Multi-ensemble Machine Learning based approach for Short-term Solar Radiation Forecasting

Enzo L. Fernandes[a], Renato F. Cantão[b]

[a]Department of Computation, UFSCar, Sorocaba, Brazil
[b]Department of Physics, Chimestry and Mathematics, UFSCar, Sorocaba, Brazil

ARTICLE INFO

Keywords:
Machine Learning
Supervised Learning
Ensembles
Solar Radiation Forecast
Solar Energy
Renewable Energy

ABSTRACT

The increasing need for energy resources alongside with rising commitment towards environmental awareness have led to the development of various renewable energy source alternatives aiming to progressively replace the current fossil fuel standards. One of these alternatives is the Solar Photovoltaic Energy, a decentralized energy source that depends directly on the solar radiation intensity. This paper introduces a machine learning based approach featuring multiple ensemble predictors for short-term solar radiation forecasting systems using meteorological data. The data used were collected in meteorological stations in the State of São Paulo, Brazil, and consist of observations of meteorological variables recorded hourly in a period corresponding to 2001–2021 in 56 meteorological stations. The proposed system consists of a 2-procedure machine learning pipeline in which models for site-specific and generalization predictors are trained. In each one of these procedures a selection of the best fitted models with the respective hiperparameters is conducted comparing models trained with Random Forests, Extreme Gradient Boosting, Extra Trees, Support Vectors Machines and Neural Networks as well as the Stacking-based ensemble that incorporates the previous algorithms. In the first procedure site-specific models are trained and selected and a prediction set for the respective site is produced; in the second procedure a compositor model is trained on the prediction sets targeting the expected observed values for solar radiation in a *leave one out* configuration. Models of both site-specific and extrapolated predictions have been evaluated in terms of RMSE, MAE, MBE and R^2 Score to determine the best performing models for each procedure and to assess the performance of the system. In order to demonstrate the effectiveness of the proposed system, estimators produced in both procedures are compared with a solar radiation empirical model. The predictions of the generalization procedure were also compared to an Inverse Distance Weighting-based method taking into account not only distance but also the fitness of the estimators.

1. Introduction

One of the great challenges faced by mankind in the current century is achieving the balance between the seek for energy sources, driven by the progressive informatization in society, and the conscientious exploration of natural resources. Although much progress has been made in the recent years, such as the advent of electric cars and the development of novel renewable energy sources [34, 12], most of the countries in the world still predominantly rely on fossil fuels as their primary energy source REFRENCE. On the other hand, Photovoltaic Solar Energy – a renewable, low environmental impact and relatively accessible energy source – has been gaining attention of the public as a reflect of the recent development breakthroughs [29, 15] and rising adoption of this technology in houses and industries. Its versatility allows for use cases that range from household energy generation to industry-scale smart grid connected solar farms and, given its characteristics, it is considered a promissing fossil fuel replacement in the long term.

However, one of the inherent issues of this technology is the direct dependence on the amount of solar radiation avaliable at the generation site in order to generate electic energy. Certain aspects of this relation between site-specific characteristics and the amount of solar radiation have been mapped aiming to provide long term solar radiation predictions. Yet, when it comes to the the integration of the distributed generation agents into the energy grid, greater forecasting granularity is demanded in order to actively assist the management of both infrastructure and smart grid systems. In this context, the capability of accurately predicting the amount of solar radiation available for the photovoltaic panels in the short term implies not only in greater control over power supply policy scheduling by smart grid system managers but also introduces a strategic advantage for generation agents in electric energy trading markets as part of decision making processes.

One of the possibilites of the solar ration forecasting processes is to employ Machine Learning (ML) techiniques with the objective of obtaining a hypotesis function $h(\cdot)$ capable of reproducing the bahavior of the global solar radiation with respect to a given set of input variables regarding meteorological conditions of the prediction site. In this context both inputs, namely meteorological variables, and the excpected outputs – global solar radiation – are to be known *a priori* when training the ML models. Such forecasting task can be carried out in different manners taking as input different representations of the meteorological conditions aiming to predict global radiation with different forecast horizons into the future. With regard to the forecast horizon, different periods in the future can be used to forecast solar radiation. The study conducted by [24] demonstrated the effec-

⋆ Funded by FAPESP, grant number 2020/09607-9.
ORCID(s):

tiveness of models such as SVMs and Hidden Markov Models for solar radiation forecasting 5 and 30 minutes in the future using data collected from nearly 18.000 sites across Australia. The models were developed as part of a solar prediction platform taht aimed to optimize decision making in the Australian electricity market matching the prediction horizons with the respective dispatch and trading price time-windows. Another application of these techniques was demonstrated by [13] as an effort to map the geographical distribution and variation of daily global solar radiation and photovoltaic power potential in the region of Loess Plateau, Northern China. The experiments conducted used models such as Extreme Learning Machines, SVMs and Generalized Learning Neural Networks applied to hourly records of meteorological variables collected in 57 meteorological stations during the period corresponding to 1961–2016. The predictions produced were further processed using spatial interpolation techniques aiming to generalize the predictions in the geographical space obtaining mappings in different areas and seasons of the year for the study area. In different studies the data used to train the models may be arranged in different formats aiming to benefit from the cyclical behavior presented by the targeted phenomenon. One common approach is to use time series data to extract the patterns from the training data. This approach has been applied in both [2, 22] in order to predict solar radiation. The former conducted experiments using Recurrent Neural Networks in high-resolution[1] time series data collected in two solar farms in Canada whereas the latter applied SVMs, NNs, and Fuzzy Inference Systems as well as its variants to time series data in order to estimate the hourly solar radiation in Abu Musa Island, Middle East. In both studies, the empirical models used as reference have been outperformed by the ML models applied to time series data. Another approach is to use different kinds of input variables to the models other than the ones acquired by ground meteorological observations. In [8] an approach using satellite data collected by the meteorological satellites operated by EUMETSAT – European Organization for the Exploration of Meteorological Satellites – with data regarding reflective information in various visible spectral channels was conducted to estimate hourly global radiation in the region of Toledo, Spain. The satellite data were complemented by site-specific information regarding cloud index and clear sky radiation models. Throughout the experiments conducted different numbers of input variables were used to train models such as NNs, Gaussian Processes, SVMs and Extreme Learning Machines which performances were later assessed by physical models.

Considering the vast horizon of the ML landscape, different strategies and model combinations can be carried out aiming performance gains for forecasting improvements. As an example, [5] applies combinations of estimators as part of an ensemble-based model in which predictions made by trained models are used to train a *meta-learner* – a model responsible for combining the produced predictions providing better generalization capabilities – for solar radiation forecasting. With respect to the input variables for $h(\cdot)$, [25] conducts a study regarding the selection of the variables with the objective to optimize the performance of an Adaptive Neuro-Fuzzy Inference System. A similar strategy is applied by [38] when training SVMs with different sets of input variables. Other approaches consider using different techniques instead of ML as an attempt to obtain $h(\cdot)$. In particular [21] develops a Fourier-Series based generic model to address daily expected global solar radiation forecasting via "a set of sine and cosine harmonics in the temporal and spatial domain" [21]. Such model – parameterized by the day of year, altitude and latitude of the prediction site – was fitted on data collected in 53 meteorological stations around the planet. The results obtained using this approach show that the proposed model outperformed the reference empirical models used.

Given the studies accomplished by peers, the approach introduced by this paper consists in a hourly solar radiation forecasting system with the extensive use of ML-based estimators in order to both provide site-specific predictions and generalized prediction through the use of numerous ensembles; Section 2 introduces the solar prediction architecture, Section 3 introduces the ML algorithms used to train the models, Sections 4 and 5 present details the area of study and the preprocessing, respectively and Section 7 presents the results obtained during the study.

2. Solar Prediction System Architecture

The proposed ML-based forecasting system is composed of two stacked procedures: P1 and P2. In each one of these procedures Supervised Leaning techniques have been applied in order to prepare the data and obtain models capable of estimating R.

P1: Site-specific models trained and tuned with historical meteorological data collected in each one of the collection sites in the area of study. Predictions are site-specific to each collection sites.

P2: A single generalization model trained and tuned with the predictions of the site-specific models in the previous procedure. Predictions correspond to generalizations in the geographical space of predictions generated by models obtained in P1.

Figure 1 depicts the overall processes performed in both procedures.

3. Prediction Methods

3.1. Neural Networks

Artificial Neural Networks [17] are optimization-based computational models capable of simulating the behavior of biological neurons through the combination of several processing units denominated *perceptrons*. Each perceptron is responsible for computing the *activation* of the weighted sum

[1] Meteorological variables were sampled at a rate of 100Hz in each of the data collection sites.

Nomenclature

Acronyms

DOY	Day of Year
EMP	Empirical Model
IDW	Inverse Distance Weighting
ML	Machine Learning
NN	Neural Network
P1	Procedure 1 (Site-specific models)
P2	Procedure 2 (Generalization model)
PCA	Principal Component Analysis
RF	Random Forest
SVM	Support Vector Machine
XGB	Extreme Gradient Boosting

Performance Metrics

MAE	Mean Absolute Error
MBE	Mean Bias Error
R^2	R^2 Score/Coefficient of Determination
RMSE	Root Mean Squared Error

Variables

$\hat{y}_i$	i-th prediction of R in prediction set
$\overline{y}$	Mean value of y in test set
ψ	Generalized prediction of R
doy	Day of year
h	Hour of day
La	Latitude
Lo	Longitude
R	Global solar radiation
y_i	i-th observation of R in test set

of its inputs with respect to an activation function – examples of activation functions are provided by Eqs. 2 and 3. This value is then propagated though the network and used as input to other perceptrons.

$$a_p^{(l)} = g(z) \quad \text{where} \quad z = \sum_{i=1}^{n} \left(a_i^{(l-1)} w_i^{(l)} \right) + b^{(l)} \quad (1)$$

These processing units are arranged across the network as sets of layers in such way that every perceptron in the n-th layer is connected to all perceptrons in layer $n+1$. Eq. 1 presents the activation $a_p^{(l)}$ of the p-th perceptron in the l-th layer of the network as well as the weights $w_i^{(l)}$ and the *bias* term $b^{(l)}$ for layer l. In this context, NN are composed by an input layer with the same number of perceptrons as the number of features in a training instance; an output layer with the same number of perceptrons as the number of expected output values; and a variable number of intermediate layers – denominated *hidden layers* – each with an arbitrary number of perceptrons. In such achitecture, communly referred to as Dense Neural Network, activations are propagated from input to output in a process designated *feed-forward*.

$$g_1(z) = \frac{1}{1+e^{-z}} \quad (2)$$

$$g_2(z) = \max(0, z) \quad (3)$$

The training process of NNs consists in obtaining optimal values of $w_i^{(l)}$ aiming to minimize a loss function calculated with the NN output values. This process is conducted by adjusting the weights based on the partial derivative of the loss function with respect to all values of $w_i^{(l)}$ in the network from the output layer towards the input layer (except the input itself). This NN training technique is known as *back-propagation*.

3.2. Random Forests

Random Forests [4] correspond to Machine Learning ensembles composed of an arbitrary number of Decision Tree estimators. The atomic estimators are fitted in bootstrapped samples of the dataset and the predictions of each individual estimator are averaged by the Random Forest in order to provide better predictive fitness and prevent overfitting. In order to construct each Decision Tree split candidates with random subsamples of features with pre-determined size are compared to splits containing all available features selecting the candidate that provides the most descriminative split in relation to the expected values in the bootstrapped sample.

3.3. Extra Trees

Extremely Randomized Trees – Extra Trees – are Random Forests ensemble variants in which further randomness is introduced during the construction of the Decision Tree at the moment of splitting the nodes. As opposed to RFs, Extra Trees perform splitting by selecting thresholds at random in non-bootstrapped samples of the original dataset instead of evaluating each split candidate threshold with an heuristic. This increased amount of randomness is often reflected in further reduction of the variance of the ensemble although minor increase in bias may be experienced.

3.4. Extreme Gradient Boosting

eXtreme Gredient Boosting Trees [6] are supervised learning algorithms based on Gradient Boosting Decision Trees [16] and Random Forests in which a set of weak estimators is iteractively fitted through the data. This additive learning process is conducted in such way that the first estimator is fitted with the all available training instances and the subsequent estimators are iteractively added to the ensemble in order to fit the residual errors observed in the previous estimators. This process is repeated until a stopping condition

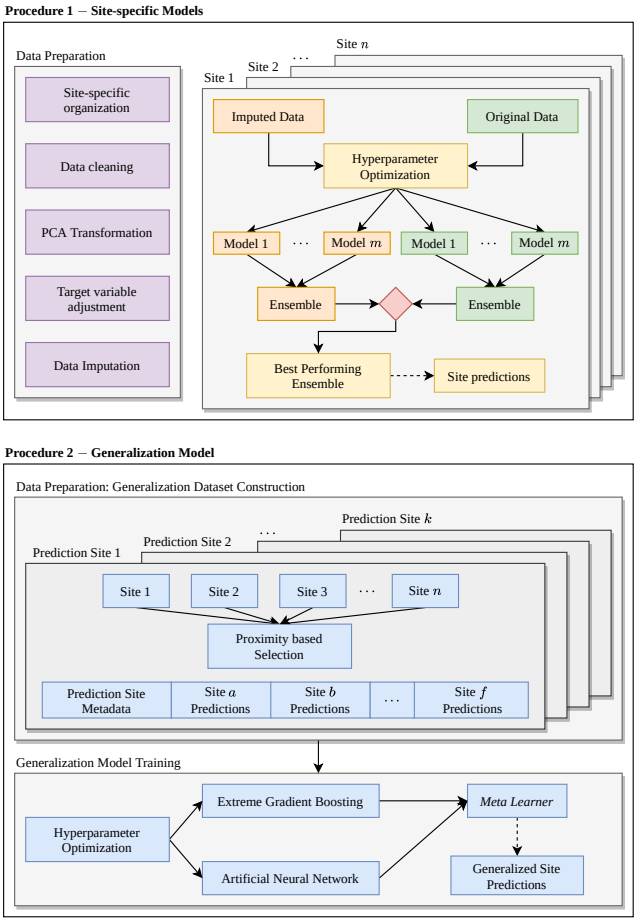

Figure 1: Schematic overview of procedures P1 (site-specific models) and P2 (geographical generalizations based on P1's predictions).

is reached.

$$\hat{y}_i = \phi(\boldsymbol{x}_i) = \sum_{k=1}^{K} h_k\left(\boldsymbol{x}_i\right) \tag{4}$$

$$L^{(i)} = \sum_{i=1}^{k} l\Big(y_i, \hat{y}_i^{(t-1)} + h_i\left(\boldsymbol{x}_i\right)\Big) + \Omega(h_i) \tag{5}$$

With the purpose of employing additive learning the regularized cost function is modified in order to account for the predictive iterations performed by previous estimators h_i in the data (Eq. 5 in which Ω is the regularization term for the estimator h_i). In this configuration, the ensemble prediction is carried out by the sum of the predictions of each one of the K atomic estimators h_i, as described by Eq. 4.

3.5. Support Vector Machine

Support Vector Machines [9] are supervised learning algorithms capable of performing highly non-linear data fitting through the transportation of training instances to higher dimensional spaces denominated *feature-spaces*. In this context, let n be the number of training instances and m the number of features in each instance $x^{(i)}$; in such transformation each one of the n training instances is transported from a m to a n-dimensional space in which fitting takes place on the new training instances $l^{(i)}$.

$$f^{(i)} = \exp\left(-\frac{\|x^{(i)} - l^{(i)}\|_2}{2\sigma^2}\right) \tag{6}$$

Once the training instances undergo the transformation the fitting process is performed on the values calculated by a *kernel-function* that estimates the *similarity* $f^{(i)}$ between each training instance and all instances in the feature-space – as en example of kernel function, Eq. 6 presents the *radial basis function*.

$$h_{\vec{\theta}}(x) = \sum_{i=1}^{n} \theta_i \cdot f^{(i)} \tag{7}$$

After the feature-space samples are obtained the fitting process is conducted by optimizing the values of $\vec{\theta}$ in order to minimize the squared error cost [36]. Once such process is complete inference takes place according to Eq. 7.

3.6. Ridge Regression

Ridge Regression corresponds to the Linear Regression [31] with the addition of regularization in the weights $\vec{\theta}$ used to perform the linear transformation $h_{\vec{\theta}}(x)$ over the input features $\vec{x}$. Under these circunstances, the loss function L with respect to the parames $\vec{\theta}$ used to optimize the weights is modified according to Eq 8 in order to include a regularization parameter α which restrics the degree of freedom of the parameters during training. As a result, regularization reduces the variance of the estimators' predictions providing robustness against overfitting [32].

$$L_{\vec{\theta}}\left(\vec{x}\right) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_{\vec{\theta}}\left(x^{(i)}\right) - y^{(i)}\right)^2 + \overbrace{\alpha \sum_{j=1}^{n} \theta_j^2}^{\text{regularization}}\right] \tag{8}$$

Once the loss function is accordingly modified the training is identical to ordinary Linear Regression and inference is performed by the dot product between $\vec{\theta}$ and $\vec{x}$.

3.7. Stacking Ensemble

Stacking ensemble generelizations correspond to supervised learning *meta-models* capable of aggregating predictions performed by indovidual estimators in order to produce a final prediction. Such aggregation is performed by an intermediate estimator called *meta-learner* trained on the outputs of the indovidual estimators targeting the excpected values for each training instance.

$$\left(h_1(\vec{x}), h_2(\vec{x}), \cdots, h_n(\vec{x})\right) \xrightarrow{meta-learner} h^*(\vec{x}) \tag{9}$$

Such approach allows one to combine different characteristics of individual algorithms into one estimator aiming to achieve better generalization capabilities [11].

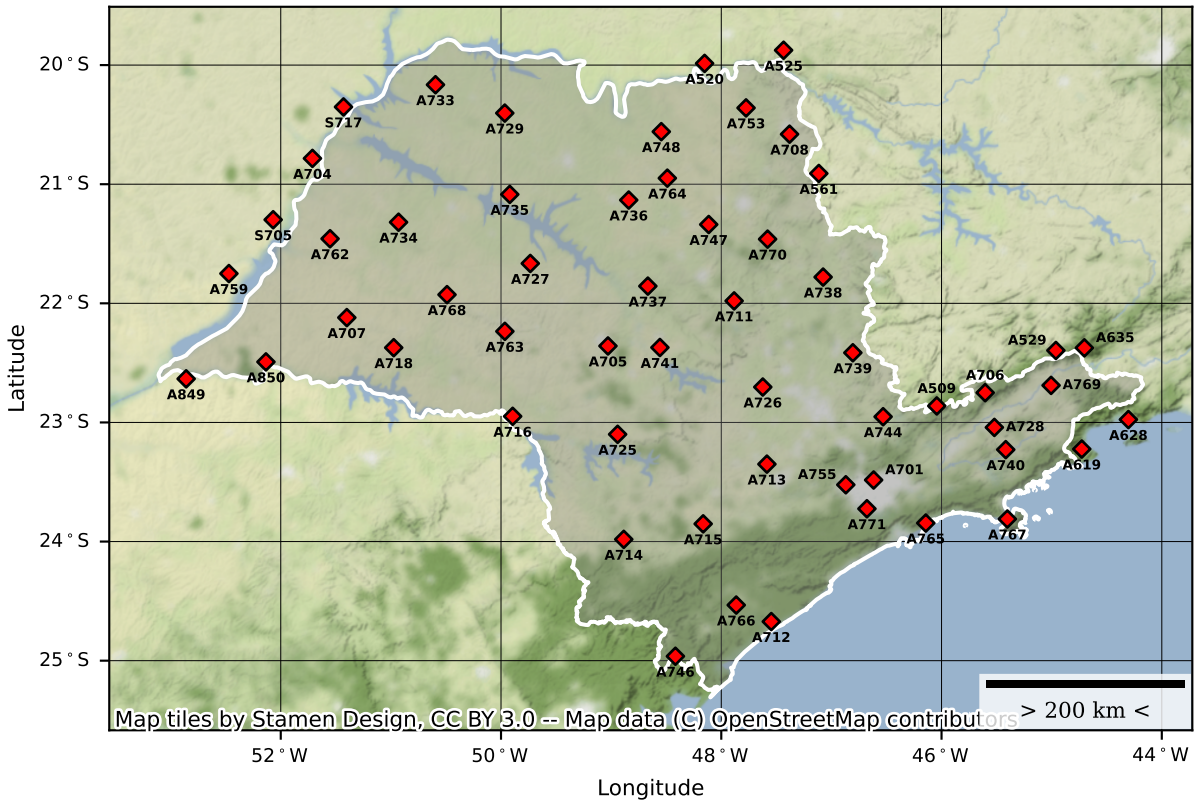

Figure 2: Overview of the study area corresponding to the State of São Paulo, Brazil. In the map, the red spots denote the location of the meteorological stations with the respective site IDs.

3.8. Inverse Distance Weighting Variant

Such method consists of an aggregation based on IDW [33] with respect to the predictions produced by the best estimators in each site of P1 taking into account not only the distance from the site-specific estimator to the prediction site but also the fitness of the estimators themselves. In this context, fitness is calculated as the arithmetic mean of the metrics RMSE and MAE for every best estimator. The weight w_i of a site s_i in relation to a prediction site s_0 is presented by Eq. 10 in which $d(\cdot)$ denotes the Haversine Distance and $e^{\star}_{s_i}$ the best estimator for s_i. In order to limit number of estimators used during weight calculation a distance threshold was introduced.

$$w_i(s_i, s_0) = \frac{1}{d(s_i, s_0)^2} \left(\frac{\mathrm{RMSE}(e^{\star}_{s_i}) + \mathrm{MAE}(e^{\star}_{s_i})}{2} \right)^{-1} \quad (10)$$

Once the weights for the predictions of the best estimator of every selected site s_i are calculated the remaining steps to obtain the generalized prediction ψ in the site s_0 are identical to those in ordinary IDW and are presented by Eq.11 in which k denotes the number of selected site predictions to aggregate.

$$\psi_{ts+1}(s_0) = \frac{\sum_{i=1}^{k} \left(\overbrace{e^{\star}_{s_i}(x_{s_i})}^{\hat{R}^i_{ts+1}} \cdot w_i(s_i, s_0) \right)}{\sum_{i=1}^{k} w_i(s_i, s_0)} \quad (11)$$

3.9. CPRG Model

The Collares-Pereira and Rabl model [7] modified by Gueymard – CPRG model – is a decomposition-based hourly solar radiation prediction model [14]. Different studies – *e.g.* [37, 3] – have compared the effectiveness of this model against its peers resulting in better performance with the CPRG model. Such model is defined by the following equation:

$$\frac{l}{H} = \frac{(a + b\cos W) \cdot r_0}{f_c} \quad (12)$$

In Eq. 12 $\frac{l}{H}$ is the ratio between the hourly and the total global solar radiation for the day and a and b are regression coefficients:

$$a = 0.4090 + 0.5016 \cdot \sin\left(W_s - 60°\right) \quad (13)$$

$$b = 0.6609 + 0.4767 \cdot \sin\left(W_s - 60°\right) \quad (14)$$

In Eq. 12 f_c and r_o are calculated according to Eqs. 15 and 16, respectively:

$$f_c = a + b \cdot \frac{\frac{\pi W_s}{180} - \sin W_s \cdot \cos W_s}{\sin W_s - \frac{\pi W_s}{180} \cdot \cos W_s} \tag{15}$$

$$r_0 = \frac{\pi}{24} \cdot \frac{\cos W - \cos W_s}{\sin W_s - \frac{\pi W_s}{180} \cos W_s} \tag{16}$$

In Eqs. 16 and 15, the solar hour angle W and the solar sunset hour W_s (both in degrees) are given by Eqs. 17 and 18, respectively:

$$W = \frac{360 \cdot (t_s - 12)}{24} \tag{17}$$

$$W_s = \arccos\big(-\tan\varphi \cdot \tan\delta\big) \tag{18}$$

In Eqs. 16 and 15, φ, δ, t_s denote the latitude (degrees), the solar declination angle (radians) and the solar time, respectively. The latter can be calculated according to Eq. 19 as described in [26].

$$t_s = LT + \frac{ET}{60} + \frac{4}{60}\big(L_s - Lo\big) \tag{19}$$

In Eq. 19 LT is the local standard time, L_s is the meridian timezone for the local time, Lo is the latitude of the prediction site (degrees) and ET is the time equation calculated according to Eq. 20.

$$ET = 9.87\sin\big(2B\big) - 7.53\cos B - 1.6\cos B \tag{20}$$

$$B = \frac{360 \cdot (doy - 81)}{365} \tag{21}$$

According to [35], the value of the solar declination angles δ (radians) in Eq. 18 can be calculated by Eq. 22.

$$\begin{aligned}\delta = \left(\frac{180}{\pi}\right) \cdot \big[0.006918 \quad &- \quad 0.399912\cos\Gamma \\ + 0.070257\sin\Gamma \quad &+ \quad 0.006758\cos 2\Gamma \\ + 0.000907\sin 2\Gamma \quad &+ \quad 0.002697\cos 3\Gamma \\ + 0.001480\sin 3\Gamma \big]\end{aligned} \tag{22}$$

$$\Gamma = \frac{2\pi(doy - 1)}{365} \tag{23}$$

In Eq. 23 Γ denotes the day angle (radians) and doy the day of the year.

4. Study Area and Data Collection

The area selected for this study corresponds to the State of São Paulo, Southeastern Brazil (Figure 2). This region is characterized by the predominance of dry-winter humid subtropical climate REFERENCE (classified as *Cwa* according to the Köppen-Geiger climate classification system REFERENCE) and known for its hot and humid Summer and mild and dry Winter.

Meteorological variable		Unit
B	Barometric pressure*	hPa
D	Dew point*	°C
H	Humidity*	%
P	Precipitation	mm
T	Temperature*	°C
R	Global solar radiation	kJ m^{-2}
W_s	Wind speed	m s^{-1}
W_d	Wind direction	°

Table 1
Meteorological variables obtained from the meteorological stations. Variables marked with (*) contain measures of the instant value as well as the maximum and minimum values of the previous hour.

The data used to train the site-specific models in P1 consisted of observations of meteorological variables such as temperature, atmospheric pressure and solar radiation, collected on an hourly basis by the National Institute of Meteorology of Brazil [27] through its meteorological stations spread across the country. Such collected data are part of a wider set of information collected by the Institute as an effort to constructively influence the decision-making processes for industries and institutions through monitoring, analysis and forecasting of the weather with the purpose to embrace sustainable development REFERENCE. The historical observations used in this study can be downloaded at no cost in INMETs website [28].

In total 56 data collection sites in the State of São Paulo have been selected to acquire the data used during training and evaluation of the predictors. Although not all meteorological stations had their historical records in the same time frame, the period of observations ranges from January of 2001 to July of 2021.

5. Preprocessing

The obtained dataset consisted of hourly observations of meteorological variables present in Table 1, as well as the timestamps for each observation, separated by station and year. Further information such as latitude, longitude and altitude have been acquired as metadata.

For P1 the preprocessing operations consisted of grouping the observations by station, removing inconsistent values, adjusting the values of R for the target variable and generating the timestamps. The features B, H, T and D have been PCA-transformed [19] in order to reduce their dimensionality from 3 dimensions (instant value, maximum and minimum measures of the previous hour) to one dimension in T, B and D and two dimensions in H. During preprocessing of this procedure samples were grouped by geographical location and arranged in time order to obtain the target variable – *i.e.* $\hat{R}_{ts+1}$ for the next hour – in relation to a timestamp ts in the format (doy, h).

$$\big(ts, B, D, H, P, T, R, W_s, W_d, R\big) \rightarrow \hat{R}_{ts+1} \tag{24}$$

The dataset used in P2 was obtained from the predictions of the selected estimators in the previous procedure (selection is further explained in Section 6.2) as well as information such as the errors of the estimators and the distance between the prediction point – given by (La, Lo) – and the locations used to train site-specific estimators. For each one of the prediction points a set of closest site-specific estimators has been obtained based on the Haversine Distance with a threshold of 120 km in order to compose the dataset with the format shown by Eq. 25. In this context, n was limited to a maximum of 4 stations and the 3-tuple $(d^n, E^n, \hat{R}^n_{ts_i})$ – denoting the Haversine Distance between the prediction point and the n-th estimator, the error of the n-th estimator and the prediction of the prediction of the i-th estimator for the n-th timestamp, respectively – were ordered by distance in such way that station 1 was the closest to the prediction point.

$$\left(ts_i, La_i, Lo_i, \underbrace{d^1, E^1, \hat{R}^1_{ts_i+1}}_{\text{station 1}}, \cdots, \underbrace{d^n, E^n, \hat{R}^n_{ts_i+1}}_{\text{station } n}\right) \rightarrow \psi_{ts+1} \quad (25)$$

When the number of stations meeting the distance requirement was inferior to 4 the 3-tuples were zero-padded to the right preserving the order based on distance. Samples with fewer than 3 stations were discarded.

5.1. Data Imputation

In an attempt to reduce the impact of corrupted and/or missing data during the training process in P1 an imputation method was introduced. Such method consisted of applying IDW interpolation REFERENCE selecting for each one of the timestamps ts with missing values of feature f in the interpolated station s_0 the closest nearby stations based on La and Lo with available records of f at ts. The distance threshold for selecting the nearby stations and the minimum number of available records in order to perform the imputation were 120 km and 3 records, respectively. This procedure, presented in Eq. 26 where S is the set of closest stations meeting the selection criteria and d denotes the Haversine Distance, was repeated for each one of the features and stations in the dataset.

$$s_0.f_{ts} = \text{IDW-interpolate}\left(d\left(s_0, S\right), S_{f_{ts}}\right) \quad (26)$$

In order to assess the effectiveness of this imputation method both interpolated and original data have been used during training in P1. The estimators produced using both datasets – *i.e* interpolated and original – were compared and selected based on performance in such way that the estimators trained by interpolated data were only used when there was a performance improvement over the ones trained with original data.

6. Train and Test Set Preparation

The train and test sets used in both P1 and P2 were split according to the timestamps of their samples. For P1, the dataset was split so that the samples recorded during the interval 2019-2021 were used during test routines, and the remaining samples – *i.e.* from the beginning of the records of each station to 2018 – were used during training. Considering the imputed data, the average train-test-holdout percentage across the selected stations was ∼85% for training and ∼15% for testing. For P2, the dataset was grouped by prediction site in a *leave-one-out* configuration with a same train-test time intervals as P1: for each prediction site in P2 a group of nearby stations was selected to compose the input dataset using as target the real observed values in the *left out* station during the test period (the same target values used for P1); this process was iteractively repeated obtaining for every timestamp corresponding to the observed values of the R in the test set of the *left out* station predictions carried out by the nearby site-specific models at the same timestamps as described in Eq. 25. The dataset for each station was later sorted by timestamp and the train and test samples were selected iteractively in such way that for every 5 consecutive samples in the ordered grouped dataset 4 samples were selected for training and 1 for testing. This set of grouped datasets was subsequently concatenated producing the test and train sets for P2. Once the splits were performed the samples in training sets of both P1 and P2 were shuffled.

6.1. Data Normalization

Many ML algorithms require the data to be scaled in restricted ranges in order to optimize the fitting process of the produced estimators REFERENCE. In this study two normalization methods were applied: in P1 data were normalized according to function $f_1(\cdot)$ (Eq. 27), often referred to as *Robust Scaling*; in P2 data were normalized according to function $f_2(\cdot)$ (Eq. 28), also known as *Min-Max Scaling* [10]. In both equations m denotes the number of features in the datasets and X_i is a vector with values of the i-th feature for every observation.

$$f_1(X_i) = \frac{X_i - Q_1(X_i)}{Q_3(X_i) - Q_1(X_i)} \quad \forall i \in \{1, \cdots, m\} \quad (27)$$

$$f_2(X_i) = \frac{X_i - \min(X_i)}{\max(X_i) - \min(X_i)} \quad \forall i \in \{1, \cdots, m\} \quad (28)$$

6.2. Model Tuning and Selection

During the execution of the training routines, methods for hyperparameter optimization were introduced in order to improve the generalization capabilities of the estimators.

For the estimators in P1, Exhaustive Grid Search was performed for every one of the selected supervised learning algorithms in each selected location; the respective routines were executed with training set using 5-fold cross-validation [23] obtaining from the search space a set of best hyperparameters. The search space of hyperparameters for each algorithm is shown in Table 2. Once the hyperparameter optimization routines were accomplished estimators of the same site (including the ensemble using all the available estimators with optimal hyperparameters obtained during Grid Search)

Algorithm	Hyperparameter	Tested Values
SVM[a]	C	1, 2, 5
	gamma − γ	scale, auto
	epsilon − ϵ	0.1, 0.15, 0.2, 0.4
Extra Trees[b]/RF[c]	min_samples_split	2, 20, 100, 250, 500
	min_samples_leaf	1, 10, 50, 150, 500
	max_features	auto, sqrt
	n_estimators	100, 150, 200, 400
XGB[d]	eta − η	0.3, 0.1, 0.15, 0.35
	gamma − γ	0, 0.05, 0.01
	sampling_method	uniform, gradient_based
	max_depth	6, 8, 10, 12
NN[e]	learning_rate	constant, adaptative
	solver	ADAM, SGD, LBFGS
	hidden_layer_sizes	(100,), (200,), (100,100), (150,150), (50,50,20), (30,30,15)
	activation	relu, sigmoid

[a] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html
[b] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesRegressor.html
[c] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
[d] https://xgboost.readthedocs.io/en/latest/parameter.html
[e] https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html

Table 2
Hyperparameter search space presented in terms of the nomenclature adopted in the respective libraries used to train the models. Further information can be obtained from the references in each training algorithm.

were compared in terms of RMSE. The selected model in each one of the training sites was the one with the smallest RMSE. This process was conducted for both original and imputed data in P1 as described in Section 5.1.

For the estimators in P2 two supervised learning algorithms have been selected to train the models in the ensemble: XBG and NN. For these two algorithms the hyperparameter search was conducted via Bayesian Optimization implemented by the Optuna Framework [1] resulting in the hyperparameters presented by Table 3. For the *meta-learner*, *Ridge Regressoin* was used performing 5-fold cross validation using three different values for α: 0.1, 1, 10.

7. Results

In the next sections predictions produced by both site-specific and generalization models are evaluated in terms of MAE (Eq. 29), RMSE (Eq. 32), MBE (Eq. 30) and R^2 (Eq. 31). With the exceptino of R^2 all metrics are presented in terms of kJ m^{-2}.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{29}$$

$$\mathrm{MBE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right) \tag{30}$$

Algorithm	Hyperparameter	Value Obtained
NN[a]	solver	sgd
	learning_rate_init	0.0470
	momentum	0.3631
	power_t	0.4926
	alpha − α	0.0298
XGB[b]	n_estimators	500
	eta − η	0.1135
	max_depth	9
	subsample	0.7748
	min_child_weight	5.8586
	alpha − α	3.9644
	lambda − λ	0.5002

[a] https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html
[b] https://xgboost.readthedocs.io/en/latest/parameter.html

Table 3
P2 ensemble estimator hyperparameters obtained after hyperparameter search routines conducted by Optuna Framework. Hyperparameters are presented in terms of terminology adopted by the respective libraries.

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{31}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{32}$$

7.1. Site-specific Models

The estimators obtained after the training processes in P1 corresponded to site specific models cabable of estimating R from the meteorological information available locally. The results in this Section show the performance metrics for the best site-specific estimators after the estimator selection. The performance metrics for each selected estimator as well as the comparison between the produced estimators and the empirical model in each site are shown in Table 4. Results obtained show that the ML models have outperformed the empirical model in terms of site-specific forecasting with emphasis in the drastic reduction in MBE. A summarized comparison between both methods is presented in Table 5.

During performance evalutation of models in P1 a particular pattern was observed regarding the residual error plots generated from the estimators predictions. Such pattern, depicted by Figure 3 using as example Station ID A711, was observed in all predictions sites in P1 regardless of the algorithm used to train the estimators. Such pattern matches the mean monthly precipitation profile illustrated by Figure 4 and exposes the fact that the increased amount of precipitation, and hence variation in the atmospheric conditions with respect to R, introduces noise that can't be properly fitted by the estimators even after the execution of hyperparameter optimization routines leading to worse performance under such ciscunstances.

Site ID	Latitude	Longitude	Performance Metrics								Imputation
			RMSE		MAE		MBE		R2 Score		
			ML	EM	ML	EM	ML	EM	ML	EM	
A726	−22.703	−47.623	358.103	584.131	229.882	471.487	−2.347	254.952	0.890	0.774	✓
A509	−22.861	−46.043	417.478	615.828	281.013	469.141	19.648	221.684	0.833	0.741	✓
A520	−19.986	−48.151	359.597	466.936	230.494	410.024	10.367	173.171	0.873	0.779	✗
A525	−19.875	−47.434	388.888	467.702	252.522	421.816	−6.211	228.672	0.872	0.843	✗
A529	−22.396	−44.962	412.346	578.425	265.618	344.054	23.415	193.879	0.846	0.681	✓
A619	−23.223	−44.727	362.237	642.057	228.034	487.656	−1.280	205.239	0.895	0.691	✓
A701	−23.483	−46.617	421.693	556.146	284.930	391.386	−52.428	217.760	0.868	0.822	✗
A704	−20.790	−51.712	352.585	796.970	226.271	722.238	−10.882	197.277	0.892	0.491	✓
A705	−22.358	−49.029	372.524	358.215	240.827	280.787	28.646	115.390	0.876	0.783	✓
A707	−22.120	−51.400	361.240	849.888	233.295	724.310	22.135	134.011	0.878	0.105	✗
A708	−20.580	−47.380	417.463	543.185	272.209	485.250	5.130	208.487	0.839	0.748	✓
A711	−21.980	−47.883	358.865	448.547	230.721	401.742	2.134	184.363	0.884	0.786	✗
A712	−24.717	−47.550	359.983	379.281	236.091	285.211	0.240	118.136	0.872	0.840	✗
A713	−23.350	−47.667	354.795	535.243	231.870	457.385	−1.942	195.488	0.885	0.775	✗
A714	−23.981	−48.885	385.079	457.955	260.001	380.039	44.693	155.982	0.875	0.792	✓
A715	−23.851	−48.164	362.568	466.570	235.982	381.818	12.802	143.792	0.886	0.769	✓
A716	−22.949	−49.894	341.682	518.743	218.826	432.533	3.772	151.890	0.895	0.625	✗
A718	−22.372	−50.974	334.468	641.741	212.978	574.036	5.757	169.343	0.896	0.653	✓
A725	−23.100	−48.946	360.657	484.249	231.852	396.253	−20.742	146.249	0.889	0.736	✗
A727	−21.665	−49.734	335.416	570.102	211.888	484.947	15.673	220.664	0.908	0.763	✗
A728	−23.042	−45.520	370.028	355.876	235.244	291.556	−18.647	172.762	0.885	0.883	✓
A729	−20.403	−49.966	339.960	477.771	218.334	413.774	35.733	165.184	0.896	0.773	✗
A733	−20.165	−50.595	333.948	633.935	209.043	546.672	22.136	227.390	0.898	0.730	✗
A734	−21.319	−50.930	346.816	670.457	220.039	601.199	−12.924	215.652	0.890	0.663	✓
A735	−21.086	−49.921	388.757	646.704	246.028	534.314	−2.927	196.913	0.891	0.684	✗
A736	−21.133	−48.840	362.137	626.573	236.549	545.364	−40.338	213.144	0.879	0.747	✓
A737	−21.856	−48.667	344.017	521.332	218.982	452.746	13.460	214.299	0.908	0.819	✗
A738	−21.780	−47.080	250.735	333.018	161.254	287.079	30.708	124.440	0.881	0.791	✗
A739	−22.415	−46.805	359.555	456.601	230.863	384.565	−20.558	190.691	0.884	0.836	✓
A740	−23.228	−45.417	361.491	426.767	228.705	353.503	−4.096	216.129	0.895	0.876	✗
A746	−24.963	−48.416	342.082	415.850	224.895	335.077	−13.427	115.725	0.884	0.736	✓
A747	−21.338	−48.114	346.865	442.103	216.820	369.305	18.765	196.750	0.886	0.850	✗
A748	−20.559	−48.545	353.802	640.973	227.071	545.170	31.125	227.999	0.888	0.722	✓
A753	−20.359	−47.775	347.917	549.392	216.777	478.413	12.509	221.882	0.889	0.798	✗
A755	−23.523	−46.869	544.999	392.011	319.757	328.477	−31.247	167.185	0.742	0.830	✓
A759	−21.751	−52.471	332.066	818.085	200.674	708.461	0.956	158.652	0.902	0.545	✓
A849	−22.634	−52.859	314.041	784.977	191.278	688.031	−9.888	165.516	0.915	0.588	✓
A850	−22.492	−52.134	339.294	809.662	211.607	699.164	−24.469	197.124	0.901	0.611	✗

Table 4
Comparision of performance metrics for P1 estimators with respect to the site-specific models and the CPRG empirical model. Prediction sites marked as (✓) correspond to those which performance was improved by the introduction of the imputation technique described by Section 5.1 whereas those marked as (✗) correspond to sites which models had better performance when using the original data.

7.2. Generalization Model

The estimator obtained after the processes of P2 corresponded to an ensemble composed by the models NN and XGB – using as *meta-learner* a Ridge Regressor – capable of estimating R recieving as input the predictions of the selected P1 estimators with respect to the Haversine Distance between the prediction point and P1 sites. Aiming to supplementary assess the effectiveness of the generalization model the predictions generated were compared to the ones generated by the CPRG empirical model and the IDW interpolation variant. The performance metrics for the predictions of these methods in each prediction site are presented by Table 6. Results show that the ensemble-based generalization model obtained in P2 was capable to provide generalization predictions with site-specific-like performance outperforming both reference methods in virtually all prediction sites of P2; Table 7 shows a summarized comparison with regard to all P2 prediction sites and Figure 6 highlights the performance improvements obtained when comparing the IDW-variant interpolation to the ensemble estimator. Such performance improvements obtained show, however, that no bias was introduced by the utilization of the ensemble estimator with respect to geographical localization of the prediction sites as depicted by Figure 5 which shows the comparison in RMSE between the ensemble estimator and the IDW-variant interpolation.

Metric		Method		$\Delta\%$ EMP-ML
		EMP	ML	
RMSE	$\bar{x}$	551.68	363.05	-34.19
	σ	136.97	43.09	-68.53
MAE	$\bar{x}$	462.23	232.34	-49.73
	σ	126.98	27.42	-78.40
MBE	$\bar{x}$	184.83	2.25	-98.78
	σ	36.26	21.31	-41.23
R^2 Score	$\bar{x}$	0.7284	0.8806	20.89
	σ	0.1370	0.0287	-79.04

Table 5
Summarization of the performance observed during P1 evaluation; $\bar{x}$ denotes the mean value and σ denotes the standard deviation.

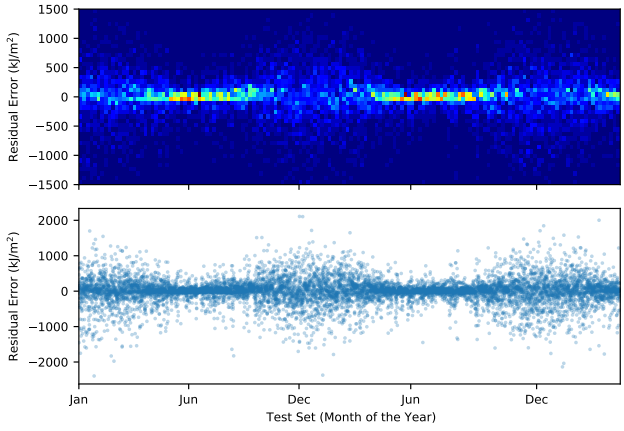

Figure 3: Example of bias in estimators obtained in P1 exemplified by Station ID A711 with respect to the period of the year.

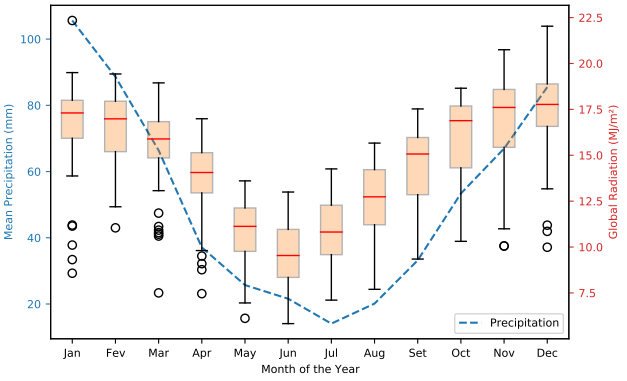

Figure 4: Relation between the variation of Global Solar Radiation – represented by the mean daily values in $\mathrm{MJ\,m^{-2}\,d^{-1}}$ – with respect to the mean monthly precipitation between 7:00 and 17:00 in mm. Both features were compared with respect to the period 2001–2020 using all obtained datasets.

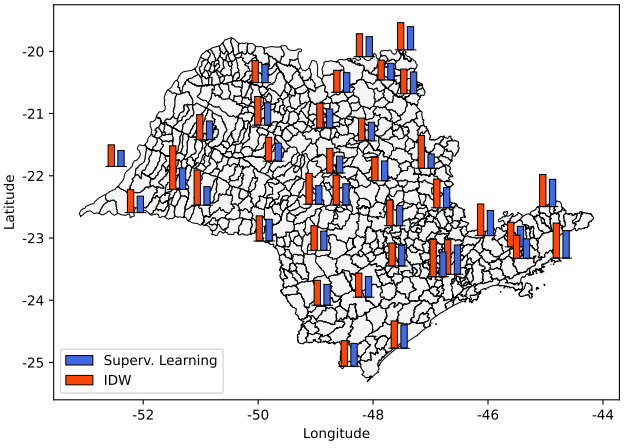

Figure 5: Comparision between RMSE for each prediction sites of P2 for IDW-variant interpolation and ensamble generalization model.

8. Conclusions

This paper introduced a multi-ensemble ML approach in order to forecast solar radiation from meteorological variables 60 minutes in the future. During its development various processes were carried out as parts of a two-procedure prediction system featuring ensemble estimators for both site-specific prediction models and the generalization ensemble estimator. The main conclusions regarding the processes implemented and the results observed are presented below.

1. Site-specific models were able to outperform the CPRG empirical models in the vast majority of prediction sites. The difference in the performance metrics for this two methods suggests that the greater complexity introduced by the ML estimators as opposed to the CPRG empirical model has proven to be an advantageoous trade-off given the perforamnce improvements observed.
2. The observation of the residual plots introduced by Section 7.1 suggests that the increased variation in meteorological conditions owing to seasonal characteristics in the prediction sites introduces noise in the datasets that can't be properly fitted by the estimators during training.
3. The introduction of the ensemble estimator in order to interpolate the predictions produced by the site-specific models showed that the ensemble estimator was able to outperform both CPRG empirical model and IDW-variant interpolation. The comparison of the values for σ and $\bar{x}$ observed in Table 7 suggest that this generalization model provided better average performance in terms of sparsity and consistency of the predictions while reducing drastically the bias error in contrast to the reference methods.
4. The imputation technique performed via IDW as part of the preprocessing conducted for P1 has provided a positive impact in the performance in 19 out of 38

Site ID	ML	IDW	EMP	ML	IDW	EMP	ML	IDW	EMP	ML	IDW	EMP
A725	365.033	454.356	484.249	253.150	341.290	396.253	−15.421	−154.201	396.253	0.883	0.820	0.736
A509	467.801	595.144	615.828	337.559	421.126	469.141	38.760	15.777	469.141	0.785	0.648	0.741
A520	385.402	441.572	466.936	257.537	294.843	410.024	34.094	35.607	410.024	0.864	0.822	0.779
A525	436.143	499.853	467.702	307.029	371.712	421.816	2.663	−36.401	421.816	0.848	0.799	0.843
A529	512.418	608.339	578.425	365.269	424.488	344.054	−5.167	−35.203	344.054	0.800	0.717	0.681
A619	513.332	662.356	642.057	371.080	463.821	487.656	65.924	206.071	487.656	0.816	0.693	0.691
A701	537.283	628.311	556.146	404.320	482.959	391.386	16.996	−53.704	391.386	0.783	0.704	0.822
A705	355.284	594.991	358.215	244.992	397.899	280.787	104.886	316.379	280.787	0.889	0.690	0.783
A707	399.919	823.542	849.888	289.203	579.853	724.310	109.154	454.659	724.310	0.855	0.384	0.105
A708	414.888	468.706	543.185	281.784	321.085	485.250	19.479	33.425	485.250	0.841	0.796	0.748
A711	381.699	443.764	448.547	253.772	313.109	401.742	9.626	−64.874	401.742	0.863	0.815	0.786
A712	434.226	509.446	379.281	316.218	368.650	285.211	17.839	73.447	285.211	0.836	0.775	0.840
A713	391.162	435.545	535.243	273.720	305.256	457.385	36.142	−9.772	457.385	0.844	0.807	0.775
A714	394.021	466.355	457.955	277.394	335.597	380.039	19.542	5.854	380.039	0.879	0.831	0.792
A715	387.398	443.704	466.570	277.117	329.657	381.818	−1.142	−24.113	381.818	0.873	0.833	0.769
A716	412.823	478.320	518.743	287.402	330.757	432.533	22.598	−50.387	432.533	0.843	0.793	0.625
A718	353.661	656.755	641.741	240.215	472.431	574.036	−0.481	−302.232	574.036	0.883	0.594	0.653
A726	382.792	468.392	584.131	264.353	356.313	471.487	0.076	−160.153	471.487	0.877	0.816	0.774
A727	348.855	448.498	570.102	233.110	318.243	484.947	8.429	−70.688	484.947	0.906	0.845	0.763
A728	390.602	471.586	355.876	268.919	318.479	291.556	23.464	18.084	291.556	0.875	0.818	0.883
A729	343.228	425.006	477.771	236.736	286.302	413.774	15.750	101.309	413.774	0.892	0.835	0.773
A734	362.563	464.721	670.457	237.774	338.495	601.199	24.397	−109.798	601.199	0.882	0.806	0.663
A735	429.000	543.535	646.704	289.515	377.156	534.314	12.931	−79.406	534.314	0.870	0.790	0.684
A736	361.753	463.467	626.573	251.708	339.242	545.364	−5.026	−17.924	545.364	0.882	0.803	0.747
A737	322.433	446.310	521.332	220.343	346.535	452.746	−10.858	−211.276	452.746	0.919	0.844	0.819
A738	273.439	594.317	333.018	185.757	461.929	287.079	93.025	442.714	287.079	0.859	0.330	0.791
A739	385.435	527.816	456.601	266.828	398.731	384.565	20.124	−265.390	384.565	0.869	0.754	0.836
A740	367.809	433.265	426.767	260.974	311.019	353.503	9.628	−82.672	353.503	0.895	0.855	0.876
A746	434.422	493.175	415.850	302.756	332.639	335.077	40.337	82.712	335.077	0.818	0.764	0.736
A747	356.837	434.566	442.103	239.779	297.200	369.305	35.220	−3.910	369.305	0.880	0.822	0.850
A748	378.292	418.052	640.973	254.245	290.889	545.170	21.070	−24.617	545.170	0.875	0.846	0.722
A753	323.462	383.848	549.392	210.056	254.409	478.413	25.463	3.976	478.413	0.904	0.866	0.798
A755	482.550	720.416	392.011	316.185	483.931	328.477	56.894	302.160	328.477	0.788	0.523	0.830
A759	320.011	400.219	818.085	194.056	294.226	708.461	27.457	−152.212	708.461	0.904	0.848	0.545
A850	325.650	422.567	809.662	203.331	313.228	699.164	10.097	−164.993	699.164	0.909	0.848	0.611

Table 6
Performance metrics for P2 models with respect to generalization predictions obtained from the ensemble estimator, the IDW-variant interpolation and NAME empirical model.

Metric		Method			Δ% EMP-ML	Δ% IDW-ML
		EMP	IDW	ML		
RMSE	$\bar{x}$	535.66	507.73	392.33	26.75	22.73
	σ	128.78	99.86	58.72	51.38	41.19
MAE	$\bar{x}$	445.94	362.10	270.69	39.30	25.24
	σ	116.51	72.15	28.69	74.51	58.85
MBE	$\bar{x}$	183.81	0.52	25.25	86.97	-4755.76
	σ	36.89	172.55	29.76	19.32	82.75
R^2 Score	$\bar{x}$	0.7392	0.7581	0.8625	-16.68	-13.77
	σ	0.1344	0.1262	0.0361	73.14	71.39

Table 7
Summarization of the performance observed during P2 evaluation with respect to the different methods used; $\bar{x}$ denotes the mean value and σ denotes the standard deviation.

site-specific models. This result shows that such technique can be effectively applied to meteorological data in order to artificially reconstruct missing values in the datasets increasing the amount of data available for training routines.

9. Acknowledgements

The processes developed during the work hereby presented have extensively used implementations provided by Pandas [] for data preprocessing; Scikit-Learn [30] for estimator training, hyperparemeter optimization and evaluation; Matplotlib [18] and GeoPandas [20] for images regarding data visualization and geographical locations respectively; and Optuna Framework [1] for further hyperparameter optimization.

References

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[2] Ahmad Alzahrani, Pourya Shamsi, Cihan Dagli, and Mehdi Ferdowsi. Solar irradiance forecasting using deep neural networks. *Procedia Computer Science*, 114:304–313, 2017.

[3] Özge Ayvazoğluyüksel and Ümmühan Başaran Filik. Estimation methods of global solar radiation, cell temperature and solar power forecasting: A review and case study in eskişehir. *Renewable and Sustainable Energy Reviews*, 91:639–653, 2018.

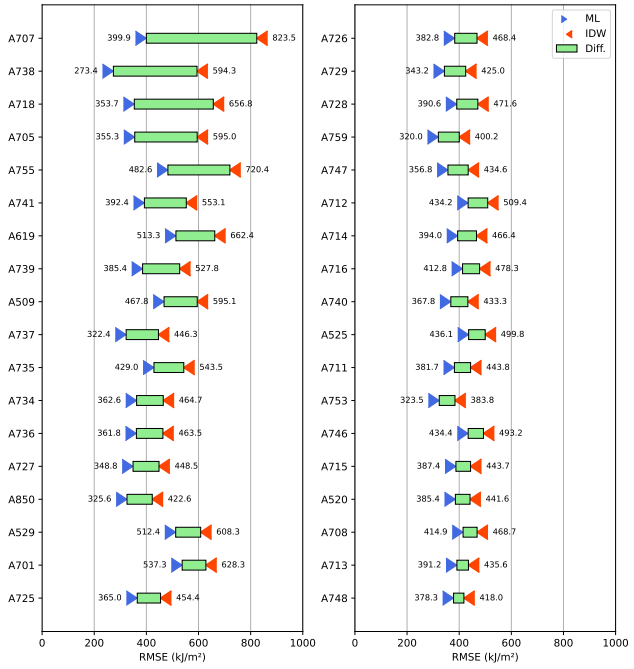

Figure 6: Performance improvement comparison for each prediction site during P2 considering IDW-variant interpolation and P2 ensemble estimator. Stations presented are sorted by the difference between the RMSE for both methods.

[4] Leo Breiman. Random forests. *Machine learning*, 45(1), 2001.

[5] Davide Cannizzaro, Alessandro Aliberti, Lorenzo Bottaccioli, Enrico Macii, Andrea Acquaviva, and Edoardo Patti. Solar radiation forecasting based on convolutional neural network and ensemble learning. *Expert Systems with Applications*, 181:115167, 2021.

[6] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[7] Manuel Collares-Pereira and Ari Rabl. The average distribution of solar radiation-correlations between diffuse and hemispherical and between daily and hourly insolation values. *Solar energy*, 22(2):155–164, 1979.

[8] L Cornejo-Bueno, C Casanova-Mateo, J Sanz-Justo, and S Salcedo-Sanz. Machine learning regressors for solar radiation estimation from satellite data. *Solar Energy*, 183:768–775, 2019.

[9] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[10] Scikit Learn Developers. Compare the effect of different scalers on data with outliers, 2020.

[11] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

[12] I Dincer and TAH Ratlamwala. Development of novel renewable energy based hydrogen production systems: a comparative study. *Energy Conversion and Management*, 72:77–87, 2013.

[13] Yu Feng, Weiping Hao, Haoru Li, Ningbo Cui, Daozhi Gong, and Lili Gao. Machine learning models to quantify and map daily global solar radiation and photovoltaic power. *Renewable and Sustainable Energy Reviews*, 118:109393, 2020.

[14] Christian Gueymard. Mean daily averages of beam radiation received by tilted surfaces as affected by the atmosphere. *Solar Energy*, 37(4):261–267, 1986.

[15] Mehreen Gul, Yash Kotak, and Tariq Muneer. Review on recent trend of solar photovoltaic technology. *Energy Exploration & Exploitation*, 34(4):485–526, 2016.

[16] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Boosting and additive trees. In *The elements of statistical learning*, pages 337–387. Springer, 2009.

[17] Simon Haykin. *Neural Networks: a Comprehensive Foundation*. Prentice Hall PTR, 1994.

[18] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[19] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[20] Kelsey Jordahl, Joris Van den Bossche, Martin Fleischmann, Jacob Wasserman, James McBride, Jeffrey Gerard, Jeff Tratner, Matthew Perry, Adrian Garcia Badaracco, Carson Farmer, Geir Arne Hjelle, Alan D. Snow, Micah Cochran, Sean Gillies, Lucas Culbertson, Matt Bartos, Nick Eubank, maxalbert, Aleksey Bilogur, Sergio Rey, Christopher Ren, Dani Arribas-Bel, Leah Wasser, Levi John Wolf, Martin Journois, Joshua Wilson, Adam Greenhall, Chris Holdgraf, Filipe, and François Leblanc. geopandas/geopandas: v0.8.1, July 2020.

[21] Eleni Kaplani, Socrates Kaplanis, and Sourav Mondal. A spatiotemporal universal model for the prediction of the global solar radiation based on fourier series and the site altitude. *Renewable Energy*, 126:933–942, 2018.

[22] A Khosravi, RNN Koury, L Machado, and JJG Pabon. Prediction of hourly solar radiation in abu musa island using machine learning algorithms. *Journal of Cleaner Production*, 176:63–75, 2018.

[23] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[24] Jiaming Li, John K Ward, Jingnan Tong, Lyle Collins, and Glenn Platt. Machine learning for solar irradiance forecasting of photovoltaic system. *Renewable energy*, 90:542–553, 2016.

[25] Kasra Mohammadi, Shahaboddin Shamshirband, Dalibor Petković, and Hossein Khorasanizadeh. Determining the most important variables for diffuse solar radiation prediction using adaptive neuro-fuzzy methodology; case study: City of kerman, iran. *Renewable and Sustainable Energy Reviews*, 53:1570–1579, 2016.

[26] Seyed Abbas Mousavi Maleki, H Hizam, and Chandima Gomes. Estimation of hourly, daily and monthly global solar radiation on inclined surfaces: Models re-visited. *Energies*, 10(1):134, 2017.

[27] INMET National Institute of Meteorology of Brazil. Main website, 2021.

[28] INMET National Institute of Meteorology of Brazil. Meteorological database, 2021.

[29] AK Pandey, VV Tyagi, A Jeyraj, L Selvaraj, NA Rahim, and SK Tyagi. Recent advances in solar photovoltaic systems for emerging trends and advanced applications. *Renewable and Sustainable Energy Reviews*, 53:859–884, 2016.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. T hirion, O. Grisel, M. Blondel, P. Prettenhofer, R. W eiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[31] Alvin C. Rencher and William F. Christensen. *Methods of Multivariate Analyysis, Third Edition*. Wiley, 2012.

[32] Ryan M Rifkin and Ross A Lippert. Notes on regularized least squares. 2007.

[33] Shashi Shekhar and Hui Xiong. *Inverse Distance Weighting*, pages 600–600. Springer US, Boston, MA, 2008.

[34] Bent Sorensen and Giuseppe Spazzafumo. *Hydrogen and fuel cells: emerging technologies and applications*. Academic Press, 2018.

[35] JW Spencer. Fourier series reprensentation of the position of the sun. *Search*, 2(5):172, 1971.

[36] Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

[37] Wanxiang Yao, Zhengrong Li, Tongbin Xiu, Yuan Lu, and Xiaobin

Li. New decomposition models to estimate hourly global solar radiation from the daily value. *Solar Energy*, 120:87–99, 2015.

[38] Jianwu Zeng and Wei Qiao. Short-term solar power prediction using a support vector machine. *Renewable Energy*, 52:118–127, 2013.