Contents lists available at ScienceDirect

Cleaner Engineering and Technology

journal homepage: www.journals.elsevier.com/cleaner-engineering-and-technology

Machine learning models to quantify the influence of PM_{10} aerosol concentration on global solar radiation prediction in South Africa

Tamara Rosemary Govindasamy *, Naven Chetty

Department of Physics, School of Chemistry and Physics, University of KwaZulu-Natal, Pietermaritzburg, South Africa

ARTICLE INFO

Keywords:
Solar energy prediction
Machine learning techniques
Air pollutant concentration
Urban pollution
Artificial neural networks

ABSTRACT

Solar prediction models are essential in developing countries such as South Africa as most meteorological stations are unable to consistently measure this quantity. Quantification of solar radiation is of great significance for the adoption and application of renewable energy systems. Machine learning models are infamous for their high prediction capacity and low input requirements. This study's foremost intention was to investigate the efficacy of using the most widespread machine learning techniques for solar radiation estimation across South Africa, by introducing PM_{10} air pollutant concentration to generalized, readily available meteorological datasets. While assessing the performance of various input parameter configurations, the techniques which were evaluated include; Artificial Neural Network (ANN), Support Vector Regression (SVR), General Regression Neural Network (GRNN), and Random Forest (RF), of which ANNs proved the most appropriate for predicting global solar radiation as indicated by the high correlation coefficients ($R^2 = 0.99852$) and low prediction errors ($RMSE = 0.22627$). The results described in this work indicate that machine learning models performed excellently for hybrid models as opposed to empirical models established in South Africa. ANN models which include PM_{10} concentration data profoundly improved the performance of relative humidity models and reduced overall error measures for models of various input parameter configurations. In addition, a poor correlation between PM_{10} concentration and air temperature was observed. This work suggests that the use of generalized, hybrid ANN models to predict solar radiation across South Africa are more functional than empirical modeling and this is indicated by the high prediction accuracy and low computational effort. The suggested model suffices as an accurate prediction model which will allow for a holistic understanding of the solar capacity available across this country while encouraging the implementation and investigation of sustainable, renewable energy technologies.

1. Introduction

The critical inquiry into the availability and accessibility of renewable energy resources is consistently evolving, especially in developing countries with emerging economies like South Africa. The rapidly escalating energy demand and inconsistent energy supply in this country is a major challenge which conventional energy producers are unable to reconcile. From the database of global fossil fuel consumption and production, South Africa still produces the largest amount of greenhouse gas emission on the African continent (Ritchie and Roser, 2017). Thus, exploration into energy production from renewable energy sources is vital for the stabilization of energy supply and essential to reduce the impact of greenhouse gas emissions and consequently global warming.

While being the 7th largest coal producer and 5th largest coal exporter globally, South Africa is still facing a dire energy crisis, with the current grid infrastructure taking strain while trying to meet day-to-day demands, and a significant proportion of the population not having access to electricity (Maleki et al., 2017). Energy security and sustainability remains one of the biggest challenges to be addressed by the South African government. The initial crisis began in late 2007, where a state of emergency called upon the implementation of the load shedding program (2008) to avoid the country facing a total black-out (Da Silva et al., 2016). At the time, this was the only manner in which ESKOM (National electricity utility) could meet the current electricity demand. Twelve years later and the load shedding schedule is still a reality. The extremity of the situation was one which had been brewing for some time and some of the contributing factors which led to this include; shortfalls in the management of the energy system, the increased demand of the continuously growing South African population, increased energy consumption from industries such as mining, agriculture and transportation and little

* Corresponding author.
E-mail addresses: 209540135@stu.ukzn.ac.za, tamz.r.g@gmail.com (T.R. Govindasamy), chettyn3@ukzn.ac.za (N. Chetty).

https://doi.org/10.1016/j.clet.2021.100042
Received 10 November 2020; Received in revised form 19 January 2021; Accepted 20 January 2021

Table 1
Geographical details of study sites.

Location	Province	Land use classification	Latitude [°S]	Longitude [°E]	Elevation [m]
Bloemfontein	Free State	Commercial/Residential/Metropolitan	29.1030	26.3263	1400
Cape Town	Western Cape	Commercial/Industrial/Port/Metropolitan/Residential	33.9630	18.4194	670
Durban	KwaZulu-Natal	Commercial/Industrial/Port/Metropolitan/Residential	29.9650	30.4849	670
Johannesburg	Gauteng	Commercial/Industrial/Metropolitan/Residential	26.1430	28.3971	1800
Pietermaritzburg	KwaZulu-Natal	Commercial/Industrial/Residential	29.6270	30.4062	750
Groblershoop	Northern Cape	Agricultural/Residential/Industrial	28.89864	22.00107	871
Nelspruit	Mpumalanga	Commercial/Industrial/Residential/Agricultural	25.45458	30.97157	673
Dendron	Limpopo	Residential/Agricultural	23.4042	29.32828	720
Mpofu	Eastern Cape	Residential/Agricultural	32.56979	26.6995	680
Delareyville	North West	Agricultural/Residential	26.7246	25.32766	1379

to no investment in the power infrastructure by government over the past decades (Maleki et al., 2017). The energy sector has been neglected for the past few decades and existing infrastructure now proves inadequate and unable to manage the current load. The above is a brief description of the conditions which have given rise to escalated energy costs in South Africa. In addition, there remain countless households which have no access to grid electricity. Approximately 11% of the South African population (as at 2012) had no access to electricity (Fu, 2003). This with the increased electricity costs only intensifies the gap in energy poverty.

The government is currently introducing numerous efforts to alleviate the limitations faced by citizens (Fu, 2003). These are but a few of the socio-economic consequences arising from the disruption of energy supply and under performance of the national grid. While resolves to electrify the country may be on the forefront, the state utility continues to face increased demand resulting in higher electricity costs for the country as a whole. The improvement of these energy challenges is essential for the economic growth of a developing country such as South Africa.

Solar radiation, amongst various other renewable energy sources, is readily available with long-term certainty. Quantified solar radiation extends its significance to the design and performance analysis of solar devices, the application of photovoltaic (PV) systems, which directly influences the global energy budget (Zhou et al., 2019). Due to the high associated costs, skill and effort required to implement radiometric measurements, many meteorological stations across the globe find it extremely challenging to accurately measure solar radiation data, while maintenance of these stations is also quite difficult. Accurate solar radiation measurements are critical for the assessment of energy resources available within a location and can extend its application to various energy management and development initiatives. To encourage the adoption of renewable energy resources, reliable sources of this quantity (or estimations thereof) are consistently being sought after. This has led to researchers developing numerous physical models for the estimation of solar radiation, many of which can be classified as; empirical models (Fan et al., 2019); machine learning techniques (Fan et al., 2018); Satellite-image models (Pinker et al., 1995) and radiation transfer models (Gueymard, 2004). These physical models which employ readily available meteorological parameters are sufficient for estimating solar radiation in regions where solar radiation data is not measured or recorded as they have lower computational costs and input data requirements (Govindasamy and Chetty, 2019). Solar radiation data is not extensively measured in most developing countries and conventional physical models require this data for validation and calibration of the models. Other meteorological factors such as air temperature, relative humidity, air pollution concentration and sunshine-duration measurements are easily accessible from weather stations.

Machine learning techniques are commonly applied by solar radiation researchers, due to their high estimation accuracy, lower computational costs, and input data requirements. Literature provides various empirical models which have been developed from available meteorological and climatological variables and are applied to estimate solar radiation globally (Besharat et al., 2013). Particularly in South Africa, models have been developed and analyzed by (Kibirige, 2018). Adeala et al. (2015) who concluded that the addition of certain input parameters improves model performance, which suggests that hybrid models are more effective for estimating global solar radiation at South African stations. Maluta and Mulaudzi (2018) found a similar result for estimating global solar radiation in the Limpopo Province. Findings by (Govindasamy and Chetty, 2019) confirmed the enhanced performance of non-linear prediction models for South African provinces.

All preceding studies on the estimation of global solar radiation in South Africa have been limited to including meteorological variables such as air temperature, relative humidity, wind speed and sunshine duration. The inclusion of aerosol concentration to such physical models has never been investigated for this country, hence the work presented in this study is novel and significant.

Several gaseous particles, aerosols and clouds constitute the atmosphere, and each of them absorb, reflect, or scatter to deplete the incident Extraterrestrial radiation (ETR, denoted by H_o) which enters the atmosphere (Viorel, 2008). Elements such as pollutants, aerosol concentrations, atmospheric gas concentrations and clouds reduce the ETR into diffuse solar radiation by absorbing the incident solar radiation or reflecting it back into space (Mulaudzi et al., 2013). Atmospheric pollutants are thus a substantial consideration when evaluating global solar radiation.

Air pollution is a huge problem in most urban areas and can often be measured at concentrations higher than regulatory levels in industrial areas (Raimondo et al., 2007). The most common air pollutants include; sulfur dioxide, nitric oxide, nitrogen dioxide, carbon monoxide, trioxygen, particulate matter (PM_{10}) which is emitted from vehicular traffic (petrol and diesel vehicles), fixed combustion processed with solid or liquid fuel, industrial emissions and natural environmental conditions. Airborne particulate matter (PM) particles vary largely in their frequency, spatial and temporal properties, as well as their chemical configuration (Nejadkoorki and Baroutian, 2012). PM_{10} particles refer to particulates which are <10 μm in size. The study of these particulates has gained substantial interest over the years resulting from their direct relation to health conditions such as increased respiratory symptoms, decreased lung function and concerningly an increase in mortality rate in certain cities (Maraziotis and Marazioti, 2008). From a health governance perspective, monitoring air quality and atmospheric aerosol levels is fundamental for understanding the biological settings of a population (Nadzri et al., 2010).

Numerous studies have been conducted by Pérez et al. (2010) to better understand and forecast the concentrations of PM_{10} levels in the atmosphere using machine learning techniques to improve health and air quality. Modified ANN and SVR models showed superior performance for the prediction of PM_{10} concentrations (Hou et al., 2014). Understanding the dynamics of these atmospheric constituents in combination with meteorological parameters provides a holistic description of the available solar radiation for a specific region.

Machine learning approaches for estimating solar radiation in South Africa have not been explored, let alone models which include particulate matter impurity data (PM_{10}). Aerosol pollutant concentration coupled with the use of generalized datasets which do not compromise the

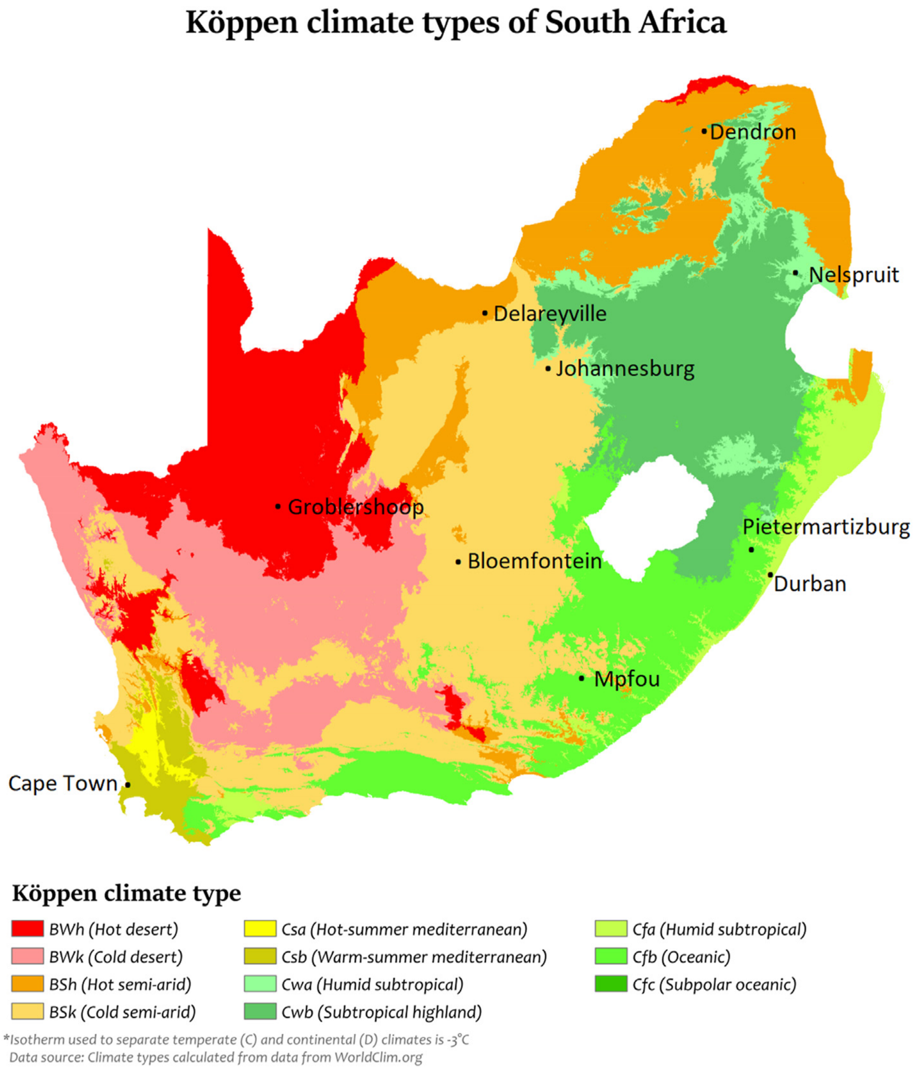


Fig. 1. Köppen climate classification of study sites (Wikipedia, 2021).

comprehensiveness of each geographical site, have not been considered in any of the earlier models developed for South Africa, making this study unique and progressive. This study essentially investigates the application of machine learning models for global solar radiation prediction using various input parameter configurations (models) as well as the effects of introducing PM_{10} data to generalized meteorological datasets. The insights provided herein are significant in describing the impact of the above considerations for South Africa.

2. Materials and methods

2.1. Meteorological datasets

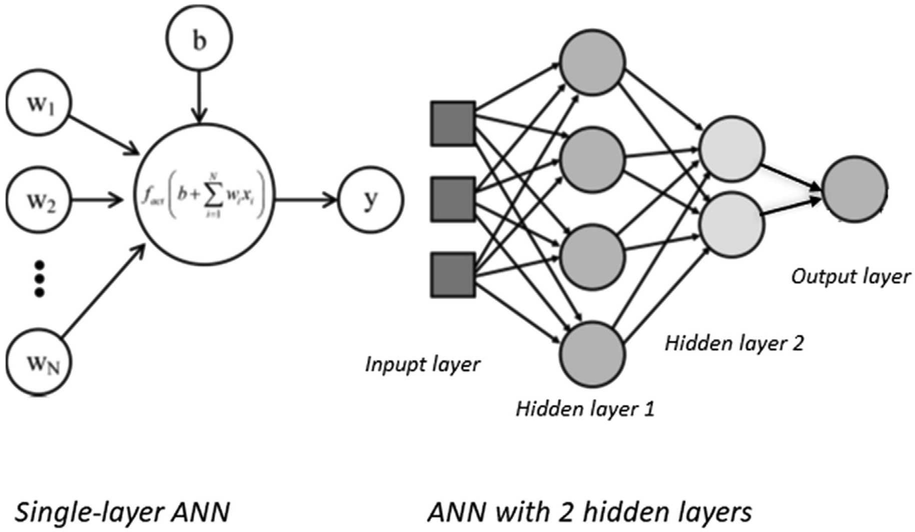
Historical meteorological datasets which include average monthly values of solar radiation, minimum and maximum air temperatures and relative humidity were obtained from the South African Weather Service (SAWS), 2020 and Agricultural Research Council (ARC) for the period January 2007–September 2020. Sunshine duration data was obtained from the International Water Management Institute (IWMI), while PM_{10} data was made available by the South African Air Quality Information System (SAAQIS), 2020. The topographic characteristics of the sites of interest for this study are presented in Table 1, while the climate classification according to Köppen are illustrated in Fig. 1 (Wikipedia, 2021).

2.2. Theory of machine learning techniques

In the recent years, machine learning (ML) techniques have been progressively developed and calibrated for the estimation of solar radiation. These algorithms are robust, high-performing, and efficient. The performance of machine learning models in comparison to empirical models for the prediction of solar radiation has proven superior (Abrahamsen et al., 2018). Most of these techniques include multiple independent variables as input parameters while they identify the contributions of the most significant parameters in predictions. Machine learning techniques are thus considerably suitable for the estimation of global solar radiation due to their high level of accuracy and minimal computational effort. In the following we briefly introduce the techniques considered and provide implementation procedures for each of the models in section 2.3.

2.2.1. Support Vector Regression (SVR)

Support vector machines (learning systems applied to a higher dimensional space) which support linear and non-linear regression and are referred to as Support Vector Regression (SVR). A key feature of SVR is that its computational effort is independent of the input dataset's dimensionality (Debasish et al., 2007). Literature on the application of SVR to predict global solar radiation recommends that this model yields precise results which are often superior to the performance of empirical



Single-layer ANN

ANN with 2 hidden layers

Fig. 2. Structure of a Single-Layer Perceptron and ANN with 2 hidden layers.

models (Feng et al., 2020). Studies also indicate that the SVR algorithm retains accuracy when transferred to sites with similar climatic conditions, while the best performing SVR models are those which input sunshine duration (Wei-Yin, 2011).

2.2.2. Random Forest (RF)

Random forests constitute a team of decision trees and are classified as an ensemble learning technique. Ensemble learning models make use of multiple algorithms or the same algorithm multiple times over to produce a more powerful model. RFs consist of random vectors which have the same distribution. RF is robust to fluctuations in the training data, even for a very large number of input parameters (Rokach and Maimon, 2005). RF is well suited for predicting global solar radiation as the model implicitly applies feature selection and interaction, while also offering the ability to consider the order of the training set data.

2.2.3. Generalized regression Neural Network (GRNN)

GRNN is a Radial Basis Function RBF (Gaussian) model which is based on probability. This technique has been widely used for the purpose of regression modeling where the parameters to be predicted are continuous (Yeboah et al., 2015). GRNN is a derivative of an ANN which operates on the concept of non-linear regression as opposed to the procedures and structure of a standard Feed forward Neural Network (FFNN) which generally consists of three layers (Feng et al., 2020). GRNN easily identifies the regression function which exists between the input and target variables in the training dataset. This makes the model simple to implement, with minimal computation time (Feng et al., 2020).

2.2.4. Artificial Neural Network (ANN)

In current solar radiation prediction trends, ANNs have been substantially investigated and have verified their predictive power consistently. Scientific findings reported by Landeras et al. (2012) demonstrate the superiority of ANNs in predicting solar radiation when compared to other machine learning techniques. The ease of use, automatic learning technique and great level of accuracy and proficiency of this model make it desirable for use by researchers across the globe.

Application of ANNs to estimate solar radiation has demonstrated performance sensitivity to the input data set, with minimal error metrics (Abrahamsen et al., 2018). A comparative study conducted by Tymvios et al. (2005) between the Angstrom-Prescott empirical model and ANNs indicate that ANNs have a higher performance. To describe the structure and learning technique of an ANN, we consider a Single-Layer FFNN (Perceptron) with one hidden layer as shown in Fig. 2. Training set data which contains meteorological variables such as air temperature, sunshine duration and relative humidity serve as the nodes in the input layer. These variables feed into the hidden (first) layer which contains a non-linear activation function of the form;

$$\Phi_i = \sum_{i=1}^{m} X_i W_i$$

Where X_i represents the i^{th} independent variable and W_i represents the corresponding weight of the variable.

Neurons in the hidden layer combine the criteria of all the input variables to create a new attribute which is stored in this layer. This hidden layer also allows for an increase in flexibility of the model by requiring specific conditions to be met. The output layer corresponds to the input variables for that specific observation and can be either continuous, binary or a categorical output depending on the ANN being implemented. Activation functions are dependent on the problem being solved and can be either of the following; Threshold function, Sigmoid function, Rectifier function or Hyperbolic tangent (tanh) function (Abrahamsen et al., 2018). At the output layer, after the initial predictions are made, a loss function is computed. This function computes the error in the prediction and feeds this information back to the hidden layer to adjust the weights associated with the input variables accordingly. This process is called back propagation and it enables the ANN to minimize the error between the observed and predicted variables by simultaneously adjusting the weights of all input variables. Other learning methods for ANNs include gradient descent and stochastic gradient descent. Depending on the training conditions specified by the user, the above process is repeated for a set number of epochs (training iterations), until the error is minimized. For our purposes we have considered a Multi-Layer Perceptron (MLP) which consists of two hidden layers with the non-linear rectifier activation function and which employs the standard version of back propagation algorithm.

2.3. Methodology and procedure

Meteorological parameters across the nine provinces for the period Jan 2007–Sep 2020 were combined to form the initial generalized dataset. The second approach of including PM_{10} data employed a dataset for the period Jan 2018–Sep 2020 for which impurity data was consistently available. Due to station and equipment limitations, pollutant concentrations and meteorological conditions are measured periodically for almost all cities in South Africa, this resulted in abundant missing datapoints and suggested the use of generalized monthly datasets. For implementation of the machine learning analysis python was employed

Table 2
Input parameters for models developed.

Type	Model	Input parameters
Sunshine	M1	$H_o, (S/S_o), (S/S_o)^{0.5}, (S/S_o)^2, (S/S_o)^3, (S/S_o)^4$
	M2	$H_o, (S/S_o), (S/S_o)^{0.5}, (S/S_o)^2, (S/S_o)^3$
	M3	$H_o, (S/S_o), (S/S_o)^{0.5}, (S/S_o)^2$
	M4	$H_o, (S/S_o), (S/S_o)^{0.5}$
	M5	$H_o, (S/S_o)$
	M6	$H_o, (S/S_o)^{0.5}$
Temperature	M1	$H_o, T_{ave}, T_{max}, T_{max}^2, \Delta T, (\Delta T)^{0.5}, (\Delta T)^{1.5}, (\Delta T)^2, (\Delta T)^{2.5}$
	M2	$H_o, T_{ave}, T_{max}, T_{max}^2, \Delta T, (\Delta T)^{0.5}, (\Delta T)^{1.5}, (\Delta T)^2$
	M3	$H_o, T_{ave}, T_{max}, T_{max}^2, \Delta T, (\Delta T)^{0.5}, (\Delta T)^{1.5}$
	M4	$H_o, T_{ave}, T_{max}, T_{max}^2, \Delta T, (\Delta T)^{0.5}$
	M5	$H_o, T_{ave}, T_{max}, \Delta T, (\Delta T)^{0.5}$
	M6	$H_o, T_{ave}, \Delta T, (\Delta T)^{0.5}$
	M7	$H_o, \Delta T, (\Delta T)^{0.5}$
	M8	$H_o, (\Delta T)^{0.5}$
Relative humidity	M1	$H_o, RH_{ave}, \Delta RH, (\Delta RH)^{0.5}, (\Delta RH)^{0.6667}, (\Delta RH)^{1.5}, (\Delta RH)^2$
	M2	$H_o, RH_{ave}, \Delta RH, (\Delta RH)^{0.5}, (\Delta RH)^{0.6667}, (\Delta RH)^{1.5}$
	M3	$H_o, RH_{ave}, \Delta RH, (\Delta RH)^{0.5}, (\Delta RH)^{0.6667}$
	M4	$H_o, \Delta RH, (\Delta RH)^{0.5}, (\Delta RH)^{0.6667}$
	M5	$H_o, (\Delta RH)^{0.5}, (\Delta RH)^{0.6667}$
	M6	$H_o, (\Delta RH)^{0.6667}$
Hybrid - Linear	M1	$H_o, (S/S_o), T_{ave}, T_{max}, \Delta T, RH_{ave}, \Delta RH$
	M2	$H_o, (S/S_o), T_{max}, \Delta T, RH_{ave}, \Delta RH$
	M3	$H_o, (S/S_o), \Delta T, \Delta RH$
	M4	$H_o, (S/S_o), T_{max}, RH_{ave}$
Hybrid – Non-linear	M1	$H_o, (S/S_o), (S/S_o)^{0.5}, (S/S_o)^2, (S/S_o)^3, (S/S_o)^4, T_{ave}, T_{max}, T_{max}^2, \Delta T, (\Delta T)^{0.5}, (\Delta T)^{1.5}, (\Delta T)^2, (\Delta T)^{2.5}, RH_{ave}, \Delta RH, (\Delta RH)^{0.5}, (\Delta RH)^{0.6667}, (\Delta RH)^{1.5}, (\Delta RH)^2$
	M2	$H_o, (S/S_o), (S/S_o)^{0.5}, T_{ave}, T_{max}, T_{max}^2, \Delta T, (\Delta T)^{0.5}, (\Delta T)^{1.5}, (\Delta T)^2, (\Delta T)^{2.5}, RH_{ave}, \Delta RH, (\Delta RH)^{0.5}, (\Delta RH)^{0.6667}, (\Delta RH)^{1.5}, (\Delta RH)^2$
	M3	$H_o, (S/S_o), (S/S_o)^{0.5}, \Delta T, (\Delta T)^{0.5}, RH_{ave}, \Delta RH, (\Delta RH)^{0.5}, (\Delta RH)^{0.6667}, (\Delta RH)^{1.5}, (\Delta RH)^2$
	M4	$H_o, (S/S_o), (S/S_o)^{0.5}, \Delta T, (\Delta T)^{0.5}, \Delta RH, (\Delta RH)^{0.5}, (\Delta RH)^{0.6667}$
	M5	$H_o, (S/S_o), (S/S_o)^{0.5}, (S/S_o)^2, \Delta T, (\Delta T)^{0.5}, \Delta RH, (\Delta RH)^{0.6667}$
	M6	$H_o, (S/S_o), (S/S_o)^{0.5}, (S/S_o)^2, (S/S_o)^3, T_{max}, \Delta T, (\Delta T)^{0.5}, \Delta RH, (\Delta RH)^{0.6667}$
	M7	$H_o, (S/S_o)^{0.5}, \Delta T, (\Delta T)^{0.5}, \Delta RH, (\Delta RH)^{0.6667}$
	M8	$H_o, (S/S_o)^{0.5}, (\Delta T)^{0.5}, (\Delta RH)^{0.6667}$
	M9	$H_o, (S/S_o)^{0.5}, (\Delta T)^{0.5}$
	M10	$H_o, (S/S_o)^{0.5}, (\Delta RH)^{0.6667}$
	M11	$H_o, (S/S_o)^{0.5}, \Delta T, (\Delta T)^{0.5}$
	M12	$H_o, (S/S_o)^{0.5}, (\Delta RH)^{0.5}, (\Delta RH)^{0.6667}$
Jan 2018 – Sep 2020		**PM_{10} (μg/m3) Models**
Sunshine	M1	$H_o, (S/S_o), (S/S_o)^{0.5}, (S/S_o)^2, PM_{10}$
	M2	$H_o, (S/S_o), (S/S_o)^{0.5}, PM_{10}$
	M3	$H_o, (S/S_o)^{0.5}, PM_{10}$
Temperature	M1	$H_o, \Delta T, (\Delta T)^{0.5}, PM_{10}$
	M2	$H_o, (\Delta T)^{0.5}, PM_{10}$
Relative Humidity	M1	$H_o, RH_{ave}, \Delta RH, (\Delta RH)^{0.5}, (\Delta RH)^{0.6667}, (\Delta RH)^{1.5}, (\Delta RH)^2, PM_{10}, (PM_{10})^{0.5}, (PM_{10})^2$
	M2	$H_o, \Delta RH, (\Delta RH)^{0.5}, (\Delta RH)^{0.6667}, (\Delta RH)^{1.5}, (\Delta RH)^2, PM_{10}, (PM_{10})^{0.5}$
	M3	$H_o, \Delta RH, (\Delta RH)^{0.5}, (\Delta RH)^{0.6667}, PM_{10}$
	M4	$H_o, (\Delta RH)^{0.5}, (\Delta RH)^{0.6667}, PM_{10}$
Linear Hybrid	M1	$H_o, (S/S_o), T_{ave}, \Delta T, RH_{ave}, \Delta RH, PM_{10}$
	M2	$H_o, (S/S_o), \Delta T, \Delta RH, PM_{10}$
Non-linear Hybrid	M1	$H_o, (S/S_o), (S/S_o)^{0.5}, (S/S_o)^2, (S/S_o)^3, (S/S_o)^4, T_{ave}, \Delta T, (\Delta T)^{0.5}, (\Delta T)^{1.5}, (\Delta T)^2, (\Delta T)^{2.5}, RH_{ave}, \Delta RH, (\Delta RH)^{0.5}, (\Delta RH)^{0.6667}, (\Delta RH)^{1.5}, (\Delta RH)^2, PM_{10}, (PM_{10})^{0.5}, (PM_{10})^2$
	M2	$H_o, (S/S_o), (S/S_o)^{0.5}, (S/S_o)^2, T_{ave}, \Delta T, (\Delta T)^{0.5}, (\Delta T)^2, RH_{ave}, \Delta RH, (\Delta RH)^{0.5}, (\Delta RH)^{0.6667}, PM_{10}, (PM_{10})^{0.5}, (PM_{10})^2$
	M3	$H_o, (S/S_o), (S/S_o)^{0.5}, \Delta T, (\Delta T)^{0.5}, \Delta RH, (\Delta RH)^{0.5}, (\Delta RH)^{0.6667}, PM_{10}$
	M4	$H_o, (S/S_o)^{0.5}, \Delta T, (\Delta RH)^{0.5}, (\Delta RH)^{0.6667}, PM_{10}$
	M5	$H_o, (S/S_o)^{0.5}, (\Delta T)^{0.5}, (\Delta RH)^{0.5}, (\Delta RH)^{0.6667}, PM_{10}$

Where H_o (MJ/m^2) represents ETR, (S/S_o) sunshine duration, ΔT (°C) = T_{max} - T_{min} ($T_{max,}$ T_{min} are the minimum and maximum monthly temperatures respectively), T_{ave} (°C) is the average monthly temperature, RH_{ave} (%) is the average monthly relative humidity, ΔRH (%) = RH_{max} - RH_{min} ($RH_{max,}$ RH_{min} are the minimum and maximum monthly relative humidity recordings respectively), PM_{10} (μg/m3) is the particulate matter concentration.

Table 3
Details of the statistical tests.

Abbreviation	Statistical test	Expression
MBE	Mean Bias Error	$MBE = \frac{1}{n}\sum_{i=1}^{n}((O_i - P_i))$
MPE	Mean percentage error	$MPE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(O_i - P_i)}{O_i}\right) \times 100$
RMSE	Root mean square error	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2}$
RRMSE	Relative root mean square error	$RRMSE = \frac{\left[\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2\right]^{1/2}}{\frac{1}{n}\sum_{i=1}^{n}O_{ave}}$
R^2	Coefficient of determination	$R^2 = 1 - \left[\frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - O_{ave})^2}\right]$

Where O_i *and* P_i represent the i^{th} observed and predicted values, respectively.

through the Spyder IDE (Integrated Development Environment) - v4.0.1.

As part of the data preprocessing steps, we made use of the Pandas libraries in Python to read the input datasets and the Sklearn package to process the data in the following steps. Missing values within the independent variable datasets were imputed using the mean of the corresponding variable. This enables a richer dataset without the loss of data points with valuable information. Categorical data such as the months of the year and city names were encoded using the label encoder modules in Python. This accounts for any seasonality which may be present in the dataset. Since most ML techniques (especially those dependent on Euclidean distance, e.g. GRNN) are sensitive to the size and noise of input parameters, we have applied feature scaling to all datasets.

The initial generalized datasets were split into training and test data subsets (test size = 0.3) for model evaluation. Neural network models (GRNN and ANN) were trained on a representative subset of the initial dataset used for the model selection.

ML techniques discussed above were executed for each of the four models according to input parameter configurations described in Table 2 below, and error metrics were evaluated for their performance. A summary of the procedures applied for each model is mentioned below.

1. SVR: For this technique we employed the SVM module and imported the SVR class to create a regressor with the following hyper parameter arrangement; SVR (kernel = 'rbf', gamma = 'scale', epsilon = '0.1'.
2. RF: The regressor was created using the RandomForestRegressor with n = 10 estimators from the Ensemble module in Sklearn.
3. GRNN: This model required the Neupy package in python along with the Algorithms and Core modules to implement GRNN with standard deviation (std = 0.1).
4. ANN: ANN in python requires the use of the Keras and TensorFlow packages. The modules used were Sequential, Dense and Activation. We created an ANN with two hidden layers which works with a non-linear rectifier activation function. The optimizer (optimization algorithm) used for this model was 'adam' and the loss function computed was 'mean_absolute_error'.

Table 4
Model performance and error statistics.

SVR	Model	MBE	MPE (%)	MAE	MSE	RMSE	R^2
Sunshine	M1	0.11571	−3.88325	0.28919	0.24527	0.49252	0.99290
	M2	0.11653	−3.87358	0.28070	0.23627	0.48073	0.99316
	M3	0.11695	−3.82970	0.26930	0.22552	0.47489	0.99347
	M4	0.10650	−3.80642	0.27634	0.23435	0.47410	0.99322
	M5	0.00136	−8.77104	2.19148	7.54151	2.74618	0.78173
	M6	0.01975	−5.27628	1.22981	2.41643	1.55449	0.93006
Temperature	M1	−0.06453	−2.32353	0.45970	0.91743	0.95782	0.97345
	M2	−0.04639	−2.32796	0.42381	0.78260	0.88465	0.97735
	M3	−0.02880	−2.33754	0.38875	0.61034	0.78125	0.98233
	M4	−0.00275	−2.44332	0.36023	0.46910	0.68491	0.98642
	M5	0.00633	−2.38893	0.43107	0.40969	0.64007	0.98814
	M6	−0.03143	−2.18167	0.51648	0.66200	0.81363	0.98084
	M7	−0.13267	−2.32290	0.51648	0.66200	0.81363	0.98084
	M8	0.23852	−8.12466	2.24092	8.37825	2.89452	0.75751
Relative humidity	M1	0.05242	−8.21981	1.79547	6.51526	2.55250	0.81143
	M2	0.06201	−8.14698	1.78209	6.44323	2.53835	0.81352
	M3	0.09072	−8.12872	1.75332	6.27178	2.05435	0.81848
	M4	0.16344	−8.34186	1.78466	6.54178	2.55769	0.81066
	M5	0.22304	−8.39301	1.74499	6.35908	2.52172	0.81595
	M6	−0.03586	−9.49537	2.24794	9.53754	3.08829	0.72396
Hybrid - Linear	M1	0.09856	−5.14356	1.30640	2.91300	1.70918	0.91545
	M2	0.06249	−6.30083	1.54362	4.12116	2.03006	0.88072
	M3	0.10880	−8.86583	2.03694	6.42890	2.53553	0.81393
	M4	0.07759	−6.27554	1.55960	4.13723	2.03402	0.88026
Hybrid – Non-linear	M1	0.07367	−3.54630	0.31274	0.26472	0.51451	0.99234
	M2	0.06030	−3.53722	0.39473	0.35857	0.59881	0.98962
	M3	0.06766	−4.04387	0.39529	0.37887	0.61552	0.98903
	M4	0.07907	−3.81184	0.32238	0.27747	0.52676	0.99197
	M5	0.08871	−3.72396	0.29098	0.23855	0.48842	0.99310
	M6	0.10192	−3.75491	0.29125	0.23368	0.48340	0.99324
	M7	0.05828	−3.96297	0.60168	0.65400	0.80870	0.98107
	M8	0.15741	−5.70887	1.10749	1.92504	1.38746	0.99428
	M9	0.15039	−5.61585	1.12482	1.96033	1.40012	0.94326
	M10	0.05496	−5.35211	1.13007	2.15343	1.46746	0.93767
	M11	0.05382	−3.63043	0.46410	0.44166	0.66458	0.98722
	M12	0.03861	−5.35217	1.14040	2.18093	1.47680	0.93688

RF	**Model**	**MBE**	**MPE (%)**	**MAE**	**MSE**	**RMSE**	**R2**
Sunshine	M1	−0.09485	−3.17567	0.72982	0.99289	0.99644	0.97126
	M2	−0.08899	−3.35422	0.73085	0.98018	0.99004	0.97163
	M3	−0.07059	−3.17715	0.74064	1.01086	1.00541	0.97074
	M4	−0.05531	−3.26521	0.81502	1.15914	1.07663	0.96645
	M5	−0.07143	−5.16969	2.02593	6.90980	2.62865	0.80001
	M6	0.00002	−3.63872	1.20516	2.42170	1.55618	0.92991
Temperature	M1	0.02008	−4.51193	1.02121	1.83216	1.35357	0.94697
	M2	0.02825	−4.44696	1.03526	1.83867	1.35597	0.94678
	M3	0.04331	−4.55109	1.05105	1.95654	1.39876	0.94337
	M4	0.05115	−4.24183	1.06377	2.02648	1.42354	0.94135
	M5	0.08451	−4.47268	1.10704	2.16329	1.47081	0.93739
	M6	0.04043	−4.98418	1.10012	2.21700	1.48896	0.93583
	M7	0.01845	−5.21401	0.97408	2.01423	1.41924	0.94170
	M8	0.04002	−6.13688	1.95769	7.20143	2.68354	0.79157
Relative humidity	M1	−0.14091	−4.48399	1.03870	3.65921	1.91291	0.89409
	M2	−0.14922	−4.45264	1.11050	3.66011	1.91314	0.89407
	M3	−0.14090	−4.57532	1.12186	3.56890	1.88915	0.89671
	M4	−0.03672	−4.86855	1.14889	3.66044	1.91323	0.89406
	M5	−0.01530	−5.07213	1.28758	4.35125	2.08597	0.87406
	M6	−0.04260	−5.50381	1.85151	6.90036	2.62685	0.80029
Hybrid - Linear	M1	0.05994	−4.91259	1.02864	1.83501	1.35463	0.94689
	M2	0.05857	−4.90174	1.02895	1.86135	1.36431	0.94613
	M3	0.10171	−5.42725	1.53226	4.11227	2.02787	0.88098
	M4	0.05031	−4.84559	1.19121	2.41371	1.55361	0.93014
Hybrid – Non-linear	M1	−0.01971	−3.92043	0.80902	1.13461	1.06518	0.96716
	M2	−0.01069	−3.69779	0.82705	1.19383	1.09262	0.96545
	M3	−0.00846	−3.88702	0.81783	1.25428	1.11995	0.96398
	M4	−0.01067	−3.96328	0.88005	1.40006	1.18324	0.95948
	M5	−0.04612	−3.85286	0.87850	1.37745	1.17365	0.96013
	M6	−0.03241	−3.89112	0.82593	1.18329	1.08779	0.96575
	M7	−0.02848	−3.86444	0.99368	1.69256	1.30098	0.95101
	M8	−0.03319	−3.80063	1.02545	1.87113	1.36789	0.94584
	M9	−0.04559	−3.54688	1.10791	2.03483	1.42648	0.94111
	M10	−0.01077	−3.91496	1.05365	1.94469	1.39452	0.94372
	M11	−0.05015	−3.66460	1.07251	1.90005	1.37842	0.94501
	M12	−0.03168	−3.64484	1.02484	1.80921	1.34507	0.94767

(*continued on next page*)

Table 4 (*continued*)

SVR	Model	MBE	MPE (%)	MAE	MSE	RMSE	R^2
ANN	**Model**	**MBE**	**MPE (%)**	**MAE**	**MSE**	**RMSE**	**R2**
Sunshine	M1	−0.01029	−0.68766	0.15897	0.06115	0.24729	0.99823
	M2	0.02627	−0.17638	0.25314	0.11883	0.34472	0.99656
	M3	−0.09183	−0.45533	0.23044	0.13198	0.36329	0.99618
	M4	−0.02812	−0.14939	0.19520	0.07141	0.26722	0.99793
	M5	−0.17026	−4.36993	2.12945	7.27326	2.69690	0.78949
	M6	0.19708	−2.34664	1.14697	2.09102	1.44604	0.93948
Temperature	M1	−0.02168	0.40338	0.39238	0.35128	0.59269	0.98983
	M2	−0.06925	0.09462	0.26080	0.37434	0.61183	0.98983
	M3	0.01442	−0.04323	0.29951	0.26911	0.51875	0.99221
	M4	0.01981	0.08782	0.25972	0.36822	0.60681	0.98934
	M5	−0.16419	0.88478	0.38819	0.32139	0.56691	0.99070
	M6	−0.01480	−2.43535	0.32545	0.29360	0.54163	0.99151
	M7	0.05298	−1.33830	0.25230	0.17528	0.41866	0.99493
	M8	0.15227	−5.43632	2.05140	7.05380	2.65590	0.79585
Relative humidity	M1	0.23362	−3.96057	0.68454	2.34072	1.52994	0.93225
	M2	0.15025	−4.42033	0.69001	2.38106	1.54307	0.93109
	M3	0.34796	−3.19611	0.83237	2.75455	1.65968	0.92028
	M4	0.21581	−5.45037	0.78142	2.54317	1.59473	0.92639
	M5	0.56617	−6.24219	1.47195	5.01814	2.24012	0.85476
	M6	0.29088	−5.43279	1.85083	7.00042	2.64583	0.79739
Hybrid - Linear	M1	0.38964	−2.91133	1.11020	2.11722	1.45507	0.93872
	M2	0.14334	−3.67534	1.28395	2.85818	1.69061	0.91728
	M3	0.15817	−4.84855	1.90700	3.72612	2.39293	0.83427
	M4	0.34829	−4.45297	1.34307	3.04457	1.74487	0.91188
Hybrid – Non-linear	M1	−0.02867	0.85567	0.18891	0.08615	0.29352	0.99751
	M2	0.08299	−0.37390	0.23456	0.15025	0.38763	0.99565
	M3	0.01901	−0.58193	0.18106	0.07029	0.26512	0.99797
	M4	−0.11159	0.03559	0.19015	0.07925	0.28151	0.99771
	M5	−0.01706	−0.30242	0.16430	0.05120	0.22627	0.99852
	M6	−0.04798	0.31356	0.15818	0.05227	0.22863	0.99849
	M7	0.07795	0.16584	0.30611	0.19550	0.44215	0.99434
	M8	−0.06457	−0.78440	0.99679	1.55394	1.24657	0.95503
	M9	0.04378	−2.05695	1.02561	1.63009	1.27675	0.95282
	M10	0.27871	−3.95900	1.16023	2.26829	1.50608	0.93435
	M11	0.08624	0.09600	0.47600	0.40570	0.63694	0.98826
	M12	−0.07173	−0.73821	0.95318	1.59841	1.26428	0.95374
GRNN	**Model**	**MBE**	**MPE (%)**	**MAE**	**MSE**	**RMSE**	**R2**
Sunshine	M1	−0.05787	−2.93532	0.75780	1.18620	1.08914	0.96567
	M2	−0.06061	−2.88053	0.82268	1.44557	1.20232	0.95817
	M3	−0.00849	−3.27359	0.74356	1.17487	1.08391	0.96600
	M4	−0.00260	−3.26825	0.80210	1.31009	1.14459	0.96208
	M5	−0.02414	−4.65324	2.02889	6.92977	2.63244	0.79943
	M6	0.00336	−3.29149	1.22512	2.49549	1.57971	0.92778
Temperature	M1	0.09939	−4.06990	1.33930	3.36911	1.83551	0.90249
	M2	0.15541	−4.39300	1.32510	3.24620	1.80172	0.90604
	M3	0.17535	−4.58369	1.33639	3.24076	1.80021	0.90620
	M4	0.13417	−4.32191	1.31004	3.05983	1.74924	0.91144
	M5	0.24230	−4.70520	1.20835	2.69008	1.64015	0.92214
	M6	0.22024	−4.44829	1.02616	2.18453	1.47802	0.93677
	M7	0.02815	−3.58818	0.90734	1.77969	1.33405	0.94849
	M8	0.13712	−5.99598	2.02144	7.36621	2.71408	0.78680
Relative humidity	M1	−0.16041	−3.24600	1.13519	3.59826	1.89690	0.89586
	M2	−0.16135	−3.24656	1.13719	3.60746	1.89933	0.89559
	M3	−0.16209	−3.26031	1.14719	3.63488	1.90654	0.89479
	M4	−0.06211	−3.79153	1.12445	3.31578	1.82093	0.90403
	M5	−0.02068	−4.25500	1.28871	3.84240	1.95045	0.88990
	M6	−0.09261	−4.52372	1.75614	6.02561	2.49111	0.82039
Hybrid - Linear	M1	−0.04109	−3.73008	1.44176	3.79983	1.94931	0.89002
	M2	−0.01638	−4.00922	1.51215	3.99768	1.99942	0.88430
	M3	−0.08388	−3.69487	1.80429	5.80636	2.40964	0.83190
	M4	0.05291	−4.23651	1.56293	4.24119	2.05942	0.87725
Hybrid – Non-linear	M1	0.05456	−4.69487	1.62799	5.17232	2.27427	0.85030
	M2	−0.03966	−3.85953	1.45348	4.24747	2.06094	0.87707
	M3	−0.00463	−3.99808	1.16162	2.89819	1.70241	0.91612
	M4	−0.01540	−3.99236	1.15815	2.85099	1.68849	0.91749
	M5	−0.03658	−3.22214	1.13645	2.30786	1.51916	0.93320
	M6	0.05671	−4.15596	1.30607	3.13597	1.77087	0.90924
	M7	−0.02087	−3.18525	1.11576	2.28550	1.51179	0.93385
	M8	0.03941	−3.53861	1.12539	2.30094	1.51688	0.93340
	M9	−0.00025	−3.39571	1.25533	2.75600	1.66012	0.92023
	M10	0.04521	−3.48791	1.15930	2.38042	1.54286	0.93100
	M11	0.01176	−3.48622	1.14535	2.43470	1.56035	0.92953
	M12	0.04532	−3.48696	1.13226	2.25256	1.50085	0.93481

Table 5
Models including PM_{10} ($\mu g/m^3$) for the period Jan 2018–Sep 2020

SVR	Model	MBE	MPE (%)	MAE	MSE	RMSE	R^2
Sunshine + PM_{10}	M1	−5.14243	0.47943	1.13598	2.30073	1.51820	0.90958
	M2	−5.46464	0.47003	1.36983	3.28572	1.81266	0.87087
	M3	−5.43828	0.46042	1.33047	3.14480	1.77336	0.87640
Temperature + PM_{10}	M1	−5.07561	0.02714	2.27651	8.92486	2.98745	0.64924
	M2	−4.85535	0.02759	2.16550	8.14530	2.85400	0.67988
Relative humidity + PM_{10}	M1	−3.95342	0.42865	1.08919	2.47020	1.57440	0.90257
	M2	−2.44861	0.13268	1.03237	1.91262	1.38297	0.92748
	M3	−2.69415	0.20440	1.00413	1.81226	1.34620	0.92878
	M4	−2.65694	0.21735	1.04681	2.09907	1.44882	0.91750
Hybrid – Linear + PM_{10}	M1	−3.86579	0.42532	1.20209	2.74884	1.65796	0.89197
	M2	−3.57885	0.11990	1.48159	3.83727	1.95889	0.84919
Hybrid – Non-linear + PM_{10}	M1	−2.29634	0.23266	0.66889	0.80150	0.89526	0.96850
	M2	−1.85421	0.20174	0.61147	0.71583	0.84607	0.97187
	M3	−1.30956	0.07835	0.76478	0.93103	0.96490	0.96341
	M4	−1.46916	0.08948	0.83626	1.18173	1.08707	0.95356
	M5	−1.45760	0.07369	0.86565	1.19289	1.09220	0.95312
RF	**Model**	**MBE**	**MPE (%)**	**MAE**	**MSE**	**RMSE**	**R^2**
Sunshine + PM_{10}	M1	−1.29860	−0.03359	1.66525	4.80471	2.19196	0.81117
	M2	−0.80506	−0.09420	1.70115	5.16135	2.27186	0.79715
	M3	−1.25641	0.02835	1.58895	4.39383	2.09615	0.82731
Temperature + PM_{10}	M1	−1.19704	−0.05428	1.87084	6.05641	2.46098	0.76197
	M2	0.08302	−0.31011	2.17805	7.68162	2.77157	0.69810
Relative humidity + PM_{10}	M1	−6.17868	0.75992	1.72357	4.09223	2.22491	0.80545
	M2	−3.62899	0.40288	1.45631	3.72191	1.92922	0.85372
	M3	−2.59892	0.30141	1.31994	2.98475	1.72764	0.82269
	M4	−2.35900	0.23799	1.30238	2.92857	1.71132	0.88490
Hybrid – Linear + PM_{10}	M1	−3.54889	0.31103	1.45841	3.55151	1.88454	0.86042
	M2	−2.58852	0.19552	1.66962	4.85753	2.20398	0.80909
Hybrid – Non-linear + PM_{10}	M1	−2.98896	0.21360	1.42735	3.55151	1.88454	0.86042
	M2	−4.44447	0.45624	1.50350	3.74494	1.93519	0.85282
	M3	−1.99872	0.07033	1.45790	3.67762	1.91771	0.85546
	M4	−0.78509	−0.06030	1.33530	2.97148	1.72380	0.88322
	M5	−1.70605	0.03193	1.50536	3.90056	1.97701	0.84639
ANN	**Model**	**MBE**	**MPE (%)**	**MAE**	**MSE**	**RMSE**	**R^2**
Sunshine + PM_{10}	M1	−2.33619	0.29492	0.81847	1.22919	1.10869	0.94114
	M2	−4.46485	0.48770	1.42166	3.26468	1.80684	0.87169
	M3	−3.01102	0.31716	1.12691	2.06207	1.43599	0.91896
Temperature + PM_{10}	M1	−3.94781	−0.05911	2.15486	8.70044	2.94965	0.65806
	M2	−0.15033	−0.45325	1.89813	6.64646	2.57807	0.73878
Relative humidity + PM_{10}	M1	−2.42111	0.34087	0.87188	1.33550	1.15565	0.94751
	M2	0.12375	−0.03257	0.54052	0.47356	0.68815	0.98139
	M3	−2.21804	0.16768	1.06202	1.90866	1.38154	0.92499
	M4	−0.52435	−0.02991	0.93607	1.54566	1.24325	0.93925
Hybrid – Linear + PM_{10}	M1	−2.43404	0.30761	1.07160	1.86642	1.36617	0.92665
	M2	−0.28229	−0.22448	1.44165	3.37822	1.83799	0.86723
Hybrid – Non-linear + PM_{10}	M1	−0.94741	0.14185	0.55971	0.54532	0.73846	0.97857
	M2	−0.36948	0.06199	0.60855	0.50991	0.71409	0.97996
	M3	−0.61563	0.00582	0.82982	1.09209	1.04503	0.95708
	M4	0.74198	−0.12327	0.80971	0.96792	0.98383	0.96196
	M5	−1.54135	0.15352	0.85224	1.16341	1.07862	0.95428
GRNN	**Model**	**MBE**	**MPE (%)**	**MAE**	**MSE**	**RMSE**	**R^2**
Sunshine + PM_{10}	M1	0.71447	−0.32102	1.89671	7.36057	2.71308	0.71072
	M2	−1.10826	−0.09885	1.73216	3.49361	2.34384	0.78409
	M3	−1.80477	0.03258	1.53815	4.39118	2.09551	0.82742
Temperature + PM_{10}	M1	−0.71813	−0.15548	2.28890	10.13310	3.18325	0.60175
	M2	−1.80419	−0.08625	2.42668	12.12336	3.48186	0.52353
Relative humidity + PM_{10}	M1	−1.15795	0.06248	1.93486	8.03859	2.83524	0.68407
	M2	−0.81144	−0.01122	1.66256	5.42927	2.33008	0.78662
	M3	−0.83217	−0.01842	1.55954	5.09283	2.25673	0.78662
	M4	−0.76169	−0.03206	1.68826	5.15755	2.22567	0.77984
Hybrid – Linear + PM_{10}	M1	−0.61155	0.00657	1.48199	4.81245	2.19373	0.81086
	M2	−2.54424	0.13368	2.19159	9.83360	3.13670	0.61332
Hybrid – Non-linear + PM_{10}	M1	−0.33441	−0.14409	2.14111	9.35688	3.05890	0.63260
	M2	−1.41922	−0.09043	1.89030	7.55656	2.74892	0.70301
	M3	−0.24565	−0.29808	1.85090	7.74896	2.78370	0.69545
	M4	−1.59541	0.08772	1.50954	4.70720	2.16975	0.81497
	M5	0.74272	−0.31244	1.58933	4.82235	2.19598	0.81047

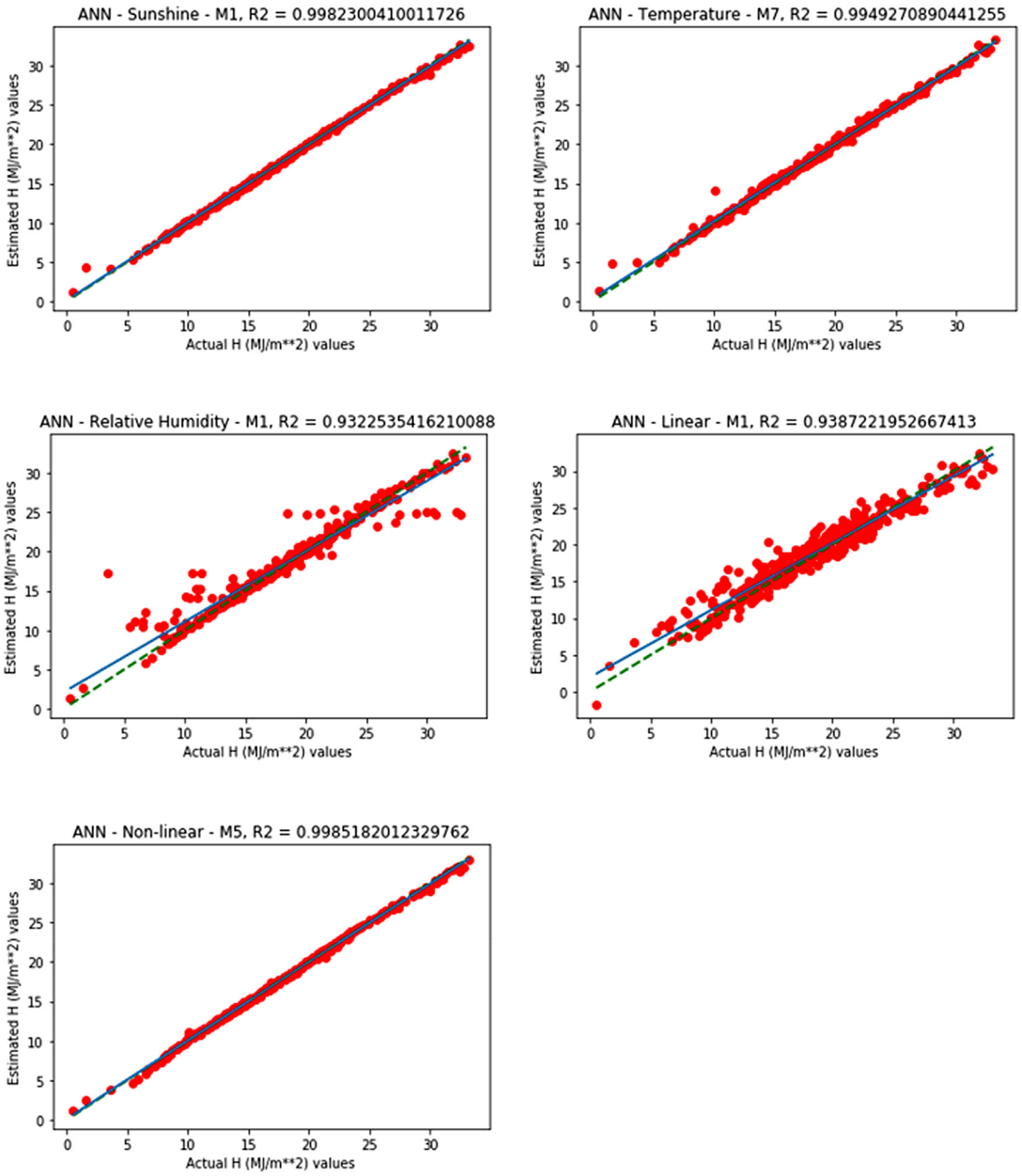


Fig. 3. Best performing ANN models for each parameter configuration type (excluding PM_{10}).

2.4. Statistical error analysis

An understanding of the accuracy of our analysis and proposed models is quantified in terms of the statistical error analysis. The mean bias error (MBE), specifies the average deviance of the calculated values from observed values and is an indicator of a model's long-term performance. The root mean square error (RMSE), gives insight into the short-term performance of a correlation. The coefficient of determination (R^2), is a measure of the correlation between the variance in the dependent (predicted) variables that can be described by the independent (measured) variables. Low values for all statistical error measures are desired, along with high correlation indicators. Previous studies propose that percentage errors between −10 and 10% are acceptable (Marwal et al., 2012). Statistical analysis reported in the present study was calculated using the error metrics presented in Table 3.

3. Results

Model performance and error metrics are described in Table 4 according to the model type (input parameters) and machine learning technique. Table 5 provides a view of model enhancement by including PM_{10} data.4.

3.1. Prediction models excluding PM_{10}

Generally, the ANN models outperformed the RF, SVR and GRNN prediction models, with Hybrid-Non-linear – M5 model indicating the best performance ($R^2 = 0.99852$, MBE = -0.01706). All error metrics for this model were well within the accepted range and showed strong correlation, R^2 in the range [0.93435: 0.99852] for the 12 non-linear hybrid models assessed. Non-linear model configurations were the best performing and this agrees with work completed in (Govindasamy and Chetty, 2019) which suggests that hybrid, non-linear models show improved performance when compared to linear, single parameter prediction models. The ANN models are flexible in a sense that the hidden layers can be configured and modified based on the input dataset. The subsequent best performing machine learning technique for hybrid models was the SVR- M8, $R^2 = 0.99428$, MPE = −5.70887%.

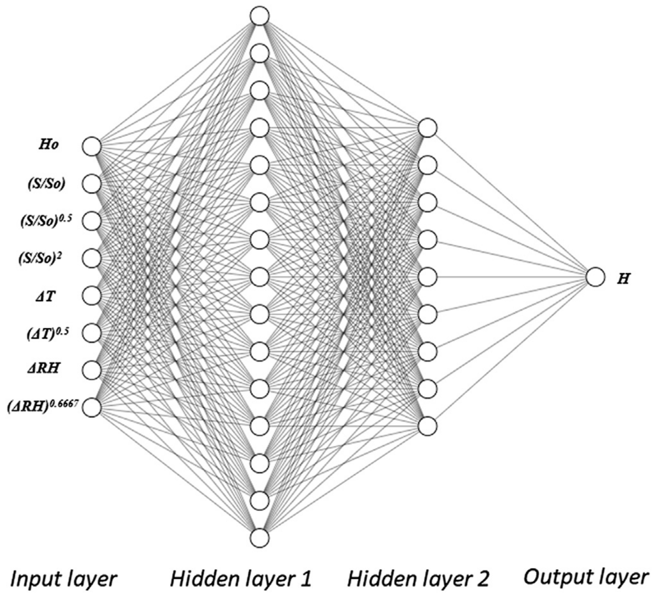


Fig. 4. ANN architecture for Hybrid – Non-linear -M5 (Publication-ready NN-architecture schematics, 2020). Where H (MJ/m^2) represents the global solar radiation.

Literature suggests that SVR models retain accuracy when transferred to sites with similar climatic conditions and produce superior predictions for models which include sunshine duration (Pérez et al., 2010). This is indicated in the results obtained above. Numerous publications have concluded that temperature-based models exhibit greater predictive capacity than models based on relative humidity parameters (Fan et al., 2019) and the above models verify this conclusion. Overall, the machine learning techniques were able to accurately describe great correlations for sunshine-based and temperature-based models. Relative humidity models still lag in performance but indicate more than acceptable performance using ANN and GRNN models. Fig. 3 illustrates the best performing ANN models for each of the input parameter groupings, while Fig. 4 depicts the ANN architecture for the best performing hybrid model, M5.

3.2. *Prediction models including PM$_{10}$ (Jan 2018–Sep 2020)*

Measured PM$_{10}$ records did not show reduced levels during the most recent COVID-19 pandemic lockdown period (Apr–Sep 2020) as expected. This may be due to the main sources of this emission still being operational during this time (essential services continued to function). With the inclusion of PM$_{10}$ data in the meteorological dataset, ANN models again showed the best performance with hybrid, non-linear M2 outperforming the other models; R^2 = 0.97996, MPE = 0.06199%, as presented in Table 5 above. The evaluated error metrics for the ANN models + PM$_{10}$ are significantly lower than the models which exclude PM$_{10}$. This indicates model improvement by introducing this parameter. Relative humidity models were substantially enhanced by the inclusion of PM$_{10}$, M2: R^2 = 0.98139, MPE = −0.03257%, while the temperature-based models experienced a decrease in performance. This suggests that PM$_{10}$ is sensitive to both temperature and relative humidity, which influences the composition of the pollutant particulates. Relative humidity has a proportionate relationship with particulate matter, and this is evident in the relative humidity models. The SVR model indicated great performance again, with M2 non-linear, hybrid model being the best. Fig. 5 below illustrates the best performing ANN models including PM$_{10}$ data.

4. Discussions and implications

The findings above describe a cost-effective method for predicting solar radiation with minimal computational effort and a high level of accuracy. The machine learning techniques investigated can be employed to sufficiently evaluate the solar capacity available across South Africa and to assist in governmental endeavors to progress towards cleaner energy sources. This country's dependency on non-renewable energy sources is still astonishingly high, while most initiatives on the implementation of renewable energy technologies are still in investigative stages. Knowledge of the available energy potential for specific areas will help identify optimal capture sites, while generalized models will help facilitate solar capacity predictions which are essential for the design and implementation of solar technologies which extend their use across the industrial, commercial and residential sectors. This consequently influences the structure of national energy budgets and renewable energy strategies.

In terms of ease of use, the suggested methods require the least amount of expertise or skill to implement. Further, the machine learning techniques described in this study can be easily transferred to other countries with similar climatic and meteorological conditions. The use of generalized datasets and models allow for solar energy assessment at a national and provincial level, without compromising the comprehensiveness of the individual sites of interest. Following the various advantages associated with the use of ML techniques, this study proposes a simple and robust method for evaluating solar radiation data, specifically in locations which experience the financial and technical limitations related to maintaining meteorological stations which measure solar radiation.

The implications of having a reliable model for approximating solar radiation data (which includes remote locations) allows for a more inclusive assessment of the energy resources and security within the country and thus impacts all future energy proposals and projects. Such models enable and encourage the use of renewable solar energy technologies, which will help reduce the existing burden on the state's energy infrastructure and reduce the costs of electricity. While provincial municipalities attempt to improve living conditions by providing electricity to rural households, the methods described above will provide an easy assessment for the viability of solar technologies within any location. Access to secure electricity supply and a decrease in the strain being experienced by ESKOM are essential components for the development of South Africa, thus creating employment opportunities and supporting the country's economic growth.

5. Conclusion

In this work we provide a unique analysis on the capability of machine learning techniques as an improved method to predict global solar radiation in South Africa, while highlighting comparisons between previously developed empirical models. The results presented compare favorably with that described in literature (Vakili et al., 2016), which conclude that machine learning techniques provide reliable and accurate predictions and are considerably suitable for modeling particulate matter concentration. ANN models have proven superior for the estimation of global solar radiation in South Africa, due to its excellent prediction accuracy and minimal error measures. The ease of implementational and minimal computational effort make this technique most desirable. Non-linear hybrid models, specifically with the input data configuration of M5 have demonstrated most appropriate considering the meteorological parameters measured across the nine provinces of South Africa. M5 which does not include PM$_{10}$ concentration indicated the highest correlation coefficient, R^2 = 0.99852 with MBE = −0.01706.

The inclusion of PM$_{10}$ concentration to the generalized dataset boosted model efficiency by reducing the error metrics and this follows the conclusions made by Suleiman et al. (2018) which indicate improved model performance as a result of including particulate matter

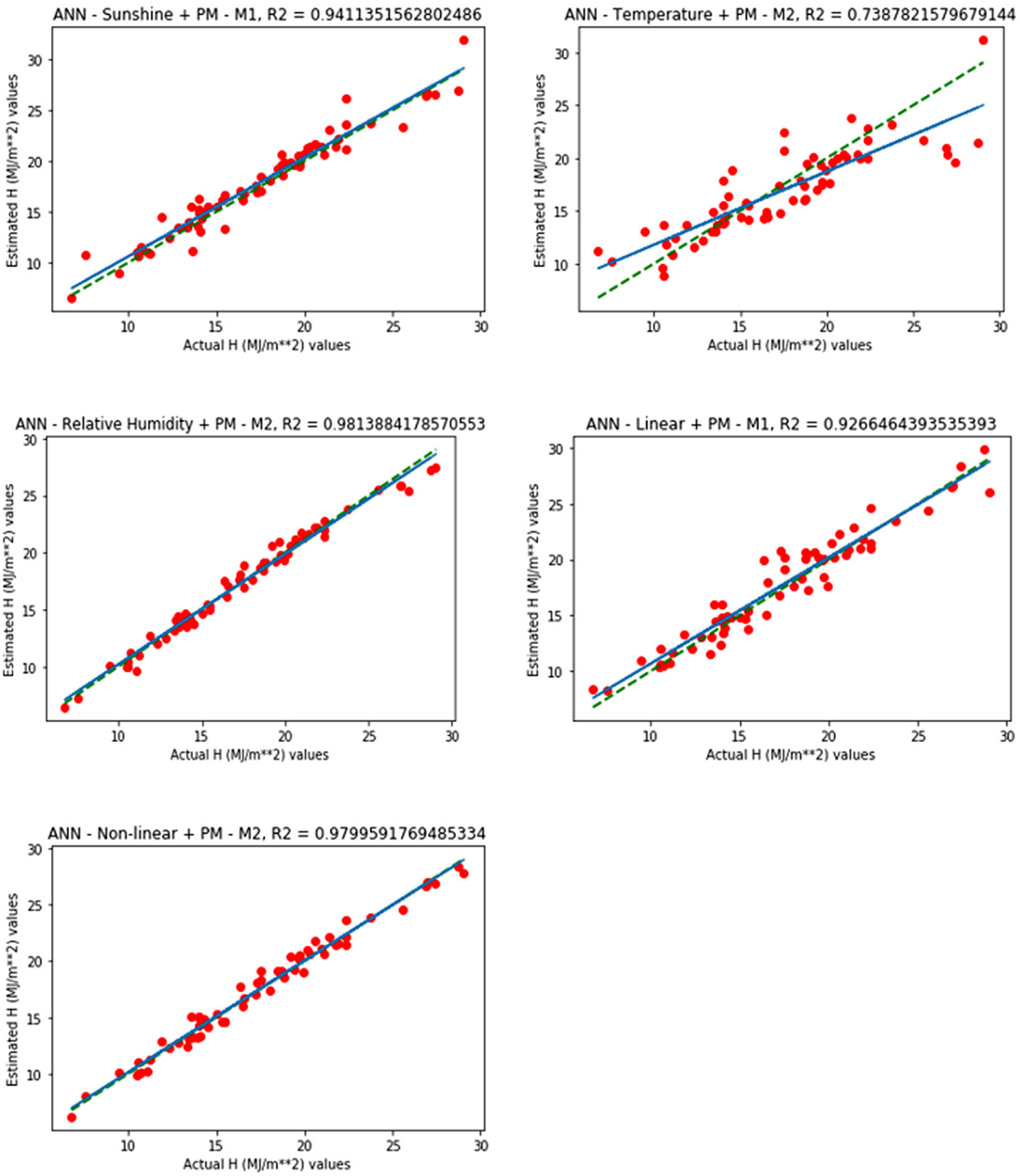


Fig. 5. Best performing ANN models which include PM_{10} data for each parameter configuration type.

concentration as a variable in solar radiation prediction. M2 (includes PM_{10} concentration) highlights this model enhancement through error reduction, resulted in the highest correlation coefficient with the lowest MPE; R2 = 0.97996, MPE = 0.06199%. For improved performance of relative humidity prediction models, it is suggested that PM_{10} data be considered, as shown by the results of this study which verify the proportionate relationship between relative humidity and particulate matter concentration. SVR models are also viable for the prediction of global solar radiation as this model is transferrable to sites with similar climatological conditions.

Machine learning techniques have shown a greater predictive capacity as opposed to the empirical models developed for South Africa. These empirical models are computed for each province and are thus dependent on the site latitude. With the introduction of a generalized dataset, the machine learning techniques are trained on multiple latitudes (all provinces) across the country to provide a comprehensive model for South Africa, without compromising the intrinsic descriptive characteristics of each location. This enables greater predictive capacity and model transferability. Machine learning models are trained on historic data and do not have many dataset limitations, this allows for easier modeling with less computational effort and input requirements.

As a future investigation, it may be worthwhile considering other pollutant concentrations as well as variations of PM size to improve global solar radiation prediction models for South Africa. Modifications to the ANN and SVR models can be introduced through processes such as stacking and boosting. Datasets for periods longer than 5 years (specifically for PM_{10} and other pollutant concentrations) will provide a more significant training set and enhance predictions, provided that this meteorological variable is consistently measured rather than periodically. The availability of reliable meteorological and aerosol data was one of the significant limitations experienced in this study.

Implementation of the proposed hybrid, ANN models through a generalized dataset covering all nine provinces to predict solar radiation will not only enhance the solar reporting capabilities for South Africa, but also allow for an all-inclusive assessment of the energy capacity available across this country. Considering estimations made by this model, the government may benefit by being able to optimally forecast solar resources for various locations including remote and rural areas. The innovative technology highlighted in this work can be used to supplement municipal initiatives currently underway by easily approximating available solar radiation and the feasibility of solar technologies within specific regions, and in turn help decrease the energy poverty gap. The machine learning techniques described in this work are able to sufficiently model the solar resources available in South Africa which can be used to influence various renewable energy strategies aimed at decreasing our consumption of non-renewable resources and the demand currently placed on the state's utility grid. Advances in renewable energy research within South Africa will encourage the inquiry into this unlimited, abundant resource, while helping to sustainably improve the country's energy security and economic growth.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors wish to acknowledge International Water Management Institute (IWMI), Agricultural Research Council (ARC), South Africa Weather Service (SAWS) and South African Air Quality Information System (SAAQIS) for providing the dataset used throughout this study.

References

Abrahamsen, E.B., Brastein, O.M., Lie, B., 2018. Machine learning in Python for weather forecast based on freely available weather data. Proceedings of The 59th Conference on Simulation and Modelling (SIMS 59, 26–28. https://doi.org/10.3384/ecp18153169.

Adeala, A.A., Huan, Z., Enweremadu, C.C., 2015. Evaluation of global solar radiation using multiple weather parameters as predictors for South Africa Provinces. Therm. Sci. 19, 495–509. https://doi.org/10.2298/TSCI130714072A, 03459836.

Besharat, F., Dehghan, A.A., Faghih, A.R., 2013. Empirical models for estimating global solar radiation: a review and case study. Renew. Sustain. Energy Rev. 21, 789–821. https://doi.org/10.1016/j.rser.2012.12.043.

Da Silva, V.J., Da Silva, C.R., Almorox, J., Junior, J.A., 2016. Temperature based solar radiation models for use in simulated soybean potential yield. Aust. J. Crop. Sci. 10 (7), 926–932. https://doi.org/10.21475/ajcs.2016.10.07.p7301.

Debasish, B., Srimanta, P., Dipak, P., 2007. Support vector regression. Neural Information Processing – Letters and Reviews 11. https://www.researchgate.net/publication/228537532_Support_Vector_Regression.

Fan, J., Wu, L., Zhang, F., Cai, H., Wang, X., Lu, X., Xiang, Y., 2018. Evaluating the effect of air pollution on global and diffuse solar radiation prediction using support vector machine modeling based on sunshine duration and air temperature. Renew. Sustain. Energy Rev. 94, 732–747. https://doi.org/10.1016/j.rser.2018.06.029.

Fan, J., Wu, L., Zhang, F., Cai, H., Zhang, W., Wang, X., Zou, H., 2019. Empirical and machine learning models for predicting daily global solar radiation from sunshine duration: a review and case study in China. Renew. Sustain. Energy Rev. 100, 186–212. https://doi.org/10.1016/j.rser.2018.10.018.

Feng, Y., Hao, W., Li, H., Cui, N., Gong, D., Gao, L., 2020. Machine learning models to quantify and map daily global solar radiation and photovoltaic power. Renewable and Sustainable Energy Reviews. https://doi.org/10.1016/j.rser.2019.109393.

Fu, Q., 2003. Radiation (Solar). University of Washington, Seattle, WA. USA, Elsevier, pp. 1859–1863.

Govindasamy, T.R., Chetty, N., 2019. Non-linear multivariate models for estimating global solar radiation received across five cities in South Africa. J. Energy South Afr. 30 (2), 38–51. https://doi.org/10.17159/2413-3051/2019/v30i2a6076.

Gueymard, C.A., 2004. The sun's total and spectral irradiance for solar energy applications and solar radiation models. Sol. Energy 76, 423–453. https://doi.org/10.1016/j.solener.2003.08.039.

Hou, W., Li, D., Yuhuan, Z., Xu, H., Ying, Z., Li, K., Donghui, L., Peng, W., Yan, M., 2014. Using support vector regression to predict PM10 and PM2.5. IOP Conf. Ser. Earth Environ. Sci. 17, 012268 https://doi.org/10.1088/1755-1315/17/1/012268.

Kibirige, B., 2018. Monthly average daily solar radiation simulation in northern KwaZulu-Natal: a physical approach. South Afr. J. Sci. 114 (9/10), 4452. https://doi.org/10.17159/sajs.2018/4452.

Landeras, G., López, J.J., Kisi, O., Shiri, J., 2012. Comparison of Gene Expression Programming with neuro-fuzzy and neural network computing techniques in estimating daily incoming solar radiation in the Basque Country (Northern Spain). Energy Convers. Manag. 62, 1–13. https://doi.org/10.1016/j.enconman.2012.03.025.

Maleki, S.A.M., Hizam, H., Gomes, C., 2017. Estimation of hourly, daily and monthly global solar radiation on inclined surfaces: models re-visited. Energies 134. https://doi.org/10.3390/en10010134.

Maluta, N.E., Mulaudzi, S.T., 2018. Evaluation of the temperature based models for the estimation of global solar radiation in pretoria, gauteng province of South Africa. Int. Energy J. 18, 181–190. http://www.rericjournal.ait.ac.th/index.php/reric/article/view/1668/676.

Maraziotis, E., Marazioti, N., 2008. Statistical analysis of inhalable (PM10) and fine particles (PM2.5) concentrations in urban region of Patras, Greece. Global Nest Journal 10. https://doi.org/10.30955/gnj.000496.

Marwal, V.K., Punia, R.C., Sengar, N., Mahawar, S., Dashora, P., 2012. A comparative study of correlation functions for estimation of monthly mean daily global solar radiation for Jaipur, Rajasthan (India). Indian J. Sci. Techn. 5 (5), 2729–2732. https://doi.org/10.17485/ijst/2012/v5i5.8. ISSN: 0974- 6846.

Mulaudzi, S.T., Sankaran, V., Lysko, M.D., 2013. Solar radiation analysis and regression coefficients for the Vhembe region, Limpopo Province. South Africa J Energy 24 (2–7). ISSN 2413-3051. http://www.scielo.org.za/scielo.php?script=sci_arttext&pid=S1021-447X2013000300001&lng=en&nrm=iso. ISSN 2413-3051.

Nadzri, O., Jafri, M.Z.M., Lim, H.S., 2010. Estimating particulate matter concentration over Arid region using satellite remote sensing: a case study in Makkah, Saudi Arabia. Mod. Appl. Sci. 4 https://doi.org/10.5539/mas.v4n11p131.

Nejadkoorki, F., Baroutian, S., 2012. Forecasting extreme PM10 concentrations using artificial neural networks. Int. J. Environ. Res. 6, 277–284. https://doi.org/10.22059/IJER.2011.493.

Pérez, N., Pey, J., Cusack, M., Reche, C., Querol, X., Alastuey, A., Viana, M., 2010. Variability of particle number, black carbon, and PM10, PM2.5, and PM1 levels and speciation: influence of road traffic emissions on urban air quality. Aerosol. Sci. Technol. 44 (7), 487–499. https://doi.org/10.1080/02786821003758286.

Pinker, R., Frouin, R., Li, Z., 1995. A review of satellite methods to derive surface shortwave irradiance. Remote Sens. Environ. 51, 108–124. https://doi.org/10.1016/0034-4257(94)00069-Y.

Raimondo, G., Montuori, A., Moniaci, W., Pasero, E., Almkvist, A., 2007. Machine learning tool to forecast PM 10 level, 87th AMS Annual meeting. https://ams.confex.com/ams/87ANNUAL/webprogram/Paper119708.html.

Rokach, L., Maimon, O., 2005. Decision trees. In: Maimon, O., Rokach, L. (Eds.), Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA.

Suleiman, A., Tight, M., Quinn, A., 2018. Applying machine learning methods in managing urban concentrations of traffic-related particulate matter (PM 10 and PM 2.5). Atmospheric Pollution Research. https://doi.org/10.1016/j.apr.2018.07.001.

Tymvios, F.S., Jacovides, C.P., Michaelides, S.C., Scouteli, C., 2005. Comparative study of Ångström's and artificial neural networks' methodologies in estimating global solar radiation. Sol. Energy 78, 752–762. https://doi.org/10.1016/j.solener.2004.09.007.

Vakili, M., Sabbagh-Yazdi, S.R., Khosrojerdi, S., Kalhor, K., 2016. Evaluating the effect of particulate matter pollution on estimation of daily global solar radiation using artificial neural network modeling based on meteorological data. J. Clean. Prod. 141 https://doi.org/10.1016/j.jclepro.2016.09.145.

Viorel, B., 2008. Modeling Solar Radiation at the Earth's Surface: Recent Advances. Springer. ISBN: 3540774548.

Wei-Yin, L., 2011. Classification and Regression Trees, vol. 1. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, pp. 14–23. https://doi.org/10.1002/widm.8.

Yeboah, F., Pyle, R., Hyeng, C., 2015. Predicting solar radiation for renewable energy technologies - a random forest approach. Int. J. Mod. Eng. 16, 100–107. ISSN: 1930-6628. https://ijme.us/issues/fall2015/fall_2015.htm.

Zhou, Y., Wang, D., Liu, Y., Liu, J., 2019. Diffuse solar radiation models for different climates zone in China: model evaluation and general model development. Energy Convers. Manag. 185, 518–536. https://doi.org/10.1016/j.enconman.2019.02.013.

Websites

Ritchie, H., Roser, M., 2017. CO_2 and Greenhouse Gas Emissions. Global Fossil Fuel Consumption and Production Database. OurWorldInData.org.. https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions (accessed June 2020).

International Water Management Institute (IWMI). http://www.iwmi.cgiar.org/ (accessed 1 October 2020).

South African Air Quality Information System (SAAQIS). http://saaqis.environment.gov.za/ (accessed 3 October 2020).

Publication-ready NN-architecture schematics. https://alexlenail.me/NN-SVG/ (accessed 10 October 2020).

Wikipedia. https://upload.wikimedia.org/wikipedia/commons/thumb/e/e0/South_Africa_K%C3%B6ppen.svg/1024px-South_Africa_K%C3%B6ppen.svg.png (accessed 08 January 2021).