Metaheuristic based MLP-SARIMAX Hybridization One Hour Ahead Solar Radiation Forecasting

1st Hugo Abreu Mendes
University of Pernambuco (*POLI*)
Recife, Brazil
ORCID: 0000-0002-0474-2680

2nd João Fausto Lorenzato de Oliveira
University of Pernambuco (*POLI*)
Recife, Brazil
ORCID: 0000-0002-1150-4904

3th Paulo Salgado Gomes de Mattos Neto
Federal University of Pernambuco (*CIN*)
Recife, Brazil
ORCID: 0000-0002-2396-7973

4th Alex Coutinho Pereira
CHESF (DEGS)
Recife, Brazil
alexcp@chesf.gov.br

5th Eduardo Boudoux Jatoba
CHESF (DEGS)
Recife, Brazil
ejatoba@chesf.gov.br

6th José Bione de Melo Filho
CHESF (DRPI)
Recife, Brazil
jbionef@chesf.gov.br

7th Elisabete Joana Paes Barreto
CHESF (DRPI)
Recife, Brazil
ejpaes@chesf.gov.br

8th Manoel Henrique da Nóbrega Marinho
University of Pernambuco (*POLI*)
Recife, Brazil
ORCID: 0000-0003-3129-0453

***Abstract*—Within the context of clean energy generation, solar radiation forecast is applied for photovoltaic plants to increase maintainability and reliability. Statistical models of time series like ARIMA and machine learning techniques help to improve the results. Hybrid Statistical + ML are found in all sorts of time series forecasting applications. This work presents a new way to automate the SARIMAX modeling, nesting PSO and ACO optimization algorithms, differently from R's AutoARIMA, its searches optimal seasonality parameter and combination of the exogenous variables available. This work presents 2 distinct hybrid models that have MLPs as their main elements, optimizing the architecture with Genetic Algorithm. A methodology was used to obtain the results, which were compared to LSTM, CLSTM, MMFF and NARNN-ARMAX topologies found in recent works. The obtained results for the presented models is promising for use in automatic radiation forecasting systems since it outperformed the compared models on at least two metrics.**

***Index Terms*—Solar radiation, time series, machine learning, optimization**

I. Introduction

In the last decade, the growth of photovoltaic (PV) generation has been expressive. According to the IEA (Internetional Energy Agency), global production increased from 32TWh in 2010 to more than 720TWh in 2020 [1]. In Brazil, generation still has tax incentives [2], and a great potential [3] still underutilized. Photovoltaic generation will correspond to only 3,9% of Brazilian national production by 2025 [4]. However, there are estimations predicting that this type of generation will be responsible for more than 36 % of the energy generated in this country by 2050 [5].

Estimation of photovoltaic generation potential is a topic that remains popular [6]–[9]. Some works carried out took into account the technology used by the photovoltaic cell and satellite models that aim to define the physical input parameters, such as radiation, temperature and wind speed [10]–[12].

Machine learning applied to PV generation forecasting has focused on the temporal prediction of electrical generation [9], [13]. This type of prediction is very useful regarding to the complete electrical system of a region or country, for balancing supply and demand, increasing the programmability for smart grids [14]. In general, for plant managers, less deviation between the scheduled and the produced power means avoiding penalties [9].

The photovoltaic energy forecast can be classified according to the forecast horizon [9]. Forecast horizon and applications can be summarized as follow:

- Very short-term (Minutes) - control and management of photovoltaic systems, micro-networks and electricity market.
- Short-term (72h) - control of the power system operations, economic dispatch and commitment of the unit.
- Medium-term (1 week) - maintenance.
- Long-Term (years) - Power plants planning.

The process of automating *machine learning* models can be called AutoML [15], in which *pipelines* are generated, being a series of transformations in the end-to-end data, that is, inputting raw observed data and setting the target for the pipeline, the AutoML is responsible for searching the best model, and the pipeline output is the prediction of the trained model [15]–[17]. AutoMLs in photovoltaic radiation prediction systems can help to increase adaptability. These systems are also capable of making the forecasting process more available for non specialists, since low-level design adjustments and modifications are no longer be necessary, as a

robust and automated system with a higher level of complexity would adapt to variations [18], [19].

In recent years, it has been common to implement statistical time series models combined with machine learning for general time series forecasting, combining ARIMA with RNA [20], SVR (Suport Vector Regression) [21], LSTM (Long Short-Therm Memory) [22], also several regression models together in a dynamic residual forecasting (DReF) selection [23]. To obtain better performance, optimization and intelligent algorithms are applied to Hybrid models [21], [23].

NARNN-ARMAX [24] is applied as a two stage day-ahead forecast, where the NARNN (Nonlinear Auto Regressive Neural Network) model is applied as a exogenous variable to the ARMAX model. MMFF (Multi-Model-Forecasting-Framework) is applied for Hourly-Similarity with solar data, blending SVR, GBR (Gradient Boosting Regressor), ANN (Artificial Neural Network) and RFR (Random Forest Regressor) [25]. LSTM neural networks, not combined with ARIMA models, is also found to be used for short-term solar radiation forecasting [26], as well as LSTM plus CNN (Convolutional Neural Network) [27], called CLSTM.

This work focuses on the use of machine learning to predict short-term solar radiation time series (hour) using hybrid models, and uses optimization algorithms for the creation of an AutoML pipeline. Meta-heuristics optimization algorithms in the hybrid model optimization process are also found in literature. Applications using the models that are covered in this work: Genetic Algorithm (GA) [28]–[30], Particle Swarm Optimization (PSO) [31] and Ant Colony Optimization (ACO) [32].

The use of meta-heuristic algorithms for the implementation of automated machine learning (AutoML) systems can be seen as the optimization of hyper-parameters and architectural optimization [19]. The proposed pipeline consists of SARIMAX parameters and exogenous variable optimization using bilevel (nested) [33] optimization of the PSO (Particle Swarm Optimization) on the super level and ACO (Ant Colony Optmization) on the lower level hybridized with architectural and parametric optimization of MLPs (Multi Layer Perceptron) through a genetic algorithm (GA).

The sections of this work are organized as follows. Materials and methods II the database used II-A, AutoARIMA and SARIMAX PSO-ACO Search algorithm in subsection II-B, hybrid models are also presented for residue correction in the section II-C, finally the results are presented in the section III and the conclusions IV.

II. Materials And Methods

A. Data

The Brazilian national meteorological institute (INMET) provides data from several stations around the country through its Meteorological Database (BDMEP) [34]. The data is acquired from measurement stations, with automatic and conventional ones, the latter are being discontinued [35]. INMET data can be used to obtain spatial interpolation and temporal prediction models [30].

The data obtained contains information with hourly frequency, as the objective is short-term forecasting, the series obtained has a maximum 30 days of data, to train and evaluate the models. The chosen cities are: Maceió-AL, Florianópolis-SC and Bom Jesus da Lapa-BA.

Among the present data, there are hourly time series of the following variables, where the available exogenous variables are highlighted in bold.

1) Global Radiation (W/m^2)
2) **Total Precipitation** (mm)
3) **Air temperature** (°C)
4) **Relative humidity**, maximum and minimum (%)
5) **Wind speed** (m/s)
6) **Gust speed** (m/s)

The Global Radiation is the target variable to be predicted. The importance of each variable is explained by the correlation between wind speed and solar radiation [36], [37]. Relative humidity and air temperature are also found to be correlated with solar radiation using INMET data [38]. Precipitation and gust speed used as a representation for cloudiness of possible storm can impact solar radiation intensity [39]. The use of exogenous variables in SARIMAX model is optimally chosen in the first part of the pipeline using PSO-ACO nested algorithms, this methodology is described in the next section II-B.

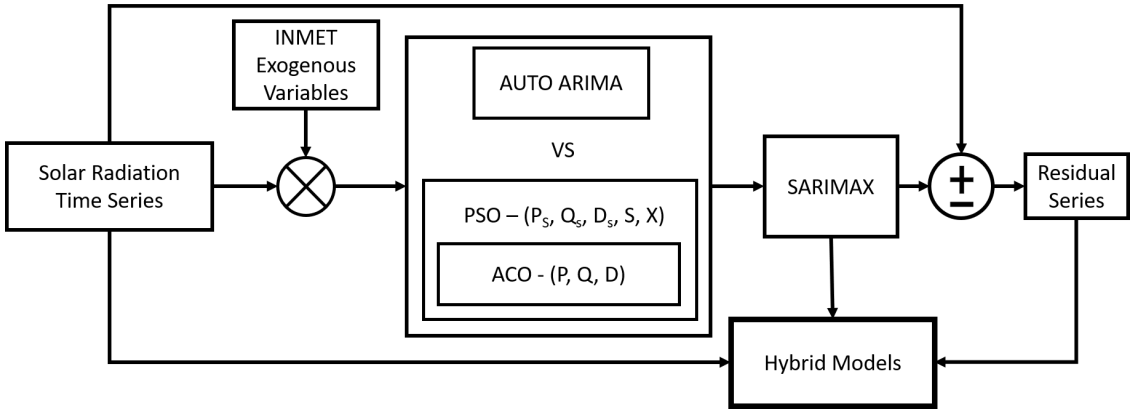
Data obtained from external sources is not always perfect, one of the common problems is the presence of empty data or inadequate format. In the case of the data processed by INMET, it was necessary to adapt the format and complement the missing data. The data were scaled after dividing the series of each variable by the maximum value found for it. At the end of the forecast, it is possible to return to the original scale. The missing data is complemented by the *forward fill* method, in which last valid value observed is used to replace empty values for each variable.

B. AutoArima versus SARIMAX PSO-ACO Search

The Box-Jenkins methodology can be used to determine the parameters for an ARIMA model, for this purpose the auto-correlation function and partial auto-correlation function are used to identify the degree of self-dependence of the time series, to identify if it is not stationary and also to see if there is a implicit seasonality. In cases of non-stationarity, a numerical differentiation is made [40]–[42]. Using statistical stationarity tests such as KPSS (Kwiatkowski-Phillips-Schmidt-Shin) [43] and ADF (*Augmented* Dickey-Fuller) [44] it is possible to automate the number of differentiation to get a stationary time series. For series with seasonal profile, a possible test is the Canova-Hansen [45]. Adding the information provided by ACF, PACF and the AIC metric, it is then possible to create algorithms for automatic modeling of time series.

To automatically obtain a SARIMAX model that represents the time series, an algorithm is needed. One of the most popular algorithms, implemented and popular in both R and python programming languages [46], that automates obtaining

Fig. 1. Flowchart for defining the SARIMAX model that is used in hybridization. Original series, SARIMAX and Residue are used for training the hybrid model.



an ARIMA model is known for the work of Hyndman-Khandakar entitled **AutoARIMA** [42], [47] uses a stepwise search methodology to find the parameters (p, d, q) e (P, D, Q, S) for a SARIMA model. In this case the seasonality, indicated by parameter S is provided by the user [46], [47]. It is also possible to add exogenous variables to the **AutoArima**, thus leading to a SARIMAX model.

The **AutoArima** does not contemplate some possibilities that could lead to better models. For SARIMAX models fitted with **AutoArima** the seasonal parameter is a parameter to chosen by the user, so an improvement is to carry out tests with different possibilities for seasonal frequencies. Another advance is to make an automated choice of the best exogenous variables among those inserted as input.

To implement the proposed improvements, one can initially think of using a grid search algorithm, from all the possibilities of iterating the parameters (p, d, q, P, D, Q, S, X), X represents the exogenous variables, therefore, all parameters can be described as positive integer variables. In this scenario, meta-heuristics ACO and PSO can be applied to solve for the SARIMAX parameters and exogenous more relevant variables. Both algorithms have agents that interact with the environment or update their solutions with each other. In this project, the algorithms are stacked and this is described below.

The PSO was formulated for use with continuous variables, however naively, it is possible to discretize the positions of the particles to solve a discrete problem, PSO discretization may involve the velocity of the particles [48], but in this work, velocity was kept as a continue variable. In the case of PSO, this algorithm was in charge of optimizing the parameters (P, D, Q, S, X), having its search space determined by the possible values that this set can receive [49]. The PSO solves for a *fitness* function that is an instance of the ACO algorithm.

The ACO receives the solution, as the (P, D, Q, S, X), found by the PSO and using the AICc (corrected Akaike Information Criteria) [50] metric between each candidate solution and the original series, the ACO algorithm searches for the (p, d, q) parameters. For example, if the solution found by PSO at one iteration is (1,1,1,1,3), the ACO uses it and search for solution of (p, d, q), one possible solution for this iteration is (1,0,1), so the resulting SARIMAX parameters are $(p = 1, d = 0, q = 1)(P = 1, D = 1, Q = 1, S = 1, X = 3)$ where $X = 3$ represents a set of possible exogenous variables. The X value is converted to base two and mapped to columns of the exogenous dataset, if $X = 3$ in base two as $X = b11$ then columns zero and one are used as exogenous variables.

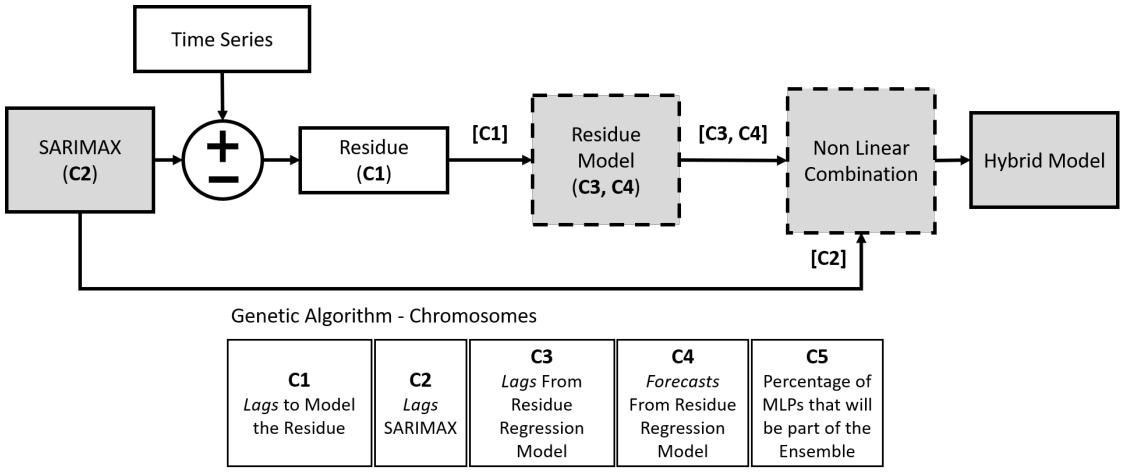
The Figure 1 shows the pipeline, with the optimization flowchart of the SARIMAX model and how it will be used later for what will be described in the next section II-C.

C. Hybrid Models for Residue Correction

In this section is discussed the hybrid models applied to the residue correction, given by the error between the original time series and the SARIMAX linear model, whose parameters and exogenous variables were defined as explained in the previous section, as being the first stage. The correction is based on the flowchart of Figure 2, as can be seen, there are four chromosomes that are used in the search for a genetic algorithm. Variables of the type *Lags* are the quantities of samples from the past used to predict the future. The role of the chromosomes is explained below:

- C1: *Lags* for the Residue Regressing - with the residue data, these samples are used in the Residue Model, which will rather be a MLP model or an ensemble of MLPs, this parameter is further described at sessions II-C1 and II-C2, respectively. C1 is the amount of data from the past of the residue series that are used for forecasting.
- C2: *Lags* SARIMAX - With the SARIMAX Model, it is possible to generate lagged samples estimation from the solar radiation time series.
- C3: *Lags* from the Residue Model - with a trained and optimized ML Model it is possible to generate the time series corresponding to the Modeled Residue and to generate lagged samples. Notice that to generate all forecasts samples from the Residue Model, this very same model must be fed back.
- C4: *Forecasts* from Residue Model - Analogous to chromosome C3, but with samples ahead in time, generated by one of the models of section II-C1 or II-C2.

Fig. 2. Schematic flowchart of the proposed residue correction models. The chromosomes are C1-C4. Two regression models are generated, the first for the residue series, while the second to combine the modeled residue with the SARIMAX forecast. The dashed models can both be either an GA optimized MLP or a GA optimized ensemble of MLPs.



- C5: Percentage of MLPs that will be part of the *Ensemble*. Only used in case of the model of Section II-C2.

Chromosomes C1-C4 are initialized as integers from a uniform distribution of lower limit 1 and upper 20. A probability of mutation and crossover of 80%. To apply the final trained model one may just use the C1 to C4 final values to adjust the shape of the input data to the trained models.

Algorithm 1: Crossover for presented hybrid algorithms

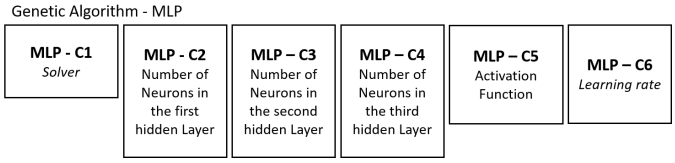
```
Input: Population, Crossover Prob.
Output: New population
begin
    Keep best individual in population;
    Sort the population based on the fitness;
    foreach individual, I do
        p = random(0,1);
        if p > Crossover Prob. then
            C = set of random cromossomes;
            population[I][C] = population[I/2][C]
        end
    end
end
```

The crossover operator works according to the Algorithm 1. Individuals are placed in order from best to worst, based on a MAE (Mean Absolute Error) metric of test data portion, so the best individual is always taken to the next population. Then, for each of the remaining individuals, a set of the four chromosomes that is received from an individual in a better position, which is given by the individual's rank symmetrically. For example, if there is 10 individuals, the second worst will receive information from the second best. The chromosome mutation occurs from the variation of the addition of an integer obtained by a uniform distribution between -2 and 2.

The second stage of the process is given by the Residue Model and the third, by the Combination Model. In both of these steps, two MLP hyper-parameterization algorithms can be chosen, based on a genetic algorithm, described in the next subsections II-C1 e II-C2.

1) AG-MLP: To optimize MLP, regarding the architecture and parameters, a genetic algorithm is used. Chromosomes are used to characterize the parameters and architecture of an MLP.The chromosomes used are shown in Figure 3.

Fig. 3. Chromosomes of the MLP Genetic Algorithm.



The chromosomes, have the following characteristics:

- MLP-C1: *Solver* - This is the optimization algorithm used to obtain the weights in neurons. Can be chosen from LBFGS, ADAM and SGD.
- MLP-C2 to MLP-C4: Number of neurons in the hidden layers 1, 2 e 3.
- MLP-C5: Activation Function - Possibility among the Identity, Logistics, Hyperbolic Tangent and Relu functions.
- MLP-C6: *Learning Rate* - This is how the MLP η Learning Rate is updated. It can be chosen from a Constant, *Invscaling* and Adaptive.

To obtain an optimized solution, witch is an instance of a trained MLP the GA algorithm firstly generates a initial population of MLPs, with random parameters and architecture, then this population is evaluated and ranked. The evaluation

is done by training each MLP and using a MAE metric on the test data. To update the population, a crossover is performed between better and worse MLPs. Because the training process of the MLPs can lead different results, the best MLP instance (parameters, architecture and weights) is always kept to the next generation. After the crossover, the numerical chromosomes are mutated. The resultant new population is re-evaluated and the cycle continues until number of epochs defined is reached. Finally the all time best MLP is returned as the optimal solution.

About MLP-C1 *Solver*, LBFGS is an acronym for *Limited-Memory Broyden-Fletcher-Goldfarb-Shanno Algorithm* which tends to lead to faster training of MLPs than the SGD algorithm, acronym for *Stochastic Gradient Descendent* [51]. In the case of using the SGD, the moment is set to 0.9. ADAM, whose name derives from *Adaptative Moment Estimation*, it is a newer algorithm than the previous ones and popular in applications involving *Deep Learning* [52].

For MLP-C6 *Learning Rate*, each of the possibilities is characterized as follows: Constant is a learning rate $\eta = 0,001$. causes the learning rate to gradually decrease, with period t using an inverse scale exponent $\eta = \eta/t^{0.5}$. Adaptive learning rate keeps learning rate constant while training loss continues to decrease. Each time two consecutive epochs fail to decrease training loss by at least $0,0001$, or fail to increase the validation score by at least $0,0001$, the current learning rate is divided by 5.

2) AG-MLP Ensemble: This model is analogous to the previous one, II-C1, in the use of MLPs as main elements, however with the difference that instead of choosing an ANNs only for the Residue Model and Combination Model, the best ANNs are chosen in order to make an average in the prediction among the best. For the correct implementation of this model, a chromosome plus C5 is added to those that have been explained and illustrated in Figure 2.

The C5 chromosome has a crossover equivalent to that demonstrated by the Algorithm 1, however its mutation occurs from the variation of the addition of an integer obtained by a uniform distribution between -10 and 10.

Based on the presented methodology and algorithms, in the next sections results were obtained for three Brazilian cities.

III. Results

To obtain the results, the following methodology for the use of the models proposed is established:

1) **AutoArima** is executed to obtain the SARIMAX model, with seasonal frequency 24 and all exogenous variables highlighted in the section II-A.
2) The PSO-ACO SARIMAX described in II-B algorithm is executed.
   a) Increase the maximum dimensionality limit for the parameters (p, d, q) and (P, D, Q) given by the **AutoArima**. Seasonality is searched in multiples of 24 for greater exploitation, and because of the natural pattern of solar radiation series.
   b) Exogenous variables are added as search variables.
   c) If the result is worse than that obtained by the **AutoArima**, the latter is maintained. The comparison is made using the AICc metric.
3) The winning SARIMAX model continues for use in the hybrid models described in subsections II-C1 e II-C2.
   a) 80% of data for training, 20% for test.

Maceió-AL is the city chosen for the first preliminary result of the defined methodology. The data set for this city is 30 days, between 03/12/2020 at 9 pm and 11/04/2020 at 8 pm from the base to the city of Maceió-AL. Having an economically important metropolitan region, with ecotourism attractions, Florianópolis-SC, it was chosen to integrate the results. This is the second time series that will be presented in the results of this work. The last 15 days of the base obtained are used, between 12/17/2019 at 0 am and 12/31/2019 at 11 pm. The third time series chosen is 15 days, between 7/17/2020 at 0 am and 7/31/2020 at 11 pm from the base to the city of Bom Jesus da Lapa-BA. This location was chosen because it contains nearby photovoltaic plants and is one of the points of greatest incidence of photovoltaic radiation according to the Brazilian Solarimetric Atlas [3].

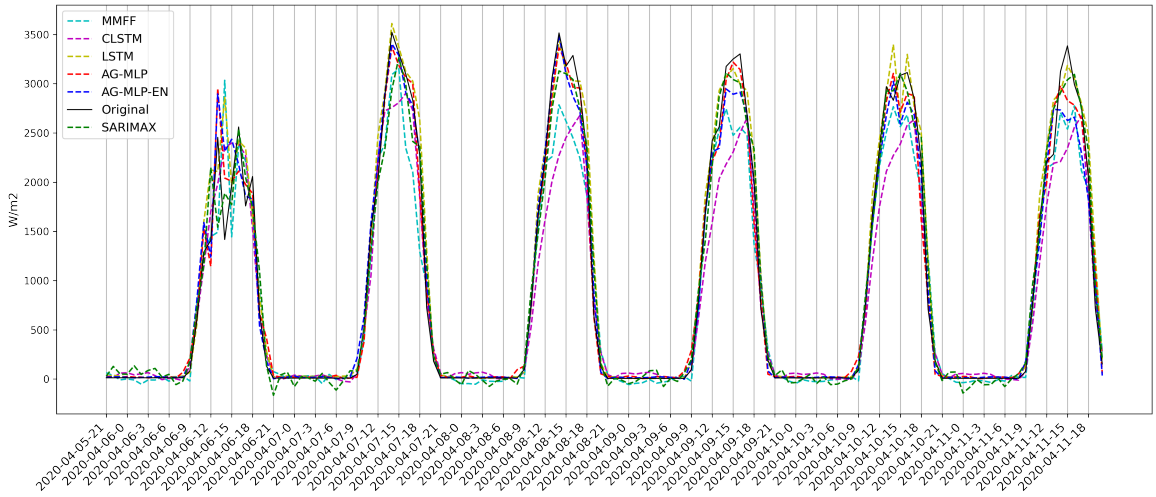
Although the number of cities/series is small, the reason difference of sample sizes length is to show that good results can be achieved for both. The Table I shows the result of the comparison between the algorithms described in the section II-B. For every model, some level of optimization was used to achieve the better results. Although some of the models may lead to variations in results, for the same random generator number seed, all algorithm would let to identical results, reason why statistical tests are not considered.

The results of the Hybrids models of the sections are shown in the Table II, A LSTM model was also tested for comparison, as it is a popular method for predicting short-term solar radiation. The topology consists of two LSTM Layers and two dense MLP layers, number of units, activation function and optimizer were defined by a simple grid search for each layer. A CLSTM model following the similar architecture as in [27] was also implemented.

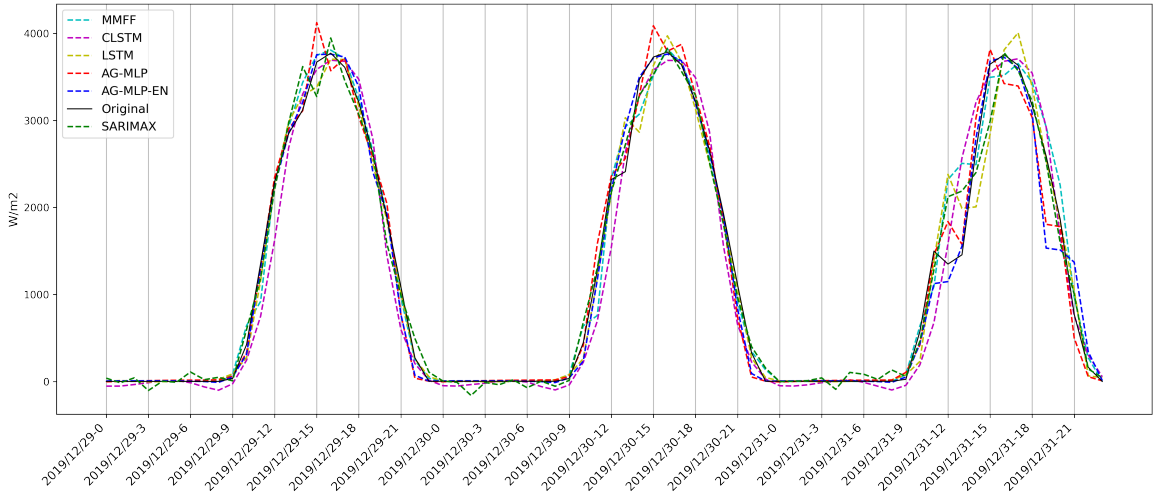
For the Maceió-AL data, hybrid models with GA results were achieved with 3 generations and population size of 12. Meanwhile, Florianópolis-SC and Bom Jesus da Lapa-BA, used both 4 generations and 15 individuals. Test data that can be seen in Figure4. It is noticed that there is correction of the time series with emphasis on the moments when the solar radiation must be zero due to the night. This correction can be numerically perceived when analyzing the MAPE metric in Table II, this result is replicated with the next time series.

IV. Conclusions

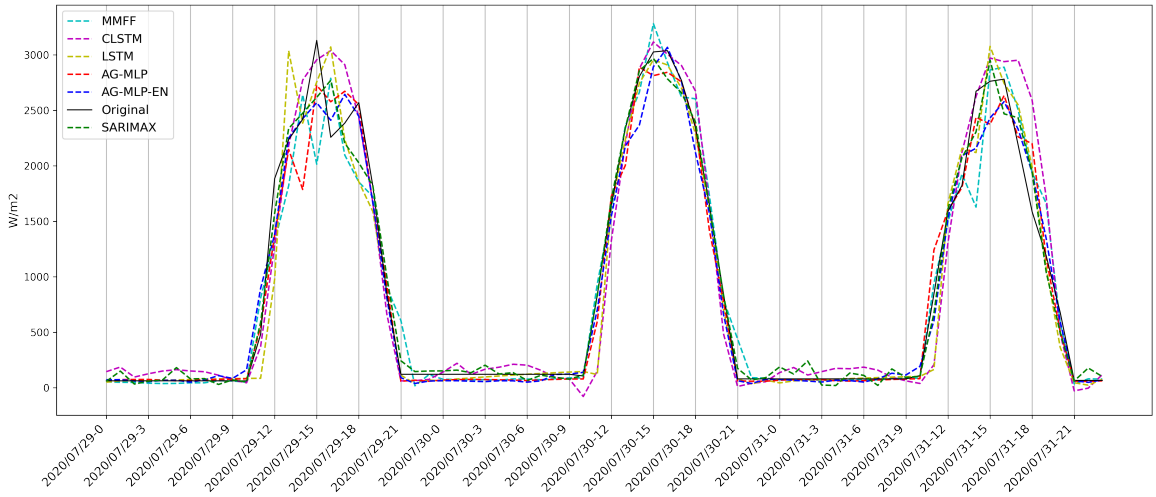
In this work, the pipeline applies in different stages three optimization algorithms PSO, ACO and GA. First, it is proposed to use two popular algorithms PSO and ACO to optimize the parameterization of SARIMAX linear models, also searching for seasonality and available exogenous variables. Right after, two ways of generating hybrid models with MLPs are shown,



(a) Maceió-AL



(b) Florianópolis-SC



(c) Bom Jesus da Lapa - BA

Fig. 4. Results data plot for models MMFF, LSTM, CLSTM, AG-MLP, AG-MLP-Ensemble, SARIMAX and Original radiation time series

TABLE I
COMPARISON OF RESULTS BETWEEN **AUTOARIMA** AND PSO-ACO FOR BRAZILIAN CITIES.

		SARIMAX			
City	Algorithm	$(p,d,q)(P,D,Q,S)$	Exogenous Variables	AICc	MAPE
Maceió-AL	AutoArima	(1,0,2)(1,0,2,24)	All	-1488.858	**5.3497**
Maceió-AL	PSO-ACO	(2,0,0)(1,0,1,24)	Air Temperature	**-1516.263**	6.7475
Florianópolis-SC	AutoArima	(0,1,1)(2,0,2,24)	All	-695.210	57.106
Florianópolis-SC	PSO-ACO	(2,0,2)(1,0,1,24)	Air Temperature Precipitation	**-754.475**	**52.316**
Bom Jesus da Lapa - BA	AutoArima	(3,0,2)(2,0,2,24)	All	-825.1507	0.9518
Bom Jesus da Lapa - BA	PSO-ACO	(1,0,0)(2,0,2,24)	Air Temperature Humidity Wind Speed	**-831.802**	**0.9233**

TABLE II
COMPARISON BETWEEN PROPOSED HYBRID MODELS, SARIMAX, NARNN-ARMAX [24], MMFF [25], LSTM [26], [53] AND CLSTM [27]. ALL METRICS ARE EVALUATED IN THE SCALE DESCRIBED IN SECTION II-A. PLEASE READ MC AS MACEIÓ, FR AS FLORIANÓPOLIS AND BJL AS BOM JESUS DA LAPA.

City	Model	MAE	MSE	MAPE
MC - AL	SARIMAX	0.0352	0.0029	2.9130
MC - AL	NARNN-ARMAX	0.0498	0.0059	2.6581
MC - AL	MMFF	0.0503	0.0079	1.7874
MC - AL	LSTM	0.0313	0.0038	0.6249
MC - AL	CLSTM	0.0546	0.0089	0.7969
MC - AL	Hybrid AG-MLP	**0.0243**	**0.0018**	0.9090
MC - AL	Hybrid AG-MLP-Ensemble	0.0268	0.0024	**0.5304**
FR - SC	SARIMAX	0.0337	0.0027	7.635
FR - SC	NARNN-ARMAX	0.0366	0.0029	8.8285
FR - SC	MMFF	0.0308	0.0036	**0.4140**
FR - SC	LSTM	0.0313	0.0036	1.7034
FR - SC	CLSTM	0.0463	0.0054	0.6717
FR - SC	Hybrid AG-MLP	0.0280	**0.0022**	1.4620
FR - SC	Hybrid AG-MLP-Ensemble	**0.0237**	0.0024	0.9386
BJL - BA	SARIMAX	0.0311	**0.0025**	0.3485
BJL - BA	NARNN-ARMAX	0.0371	0.0048	0.3008
BJL - BA	MMFF	0.0442	0.0068	0.3205
BJL - BA	LSTM	0.0427	0.0069	0.2116
BJL - BA	CLSTM	0.0526	0.0070	0.8094
BJL - BA	Hybrid AG-MLP	**0.0310**	0.0030	**0.1917**
BJL - BA	Hybrid AG-MLP-Ensemble	0.0322	0.0027	0.2519

the first using MLP to make residue correction from combination and modeling of the residue and the second using an ensemble of MLPs.

From the results obtained, it is observed that the use of PSO and ACO for parameterization of SARIMAX models is promising, with better results than **AutoARIMA**, since the proposed method search space goes beyond **AutoARIMA** and implements a nested optimization instead of the step wise algorithm of **AutoARIMA**. Highlighting the utility of the proposed PSO-ACO in optimally automating the choice of exogenous variables.

The strategy adopted to correct the residue by optimizing elements that model the error and also make a non-linear combination of the modeled residue with the SARIMAX prediction is also shown to be relevant, since for all the series chosen, it had a better performance according to at least two of the metrics adopted. In general, what was proposed in this work can be easily applied and used in autonomous time series forecasting systems with the use of exogenous variables, more specifically for forecasting solar radiation.

ACKNOWLEDGMENT

This work is a result of the research and development project entitled "Technical arrangement to increase reliability and electrical safety by applying energy storage by batteries and photovoltaic systems to the auxiliary service of 230/500 kV substations", from the Public Call - P&D+I Nº 02/2019 with financing from Companhia Hidro Elétrica do São Francisco (CHESF). This is an initiative within the scope of ANEEL's (Agência Nacional de Energia Elétrica) Research and Technological Development Program for the Electric Energy Sector, under execution by the University of Pernambuco (UPE), Edson Mororó Moura Technology Institute – ITEMM and Itaipu Technological Park Foundation – PTI. Finally, we thank Chesf and the "Superintendência de Pesquisa e Desenvolvimento e Eficiência Energética" (SPE) / ANEEL for their support in making available the resources that allowed the preparation of this work.

REFERENCES

[1] "Solar pv on track." https://www.iea.org/reports/solar-pv, 2021. Accessed: 2021-01-03.

[2] A. R. O. da Rosa and F. P. Gasparin, "Panorama da energia solar fotovoltaica no brasil," *Revista brasileira de energia solar*, vol. 7, no. 2, pp. 140–147, 2016.

[3] E. Pereira, F. Martins, A. Gonçalves, R. Costa, F. Lima, R. Rüther, S. Abreu, G. Tiepolo, S. Pereira, and J. Souza, "Atlas brasileiro de energia solar–2ª edição, são josé dos campos," 2017.

[4] "Evolução da capacidade instalada no sin." http://www.ons.org.br/paginas/sobre-o-sin/o-sistema-em-numeros, 2021. Accessed: 2021-02-20.

[5] J. Barbosa, L. P. Dias, S. G. Simoes, and J. Seixas, "When is the sun going to shine for the brazilian energy sector? a story of how modelling affects solar electricity," *Renewable Energy*, vol. 162, pp. 1684–1702, 2020.

[6] C. L. de Azevedo Dias, D. A. C. Branco, M. C. Arouca, and L. F. L. Legey, "Performance estimation of photovoltaic technologies in brazil," *Renewable Energy*, vol. 114, pp. 367–375, 2017.

[7] S. Sobri, S. Koohi-Kamali, and N. A. Rahim, "Solar photovoltaic generation forecasting methods: A review," *Energy Conversion and Management*, vol. 156, pp. 459–497, 2018.

[8] H. Wang, Z. Lei, X. Zhang, B. Zhou, and J. Peng, "A review of deep learning for renewable energy forecasting," *Energy Conversion and Management*, vol. 198, p. 111799, 2019.

[9] A. Mellit, A. Massi Pavan, E. Ogliari, S. Leva, and V. Lughi, "Advanced methods for photovoltaic output power forecasting: A review," *Applied Sciences*, vol. 10, no. 2, p. 487, 2020.

[10] H. A. Kazem and M. T. Chaichan, "Effect of humidity on photovoltaic performance based on experimental study," *International Journal of Applied Engineering Research (IJAER)*, vol. 10, no. 23, pp. 43572–43577, 2015.

[11] B. Hailegnaw, S. Kirmayer, E. Edri, G. Hodes, and D. Cahen, "Rain on methylammonium lead iodide based perovskites: possible environmental effects of perovskite solar cells," *The journal of physical chemistry letters*, vol. 6, no. 9, pp. 1543–1547, 2015.

[12] Y. Sun, F. Wang, B. Wang, Q. Chen, N. Engerer, and Z. Mi, "Correlation feature selection and mutual information theory based quantitative research on meteorological impact factors of module temperature for solar photovoltaic systems," *Energies*, vol. 10, no. 1, p. 7, 2017.

[13] C. Voyant, G. Notton, S. Kalogirou, M.-L. Nivet, C. Paoli, F. Motte, and A. Fouilloy, "Machine learning methods for solar radiation forecasting: A review," *Renewable Energy*, vol. 105, pp. 569–582, 2017.

[14] P. Chaudhary and M. Rizwan, "Energy management supporting high penetration of solar photovoltaic generation for smart grid using solar forecasts and pumped hydro storage system," *Renewable Energy*, vol. 118, pp. 928–946, 2018.

[15] M. Feurer, K. Eggensperger, S. Falkner, M. Lindauer, and F. Hutter, "Practical automated machine learning for the automl challenge 2018," in *International Workshop on Automatic Machine Learning at ICML*, pp. 1189–1232, 2018.

[16] M. Feurer, K. Eggensperger, S. Falkner, M. Lindauer, and F. Hutter, "Auto-sklearn 2.0: The next generation," *arXiv preprint arXiv:2007.04074*, 2020.

[17] F. Assunção, N. Lourenço, B. Ribeiro, and P. Machado, "Evolution of scikit-learn pipelines with dynamic structured grammatical evolution," in *International Conference on the Applications of Evolutionary Computation (Part of EvoStar)*, pp. 530–545, Springer, 2020.

[18] H. Yu, L. J. Ming, R. Sumei, and Z. Shuping, "A hybrid model for financial time series forecasting—integration of ewt, arima with the improved abc optimized elm," *IEEE Access*, vol. 8, pp. 84501–84518, 2020.

[19] X. He, K. Zhao, and X. Chu, "Automl: A survey of the state-of-the-art," *Knowledge-Based Systems*, vol. 212, p. 106622, 2021.

[20] L. Xiong and Y. Lu, "Hybrid arima-bpnn model for time series prediction of the chinese stock market," in *2017 3rd International Conference on Information Management (ICIM)*, pp. 93–97, IEEE, 2017.

[21] S. d. O. Domingos, J. F. de Oliveira, and P. S. de Mattos Neto, "An intelligent hybridization of arima with machine learning models for time series forecasting," *Knowledge-Based Systems*, vol. 175, pp. 72–86, 2019.

[22] S. Smyl, "A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting," *International Journal of Forecasting*, vol. 36, no. 1, pp. 75–85, 2020.

[23] J. F. de Oliveira, E. G. Silva, and P. S. de Mattos Neto, "A hybrid system based on dynamic selection for time series forecasting," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[24] M. Alanazi, M. Mahoor, and A. Khodaei, "Two-stage hybrid day-ahead solar forecasting," in *2017 North American Power Symposium (NAPS)*, pp. 1–6, IEEE, 2017.

[25] C. Feng and J. Zhang, "Hourly-similarity based solar forecasting using multi-model machine learning blending," in *2018 IEEE Power & Energy Society General Meeting (PESGM)*, pp. 1–5, IEEE, 2018.

[26] H. Zang, L. Liu, L. Sun, L. Cheng, Z. Wei, and G. Sun, "Short-term global horizontal irradiance forecasting based on a hybrid cnn-lstm model with spatiotemporal correlations," *Renewable Energy*, vol. 160, pp. 26–41, 2020.

[27] S. Ghimire, R. C. Deo, N. Raj, and J. Mi, "Deep solar radiation forecasting with convolutional neural network and long short-term memory network algorithms," *Applied Energy*, vol. 253, p. 113541, 2019.

[28] H. Ramchoun, M. A. J. Idrissi, Y. Ghanou, and M. Ettaouil, "Multilayer perceptron: Architecture optimization and training.," *IJIMAI*, vol. 4, no. 1, pp. 26–30, 2016.

[29] M. A. J. Idrissi, H. Ramchoun, Y. Ghanou, and M. Ettaouil, "Genetic algorithm for neural network architecture optimization," in *2016 3rd International Conference on Logistics Operations Management (GOL)*, pp. 1–4, IEEE, 2016.

[30] H. A. Mendes, H. F. Nunes, and P. S. de Mattos Neto, "Integração de dados e modelos de previsão de produção fotovoltaica do nordeste brasileiro," *Revista de Engenharia e Pesquisa Aplicada*, vol. 5, no. 2, pp. 62–72, 2020.

[31] D. Pradeepkumar and V. Ravi, "Forecasting financial time series volatility using particle swarm optimization trained quantile regression neural network," *Applied Soft Computing*, vol. 58, pp. 35–52, 2017.

[32] M. Shen, W.-N. Chen, J. Zhang, H. S.-H. Chung, and O. Kaynak, "Optimal selection of parameters for nonuniform embedding of chaotic time series using ant colony optimization," *IEEE Transactions on Cybernetics*, vol. 43, no. 2, pp. 790–802, 2013.

[33] A. Sinha, P. Malo, and K. Deb, "A review on bilevel optimization: from classical to evolutionary approaches and applications," *IEEE Transactions on Evolutionary Computation*, vol. 22, no. 2, pp. 276–295, 2017.

[34] "Bdmep-instituto nacional de meteorologia." https://bdmep.inmet.gov.br/, 2021. Accessed: 2021-01-21.

[35] E. F. Lima, R. G. S. Moraes, B. L. A. S. Fonseca, and C. M. da Silva, "Comparação de dados meteorológicos obtidos por estação convencional e automática e estimativa da evapotranspiração de referência em imperatriz/ma," *Journal of Environmental Analysis and Progress*, vol. 5, no. 2, pp. 214–220, 2020.

[36] P. S. dos Anjos, A. S. A. da Silva, B. Stošić, and T. Stošić, "Long-term correlations and cross-correlations in wind speed and solar radiation temporal series from fernando de noronha island, brazil," *Physica A: Statistical Mechanics and its Applications*, vol. 424, pp. 90–96, 2015.

[37] T.-P. Chang, F.-J. Liu, H.-H. Ko, and M.-C. Huang, "Oscillation characteristic study of wind speed, global solar radiation and air temperature using wavelet analysis," *Applied energy*, vol. 190, pp. 650–657, 2017.

[38] A. de Almeida Brito, H. A. de Araújo, and G. F. Zebende, "Detrended multiple cross-correlation coefficient applied to solar radiation, air temperature and relative humidity," *Scientific reports*, vol. 9, no. 1, pp. 1–10, 2019.

[39] S.-H. Lee, Y.-Y. Hong, C.-C. Chu, M.-Y. Lee, and H.-P. Liu, "Impacts of wind gust and cloud-shading on renewable energy generation," in *2019 IEEE 4th International Future Energy Electronics Conference (IFEEC)*, pp. 1–7, IEEE, 2019.

[40] K. W. Hipel, A. I. McLeod, and W. C. Lennox, "Advances in box-jenkins modeling: 1. model construction," *Water Resources Research*, vol. 13, no. 3, pp. 567–575, 1977.

[41] S. Atique, S. Noureen, V. Roy, V. Subburaj, S. Bayne, and J. Macfie, "Forecasting of total daily solar energy generation using arima: A case study," in *2019 IEEE 9th annual computing and communication workshop and conference (CCWC)*, pp. 0114–0119, IEEE, 2019.

[42] R. Hyndman and G. Athanasopoulos, "Forecasting: principles and practice, 2nd edn. otexts, melbourne, australia," 2020.

[43] D. Kwiatkowski, P. C. Phillips, P. Schmidt, and Y. Shin, "Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?," *Journal of econometrics*, vol. 54, no. 1-3, pp. 159–178, 1992.

[44] D. A. Dickey and W. A. Fuller, "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American statistical association*, vol. 74, no. 366a, pp. 427–431, 1979.

[45] F. Canova and B. E. Hansen, "Are seasonal patterns constant over time? a test for seasonal stability," *Journal of Business & Economic Statistics*, vol. 13, no. 3, pp. 237–252, 1995.

[46] T. G. Smith *et al.*, "pmdarima: Arima estimators for python," 2017.

[47] R. J. Hyndman, Y. Khandakar, *et al.*, *Automatic time series for forecasting: the forecast package for R*. No. 6/07, Monash University, Department of Econometrics and Business Statistics . . . , 2007.

[48] M. Clerc, "Discrete particle swarm optimization, illustrated by the traveling salesman problem," in *New optimization techniques in engineering*, pp. 219–239, Springer, 2004.

[49] E. M. Figueiredo and T. B. Ludermir, "Investigating the use of alternative topologies on performance of the pso-elm," *Neurocomputing*, vol. 127, pp. 4–12, 2014.

[50] C. M. Hurvich and C.-L. Tsai, "Regression and time series model selection in small samples," *Biometrika*, vol. 76, no. 2, pp. 297–307, 1989.

[51] Q. V. Le, J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, and A. Y. Ng, "On optimization methods for deep learning," in *ICML*, 2011.

[52] D. Yi, J. Ahn, and S. Ji, "An effective optimization method for machine learning based on adam," *Applied Sciences*, vol. 10, no. 3, p. 1073, 2020.

[53] X. Qing and Y. Niu, "Hourly day-ahead solar irradiance prediction using weather forecasts by lstm," *Energy*, vol. 148, pp. 461–468, 2018.