Deo RC, Li Y (2020) Electrical energy demand forecasting model development and evaluation with maximum overlap discrete wavelet transform-online sequential extreme learning machines algorithms. Energies 13(9):2307

Antczak K (2019) On regularization properties of artificial datasets for deep learning. Accessed from https://arxiv.org/abs/1908.07005

ASCETC (1993) Criteria for evaluation of watershed models. J Irrig Drain Eng 119(3):429–442

Ayinde BO, Zurada JM (2017) Deep learning of constrained autoencoders for enhanced understanding of data. IEEE Trans Neural Netw Learn Syst 29(9):3969–3979

Barzegar R, Aalami MT, Adamowski J (2020) Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model. Stoch Environ Res Risk Assess 34:1–19

Batey M, Green R (2000) Geometrically effective cloud fraction for solar radiation. Atmos Res 55(2):115–129

Bengio Y, Simard P, Frasconi P (1994) Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw 5(2):157–166

Byrd J, Lipton Z (2019) What is the effect of importance weighting in deep learning? In: International Conference on Machine Learning. PMLR

Cai Y et al (2019) Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. Agric for Meteorol 274:144–159

Cai S, et al. (2019) Effective and efficient dropout for deep convolutional neural networks. Accessed from https://arxiv.org/abs/1904.03392

Chen SS, Houze RA Jr (1997) Diurnal variation and life-cycle of deep convective systems over the tropical Pacific warm pool. Q J R Meteorol Soc 123(538):357–388

Chen L et al (2008) MODIS-derived daily PAR simulation from cloud-free images and its validation. Sol Energy 82(6):528–534

Chen C-Y et al (2011) Cultivation, photobioreactor design and harvesting of microalgae for biodiesel production: a critical review. Biores Technol 102(1):71–81

Chen J et al (2018) Wind speed forecasting using nonlinear-learning ensemble of deep learning time series prediction and extremal optimization. Energy Convers Manage 165:681–695

Chimmula VKR, Zhang L (2020) Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos Solitons Fractals 135:109864

Chollet F (2017) Keras (2015)

Chollet F (2018) Keras: The python deep learning library. Astrophysics Source Code Library

Crane-Droesch A (2018) Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environ Res Lett 13:114003

Deo RC, Wen X, Feng Q (2016) A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset. Appl Energy 168:568–593

Deo RC et al (2017) Very short-term reactive forecasting of the solar ultraviolet index using an extreme learning machine integrated with the solar zenith angle. Environ Res 155:141–166

Deo RC et al (2018) Adaptive neuro-fuzzy inference system integrated with solar zenith angle for forecasting sub-tropical photosynthetically active radiation. Food Energy Secur. https://doi.org/10.1002/fes3.151

Deo RC et al (2019) Adaptive neuro-fuzzy inference system integrated with solar zenith angle for forecasting sub-tropical photosynthetically active radiation. Food Energy Secur 8(1):e00151

Dev S et al (2016) Rough-set-based color channel selection. IEEE Geosci Remote Sens Lett 14(1):52–56

Dodge J, et al. (2020) Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. Accessed from https://arxiv.org/abs/2002.06305

Feng P et al (2019) Incorporating machine learning with biophysical model can improve the evaluation of climate extremes impacts on wheat yield in south-eastern Australia. Agric for Meteorol 275:100–113

Gao B et al (2020) Hourly forecasting of solar irradiance based on CEEMDAN and multi-strategy CNN-LSTM neural networks. Renew Energy 162:1665–1683

Garbin C, Zhu X, Marques O (2020) Dropout vs. batch normalization: an empirical study of their impact to deep learning. Multimed Tools Appl 79:1–39

García-Rodríguez A et al (2020) Photosynthetic active radiation, solar irradiance and the CIE standard sky classification. Appl Sci 10(22):8007

Ghimire S et al (2018) 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 Sens Environ 212:176–198

Ghimire S et al (2019) Global solar radiation prediction by ANN integrated with European Centre for medium range weather forecast fields in solar rich cities of Queensland Australia. J Clean Prod 216:288–310

Ghimire S et al (2019a) Deep solar radiation forecasting with convolutional neural network and long short-term memory network algorithms. Appl Energy 253:113541

Ghimire S et al (2019b) 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. Renew Sustain Energy Rev 113:109247

Ghonima M et al (2012) A method for cloud detection and opacity classification based on ground based sky imagery. Atmos Meas Tech 5(11):2881–2892

Gill D, Ming T, and Ouyang W (2017) Improving the Lake Erie HAB tracker: a forecasting & decision support tool for harmful algal blooms

González J, Calbó J (2002) Modelled and measured ratio of PAR to global radiation under cloudless skies. J Agric for Meteorol 110(4):319–325

Grant RH, Heisler GM (1997) Obscured overcast sky radiance distributions for ultraviolet and photosynthetically active radiation. J Appl Meteorol 36(10):1336–1345

Graves A (2013) Generating sequences with recurrent neural networks. Accessed from https://arxiv.org/abs/1308.0850

Gu L et al (2002) Advantages of diffuse radiation for terrestrial ecosystem productivity. J Geophys Res 107(D6):ACL 2-1-ACL 2-23

Gumma MK et al (2020) 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 Sens 57(3):302–322

Han J et al (2020) Prediction of winter wheat yield based on multi-source data and machine learning in China. Remote Sens 12(2):236

Hanan N et al (1995) Estimation of absorbed photosynthetically active radiation and vegetation net production efficiency using satellite data. Agric for Mateorol 76(3–4):259–276

Hengl T et al (2018) Global mapping of potential natural vegetation: an assessment of machine learning algorithms for estimating land potential. PeerJ 6:e26811v2

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780

Hohman F et al (2019) S ummit: Scaling deep learning interpretability by visualizing activation and attribution summarizations. IEEE Trans Visual Comput Graphics 26(1):1096–1106

Holdmann C, Schmid-Staiger U, Hirth T (2019) Outdoor microalgae cultivation at different biomass concentrations—assessment of different daily and seasonal light scenarios by modeling. Algal Res 38:101405

Hong Y-Y, Satriani TRA (2020) Day-ahead spatiotemporal wind speed forecasting using robust design-based deep learning neural network. Energy 209:118441

Igoe DP, Parisi AV, Downs NJ (2019) Cloud segmentation property extraction from total sky image repositories using Python. Instrum Sci Technol 47(5):522–534

Jaiswal S, Mehta A, Nandi G (2018) 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). IEEE

Jaseena KU, Kovoor BC (2021) Decomposition-based hybrid wind speed forecasting model using deep bidirectional LSTM networks. Energy Convers Manage 234:113944

Jebar MAA et al (2020) Influence of clouds on OMI satellite total daily UVA exposure over a 12-year period at a southern hemisphere site. Int J Remote Sens 41(1):272–283

Jekabsons G (2013) Adaptive regression splines toolbox for Matlab/Octave. Version 1:72

Jiang H et al (2020) Surface diffuse solar radiation determined by reanalysis and satellite over East Asia: evaluation and comparison. Remote Sens 12(9):1387

Johnson D, et al. (2015) A new quantum sensor for measuring photosynthetically active radiation. In: AGU Fall Meeting Abstracts

Kamir E, Waldner F, Hochman Z (2020) Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods. ISPRS J Photogramm Remote Sens 160:124–135

Ketkar N (2017) Introduction to keras. Deep learning with python. Springer, Cham, pp 97–111

Kim T-Y, Cho S-B (2019) Predicting residential energy consumption using CNN-LSTM neural networks. Energy 182:72–81

Konasani VR, Kadre S (2021) Machine learning and deep learning using python and tensorflow. McGraw-Hill Education, New York

Kooperberg C, Clarkson DB (1997) Hazard regression with interval-censored data. Biometrics 53:1485–1494

Kumar M et al (2018) Rapid and efficient genetic transformation of the green microalga *Chlorella vulgaris*. J Appl Phycol 30(3):1735–1745

Kuo C-CJ (2016) Understanding convolutional neural networks with a mathematical model. J vis Commun Image Represent 41:406–413

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

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444

Lee W et al (2018) Forecasting solar power using long-short term memory and convolutional neural networks. IEEE Access 6:73068–73080

Li Q, Lu W, Yang J (2011) A hybrid thresholding algorithm for cloud detection on ground-based color images. J Atmos Oceanic Tech 28(10):1286–1296

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

Li M, Soltanolkotabi M, Oymak S (2020b) Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In: International Conference on Artificial Intelligence and Statistics. PMLR

Liou K-N (1976) On the absorption, reflection and transmission of solar radiation in cloudy atmospheres. J Atmos Sci 33(5):798–805

Liu M, Zhang J, Xia X (2021) Evaluation of multiple surface irradiance-based clear sky detection methods at Xianghe—a heavy polluted site on the North China Plain. Atmos Oceanic Sci Lett 14(2):100016

Long CN et al (2006) Retrieving cloud characteristics from ground-based daytime color all-sky images. J Atmos Oceanic Tech 23(5):633–652

Lopez G et al (2001) Estimation of hourly global photosynthetically active radiation using artificial neural network models. Agric for Meteorol 107(4):279–291

Lozano IL et al (2021) Aerosol radiative effects in photosynthetically active radiation and total irradiance at a Mediterranean site from an 11-year database. Atmos Res 255:105538

Ma L, Tian S (2020) A hybrid CNN-LSTM model for aircraft 4D trajectory prediction. IEEE Access 8:134668–134680
Mahsereci M, et al. (2017) Early stopping without a validation set. Accessed from https://arxiv.org/abs/2107.12972
McCree K (1973) The measurement of photosynthetically active radiation. Sol Energy 15(1):83–87
Meka R, Alaeddini A, Bhaganagar K (2021) A robust deep learning framework for short-term wind power forecast of a full-scale wind farm using atmospheric variables. Energy 221:119759
Michalsky JJ (1988) The astronomical almanac's algorithm for approximate solar position (1950–2050). Sol Energy 40(3):227–235
Moler C (2000) Matlab incorporates LAPACK. Cleve's Corner, MATLAB News&Notes
Nwankpa C, et al. (2018) Activation functions: Comparison of trends in practice and research for deep learning. Accessed from https://arxiv.org/abs/1811.03378
Oh SL et al (2018) Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats. Comput Biol Med 102:278–287
Pankaew P, et al. (2014) Estimating photosynthetically active radiation using an artificial neural network. In: 2014 International Conference and Utility Exhibition on Green Energy for Sustainable Development (ICUE). IEEE
Parisi AV, Sabburg J, Kimlin MG (2004) Scattered and filtered solar UV measurements, vol 17. Springer, Dordrecht
Park S, Nguyen THT, Jin E (2019) Improving lipid production by strain development in microalgae: strategies, challenges and perspectives. Bioresour Technol 292:121953
Patil S, Pandit R, Lali A (2017) Responses of algae to high light exposure: prerequisite for species selection for outdoor cultivation. J Algal Biomass Utln 8:75–83
Pedregosa F et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12(Oct):2825–2830
Prasad S, Deo RC, Downs N, Igoe D, Parisi AV, Soar J (2022) Cloud affected solar UV predictions with three-phase wavelet hybrid convolutional long short-term memory network multi-step forecast system. IEEE Access. https://doi.org/10.1109/ACCESS.2022.3153475
Proskurina S et al (2019) Global biomass trade for energy—Part 2: production and trade streams of wood pellets, liquid biofuels, charcoal, industrial roundwood and emerging energy biomass. Biofuels Bioprod Biorefin 13(2):371–387
Pruvost J et al (2016) Microalgae culture in building-integrated photobioreactors: biomass production modelling and energetic analysis. Chem Eng J 284:850–861
Ramanna L, Rawat I, Bux F (2017) Light enhancement strategies improve microalgal biomass productivity. Renew Sustain Energy Rev 80:765–773
Rice L, Wong E, Kolter Z (2020) Overfitting in adversarially robust deep learning. In: International Conference on Machine Learning. PMLR
Robinson PJ (1977) Measurements of downward scattered solar radiation from isolated cumulus clouds. J Appl Meteorol 16(6):620–625
Rocha AV et al (2021) Solar position confounds the relationship between ecosystem function and vegetation indices derived from solar and photosynthetically active radiation fluxes. Agric for Meteorol 298–299:108291
Ryu Y et al (2018) MODIS-derived global land products of shortwave radiation and diffuse and total photosynthetically active radiation at 5km resolution from 2000. Remote Sens Environ 204:812–825
Sabburg JM (2000) Quantification of cloud around the sun and its correlation with global UV measurement. Queensland University of Technology
Sabburg J, Long CN (2004) Improved sky imaging for studies of enhanced UV irradiance. Atmos Chem Phys 4(11/12):2543–2552
Sabburg J, Wong J (1999) Evaluation of a ground-based sky camera system for use in surface irradiance measurement. J Atmos Oceanic Tech 16(6):752–759
Sainath TN, et al. (2015) Convolutional, long short-term memory, fully connected deep neural networks. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE
Sato M et al (2018) Application of deep learning to the classification of images from colposcopy. Oncol Lett 15(3):3518–3523
Segal M, Davis J (1992) The impact of deep cumulus reflection on the ground-level global irradiance. J Appl Meteorol 31(2):217–222
Siqueira SF et al (2020) Mapping the performance of photobioreactors for microalgae cultivation: geographic position and local climate. J Chem Technol Biotechnol 95(9):2411–2420
Slade R, Bauen A (2013) Micro-algae cultivation for biofuels: cost, energy balance, environmental impacts and future prospects. Biomass Bioenerg 53:29–38
Slater D, Long C, and Tooman T (2001) Total sky imager/whole sky imager cloud fraction comparison. In: Eleventh ARM Science Team Meeting Proceedings, Atlanta, Georgia
Song X et al (2020) Time-series well performance prediction based on long short-term memory (LSTM) neural network model. J Pet Sci Eng 186:106682
Swietojanski P, Ghoshal A, Renals S (2014) Convolutional neural networks for distant speech recognition. IEEE Signal Process Lett 21(9):1120–1124
Tang W et al (2017) An efficient algorithm for calculating photosynthetically active radiation with MODIS products. Remote Sens Environ 194:146–154
Theeuwes NE et al (2019) Persistent cloud cover over mega-cities linked to surface heat release. J Npj Clim Atmos Sci 2(1):1–6
Ullah A et al (2017) Action recognition in video sequences using deep bi-directional LSTM with CNN features. IEEE Access 6:1155–1166
Ullah FUM et al (2019) Short-term prediction of residential power energy consumption via CNN and multi-layer bi-directional LSTM networks. IEEE Access 8:123369–123380
Ullo SL, Sinha GJS (2020) Advances in smart environment monitoring systems using iot and sensors. Sensors 20(11):3113
van Der Walt S, Colbert SC, Varoquaux G (2011) The NumPy array: a structure for efficient numerical computation. Comput Sci Eng 13(2):22–30
van der Walt S, Schönberger 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:e453
Vidal A, Kristjanpoller W (2020) Gold volatility prediction using a CNN-LSTM approach. Expert Syst Appl 157:113481
Vuppaladadiyam AK et al (2018) Microalgae cultivation and metabolites production: a comprehensive review. Biofuels Bioprod Biorefin 12(2):304–324
Wagner VS (1995) Uebertragung strahlungsreleveanter wetterinformation aus punktuellen PAR- sensordaten in groesser versuchsfaechenanlagen mit hifle hemisphaerisher fotos. Allg Forst 167(1–2):34–40
Wang J, et al. (2016) 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)
Wang L et al (2016) Modeling and comparison of hourly photosynthetically active radiation in different ecosystems. Renew Sustain Energy Rev 56:436–453

Wang F et al (2018) Wavelet decomposition and convolutional LSTM networks based improved deep learning model for solar irradiance forecasting. Appl Sci 8(8):1286
Wang K, Qi X, Liu H (2019) A comparison of day-ahead photovoltaic power forecasting models based on deep learning neural network. Appl Energy 251:113315
Wu Q, Lin H (2019) Daily urban air quality index forecasting based on variational mode decomposition, sample entropy and LSTM neural network. Sustain Cities Soc 50:101657
Xie H, Zhang L, Lim CP (2020) Evolving CNN-LSTM models for time series prediction using enhanced grey wolf optimizer. IEEE Access 8:161519–161541
Yadav A, Jha CK, Sharan A (2020) Optimizing LSTM for time series prediction in Indian stock market. Procedia Comput Sci 167:2091–2100
Yu X, Guo X (2016) Hourly photosynthetically active radiation estimation in Midwestern United States from artificial neural networks and conventional regressions models. Int J Biometeorol 60(8):1247–1259
Zang H et al (2020) Short-term global horizontal irradiance forecasting based on a hybrid CNN-LSTM model with spatiotemporal correlations. Renew Energy 160:26–41
Zareipour H, Bhattacharya K, and Canizares C (2006) Forecasting the hourly Ontario energy price by multivariate adaptive regression splines. In: 2006 IEEE Power Engineering Society General Meeting. IEEE
Zhang X et al (2016) Template-oriented synthesis of monodispersed SnS2@SnO2 hetero-nanoflowers for Cr(VI) photoreduction. Appl Catal B 192:17–25
Zhang Q et al (2018a) An adaptive dropout deep computation model for industrial IoT big data learning with crowdsourcing to cloud computing. IEEE Trans Ind Inform 15(4):2330–2337
Zhang Y-D et al (2018b) Voxelwise detection of cerebral microbleed in CADASIL patients by leaky rectified linear unit and early stopping. Multimed Tools Appl 77(17):21825–21845
Zhang Y et al (2020) Genetic transformation of tribonema minus, a eukaryotic filamentous oleaginous yellow-green alga. Int J Mol Sci 21(6):2106
Zhao R et al (2017) Learning to monitor machine health with convolutional bi-directional LSTM networks. Sensors 17(2):273
Zhen Z et al (2017a) Research on a cloud image forecasting approach for solar power forecasting. Energy Procedia 142:362–368
Zhen Z et al (2017b) Research on a cloud image forecasting approach for solar power forecasting. Energy Procedia 142:362–368
Zheng Y, Zhang M, and Wu B (2016a) Using high spatial and temporal resolution data blended from SPOT-5 and MODIS to map biomass of summer maize. In: 2016a Fifth International Conference on Agro-Geoinformatics (Agro-Geoinformatics)
Zheng Y et al (2016b) Mapping winter wheat biomass and yield using time series data blended from PROBA-V 100- and 300-m S1 products. Remote Sens 8(10):824

Authors and Affiliations

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]

✉ Ravinesh C. Deo
ravinesh.deo@usq.edu.au;
https://staffprofile.usq.edu.au/profile/ravinesh-deo

Richard H. Grant
rgrant@purdue.edu

Ann Webb
ann.webb@manchester.ac.uk;
https://www.research.manchester.ac.uk/portal/ann.webb.html

Sujan Ghimire
sujan.ghimire@usq.edu.au;
https://staffprofile.usq.edu.au/Profile/Sujan-Ghimire

Damien P. Igoe
damienpaul@gmail.com

Nathan J. Downs
Nathan.Downs@usq.edu.au;
https://staffprofile.usq.edu.au/Profile/Nathan-Downs

Mohanad S. Al-Musaylh
mohanad.al-musaylh@stu.edu.iq

Alfio V. Parisi
Alfio.Parisi@usq.edu.au;
https://staffprofile.usq.edu.au/Profile/Alfio-Parisi

Jeffrey Soar
Jeffrey.Soar@usq.edu.au;
https://staffprofile.usq.edu.au/Profile/Jeffrey-Soar

[1] School of Mathematics, Physics and Computing, University of Southern Queensland, Toowoomba, QLD 4300, Australia

[2] Department of Agronomy, Purdue University, West Lafayette, IN, USA

[3] Department of Earth and Environmental Sciences, Faculty of Science and Engineering, University of Manchester, Manchester M13 9PL, UK

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

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