A precipitation forecasting model using artificial neural network in central ecotone region in Brazil

Hudson Pena Magalhães [1] e
Tiago da Silva Almeida [2]

Data de submissão: 25/11/2020. Data de aprovação: 29/1/2021.

Abstract – Precipitation forecasting may be of great value for farming, helping to reduce crop losses and irrigation costs, besides leveraging crop yield estimates. In Brazil, the state of Tocantins has a great part of its economy based on agriculture, where precipitation forecasting may help improve local production. Besides, rainfall forecasting can also contribute to urban planning through the management of water resources, among other applications Artificial Neural Networks (ANNs) have been used with relative success in precipitation forecasting for different locations and climates. With that, this work presents a method for weekly precipitation forecasting in six locations in the Brazilian state of Tocantins using ANNs and public climatic data. For that, MultiLayer Perceptron (MLP) networks were trained with data from local weather stations and El Niño Southern Oscillation (ENSO) related indices. First, input variables were selected using the forward selection algorithm. After that, ANN hyperparameters and input variables lag were optimized. The average Root Means Square Error (RMSE) of the final models was of 31.35 mm/week for the training dataset and 33.38 mm/week for the test dataset. Respectively, these values represent 9,83% and 10,46% of the maximum weekly precipitation found in the work dataset, which was of 319.1 mm. The results suggest that the created models are capable of reasonably good weekly precipitation forecasts, providing valuable information for farming, water resources management, urban planning and other related activies. Although there is possibly room for model improvement.
Keywords: Precipitation forecasting. Artificial neural network. Irrigated Agriculture. Water Resources Planning. Meteorology.

Um modelo de previsão de precipitação usando rede neural artificial na região central do ecótono no Brasil

Resumo – A previsão de precipitação pode ser de grande valor para a agricultura, ajudando a reduzir as perdas de safra e os custos de irrigação, além de alavancar as estimativas de produtividade das lavouras. No Brasil, o estado do Tocantins tem grande parte de sua economia baseada na agricultura, onde a previsão da precipitação pode ajudar a melhorar a produção local. Além disso, a previsão de precipitação também pode contribuir com o planejamento urbano por meio da gestão de recursos hídricos, entre outras aplicações. Redes Neurais Artificiais (RNAs) têm sido usadas com relativo sucesso na previsão de precipitação para diferentes locais e climas. Com isso, este trabalho apresenta um método para previsão de precipitação semanal em seis localidades do estado do Tocantins, utilizando RNAs e dados climáticos públicos. Para isso, redes *Perceptron* Multicamadas (MLP) foram treinadas com dados de estações meteorológicas locais e índices relacionados ao El Niño Oscilação Sul (ENSO). Primeiro, as variáveis de entrada foram selecionadas usando o algoritmo de seleção direta. Depois disso, os hiper parâmetros da RNA e a defasagem das variáveis de entrada foram otimizados. A raiz média do erro quadrático (RMSE) dos modelos finais foi de 31,35 mm / semana para o conjunto de dados

[1] Especialista no Programa de Pós-graduação em Sistemas de Apoio à Decisão, *Campus* Palmas, da Universidade Federal do Tocantins - UFT. hudmagalhaes@gmail.com. ORCID: https://orcid.org/0000-0002-8543-6895.
[2] Professor do curso de Ciência da Computação, *Campus* Palmas, da Universidade Federal do Tocantins - UFT. tiagoalmeida@uft.edu.br. ORCID: https://orcid.org/0000-0002-0420-4188.

de treinamento e 33,38 mm / semana para o conjunto de dados de teste. Respectivamente, esses valores representam 9,83% e 10,46% da precipitação máxima semanal encontrada no conjunto de dados do trabalho, que foi de 319,1 mm. Os resultados sugerem que os modelos criados são capazes de previsões de precipitação semanais razoavelmente boas, fornecendo informações valiosas para agricultura, gestão de recursos hídricos, planejamento urbano e outras atividades relacionadas. Embora possivelmente haja espaço para melhorias no modelo.
Palavras-chave: Previsão de precipitação. Rede Neural Artificial. Agricultura Irrigada. Gestão de Recursos Hídricos. Meteorologia.

Introduction

In meteorology, precipitation refers to the process in which water vapor condenses in the atmosphere and falls back to the surface in any form, such as rain, dew, snow and sleet. As with many human activities, agriculture can be strongly impacted by precipitation, depending on how much, for how long and when precipitation occurs. Therefore, precipitation-forecasting models can provide valuable information to support decision-making. In this regard, Asseng *et al*. (2016) suggest that short-term rainfall forecasts may benefit productivity in dryland agriculture, supporting decisions about the sowing time and application of fertilizers and fungicides. Cardoso *et al*. (2010) suggest that the use of precipitation forecast data in soybean yield estimates can lead to more reliable estimates. As suggested by Cao *et al*. (2019), rainfall forecasts can also help to optimize irrigation scheduling and reduce water expenditure.

Artificial Neural Networks (ANNs), MLP (MultiLayer Perceptron) in this research, are a type of supervised machine learning which are capable to learn nonlinear relationships between variables found in the given input data. For being a data-driven technique, no actual knowledge of the equations that represent such relationships is required when creating an ANN model. This non-parametric characteristic, along with the often-good results, make ANNs great tools for complex problems, such as precipitation forecasting.

Many works have used ANNs to forecast precipitation and rainfall using different weather variables and climatic indices as model input data. While weather variables can provide information about the atmospheric conditions of a given location, climatic indices can indicate climate anomalies that may be related to large-scale phenomena, such as the El Niño Southern Oscillation (ENSO). In a very simplified manner, ENSO is a variation of the sea surface temperature (SST) and the air pressure over the equatorial Pacific Ocean. Extreme weather conditions, such as floods and droughts, can be experienced during ENSO events in parts of South America, South Asia and Australia (SCAIFE *et al*., 2019). As a result, ENSO can cause major impact on the economy, especially in agribusiness (SCAIFE *et al*., 2019; ANDERSON *et al*., 2017).

In this regard, Abbot and Marohasy (2014) have created ANN models for rainfall forecasting using local weather variables and climatic indices, reporting better results than the General Circulation Model (GCM) used by local government institutions in Australia. Aspects of the soil can also be considered as predictor variables in these models. Soil water content, which refers to the amount of water the soil can retain, was used by Esteves *et al*. (2019) as one of the input variables of an ANN model used for rainfall forecasting. By applying ANNs to Doppler weather radar data, Dutta *et al*. (2011) was able to improve the estimation of rainfall intensity compared to traditional estimation methods. In addition, a relevant list of works on the topic, published between 2012 and 2017, is found in Abbot and Marohasy (2017).

During model construction, the ANNs learn the mechanics of the target problem directly from the data provided to the network, usually meaning that as more data is available more is learned, thereby reducing output error. When working with time series data, such as weather data, missing observations can be very common, whether due to equipment failure or staff unavailability, when working with non-automatic weather stations. There are many data gap

filling techniques, ranging from simple arithmetic mean to linear regression and the ANN models themselves. Some of these techniques were evaluated by de Oliveira *et al*. (2010) for annual precipitation data, and by Bier and Ferraz (2017) for monthly precipitation and air temperature.

One of the challenging tasks of ANN model creation is input variable selection, where the available variables are evaluated in order to find the best combination of variables for the problem. As ANNs are usually applied to nonlinear problems, some works have used the ANN models themselves to guide the variable selection process (ABBOT and MAROHASY, 2014, 2017). This approach usually requires a lot of computational power, depending on the amount of input data and variables combinations available. That said, some methods tend to use linear correlation analysis for that task, whether to select the final combination or to identify the best candidates for later selection using ANNs (LEE *et al*., 2018; AKRAMI *et al*., 2013).

GCMs are complex mathematical models that take into account the general circulation of the global atmosphere and the oceans, serving as a basis for weather forecasting and climate study. When using ANN models, it is possible to work on a much smaller scale, focusing on a small subset of variables that may impact on a large portion of the problem. In fact, reducing the number of variables, aside from reducing the model complexity, reduces the computational power required to run the model.

The state of Tocantins, in the north region of Brazil, has most of its economy based on the agribusiness, where livestock farming takes the lead, followed by grains cropping, especially soybeans. In Tocantins, the soil suitable for the cultivation of grains is found in several patches over the 277,720 Km2 of territory, where the predominant biome is the Brazilian Cerrado. This, along with precipitation conditions that vary depending on the region, lead to very dispersed producing regions (SECRETARIA DO PLANEJAMENTO E ORÇAMENTO DO ESTADO DO TOCANTINS, 2016).

In Brazil, the Meteorology National Institute (INMET) is a federal agency that publicly provides local weather information making use of dozens of weather stations across the country. Likewise, data for the ENSO climatic indices are provided by the National Oceanic Atmospheric and Administration (NOAA), which is a United States government agency. Both organizations provide time series data which can be used as input for ANN forecasting models.

Considering the exposed scenario, this work aims to create a weekly precipitation forecasting model of 24 weeks ahead in the state of Tocantins, having as output the accumulated precipitation of each of the forecasted weeks. In order to achieve this objective, ANN models were created using up to 52 weeks of lagged data as input. The input data included INMET local weather stations observations and ENSO climatic indices, as provided by NOAA. Although many works have applied ANNs to precipitation forecasting, none of the works found during review were created for locations in the state of Tocantins. It is expected that the proposed model may provide reasonably accurate information and thus contribute to the local agribusiness activities and research.

Materials and Methods

The INMET provides data from weather stations all over Brazilian territory, having 194 conventional stations and 576 automatic stations available for query on their website[3] at the time of this paper. While conventional stations require dedicated staff to take readings on site, automatic stations can transmit data automatically through wireless networks (INSTITUTO NACIONAL DE METEOROLOGIA, 2011). Another aspect observed in this stations network was that, in general, conventional stations were installed earlier in time compared to automatic stations, consequently having longer time series available.

[3] http://www.inmet.gov.br/

Covering the Tocantins state territory, there were 20 automatic and 6 conventional stations. As depicted in Figure 1, those conventional stations were dispersed throughout the state, which allowed a reasonable overview of the weather on the state. Considering the territory coverage and the longer time series, the 6 conventional stations in Tocantins were selected as forecasting target for this work. Other stations, including those located in neighbor states, were used during the gap filling process, as described later in this work. Table 1 lists all the stations used in this work.

Figure 1 - Conventional weather stations located in Tocantins state territory.

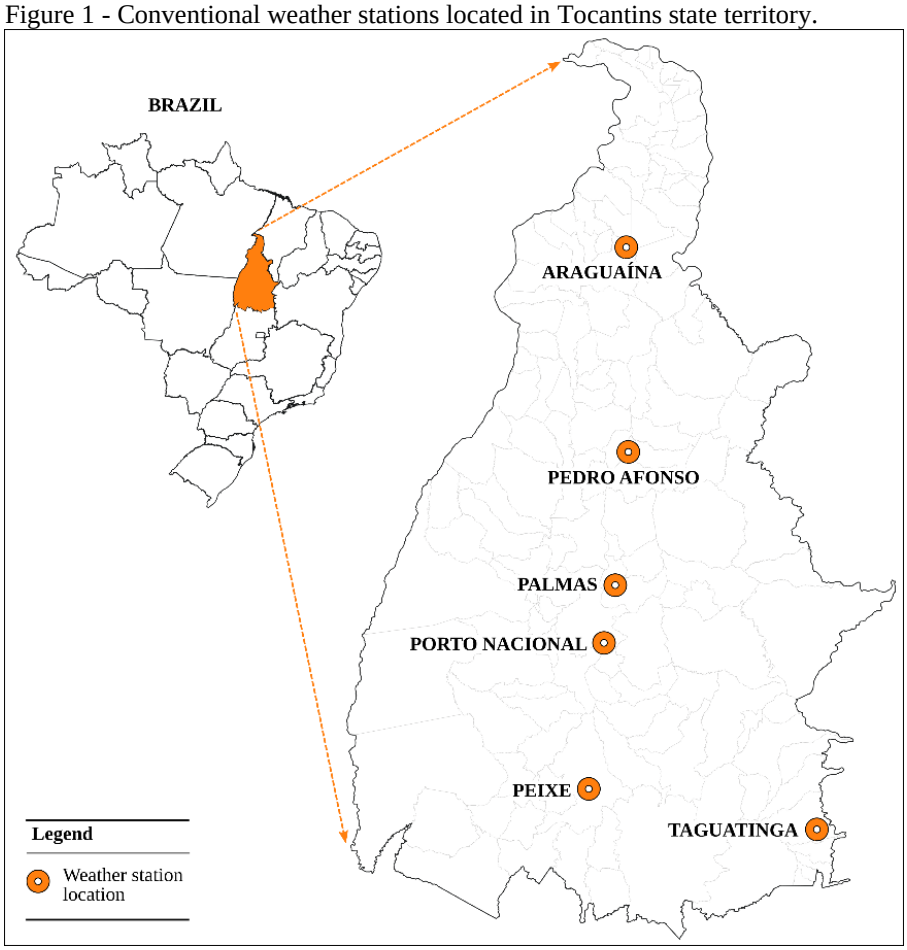


Table 1 - INMET weather stations used in this work.

No.	Area	WMO(1)	T(2)	U(3)	Latitude(º)	Longitude(º)	Altitude(m)
1	Almas/TO	86627	A	S	-11.284098	-47.212125	503.0
2	Araguaçu/TO	86648	A	S	-12.592213	-49.528738	231.85
3	**Araguaína/TO**	**82659**	**C**	**T**	**-7.103778**	**-48.20133**	**231.85**
4	Araguaína/TO	81900	A	S	-7.103954	-48.201231	231.85
5	Araguatins/TO	81821	A	S	-5.643725	-48.111839	131.0
6	Campos Lindos/TO	81902	A	S	-8.154665	-46.639323	427.0
7	Carolina/MA	82765	C	S	-7.337292	-47.459856	182.94
8	Carolina/MA	81901	A	S	-7.337269	-47.459839	183.0
9	Colinas do Tocantins/TO	81939	A	S	-8.092708	-48.478605	200.0
10	Conceição do	82861	C	S	-8.259237	-49.263816	179.02

	Araguaia/PA						
11	Conceição do Araguaia/PA	81940	A	S	-8.30361	-49.28277	176.0
12	Dianópolis/TO	86632	A	S	-11.594448	-46.847209	728.0
13	Estreito/MA	81863	A	S	-6.653272	-47.418241	183.0
14	Formoso do Araguaia/TO	86629	A	S	-11.887377	-49.608215	215.0
15	Gurupi/TO	86630	A	S	-11.745782	-49.049703	279.0
16	Imperatriz/MA	82564	C	S	-5.536521	-47.478943	126.33
17	Imperatriz/MA	81822	A	S	-5.555723	-47.459794	118.0
18	Lagoa da Confusão/TO	86602	A	S	-10.828286	-49.847882	178.0
19	Marianópolis do Tocantins/TO	81983	A	S	-9.576389	-49.72333	187.0
20	Mateiros/TO	86608	A	S	-10.434441	-45.921941	791.0
21	Monte Alegre de Goiás/GO	86670	A	S	-13.253521	-46.890326	551.0
22	**Palmas/TO**	**83033**	**C**	**T**	**-10.190897**	**-48.301822**	**291.68**
23	Palmas/TO	86607	A	S	-10.190744	-48.301811	292.0
24	Paranã/TO	86650	A	S	-12.614893	-47.871917	285.0
25	**Pedro Afonso/TO**	**82863**	**C**	**T**	**-8.968576**	**-48.177264**	**189.53**
26	Pedro Afonso/TO	81941	A	S	-8.968677	-48.177259	190.0
27	**Peixe/TO**	**83228**	**C**	**T**	**-12.015387**	**-48.544866**	**252.24**
28	Peixe/TO	86649	A	S	-12.015377	-48.544517	251.0
29	Pium/TO	86603	A	S	-10.476944	-49.629475	161.0
30	Porangatu/GO	n/a	A	S	-13.309528	-49.117478	365.0
31	**Porto Nacional/TO**	**83064**	**C**	**T**	**-10.710716**	**-48.406362**	**243.28**
32	Rio Sono/TO	81981	A	S	-9.793363	-47.132732	291.0
33	Santa Fé do Araguaia/TO	81898	A	S	-7.124191	-48.781267	171.0
34	Santa Rosa do Tocantins/TO	86631	A	S	-11.429018	-48.184889	306.0
35	São Miguel do Araguaia/GO	86646	A	S	-12.820489	-50.335969	210.0

(1) World Meteorological Organization code; (2) station type, being automatic (A) or conventional (C); (3) station usage, being forecasting target (T) or support (S).

The weather data from conventional stations was collected directly from the INMET website. This data contained daily observations realized at 09:00 AM and 09:00 PM local time (UTC-3), though, each variable was available only once a day. For instance, precipitation was available only at 09:00 AM. In order to transform those into single daily records, each observation was composed of data from the same day at 09:00 PM and data from the next day at 09:00 AM. Since each observation represents the mean or accumulated value of the last 24 hours, observations at 09:00 AM include only 9 hours of the referred day. Therefore, it was expected to get a better representation of each day with the described composition.

Historical data for the automatic stations was not available on the institute's website, having been delivered by mail, upon request. In this data, the observations were available for

each hour of the day. When transforming each variable into single daily records only those days with all 24 observations available were considered. In order to maintain consistency with the conventional stations data, the same time scheme was used here. For example, precipitation data was composed of observations from 10:00 PM on the same day up to 09:00 AM on the following day.

From INMET data the following variables were extracted: minimum air temperature (MIN_T), maximum air temperature (MAX_T), compensated mean air temperature (MEAN_T), precipitation (P), relative air humidity (HU), photoperiod (SUN), wind speed (WS) and tar evaporation (TAR).

Moving on to the ENSO indices, their relevance to the current problem is related to teleconnections, which are statistical correlations between climatic variables whose observation locations are separated by very far distances. As explained by Lee *et al*. (2018), due to the general circulation of the atmosphere and the oceans, regional climates are linked together in a global scale system. As suggested by other works (ABBOT and MAROHASY, 2014, 2017; MAROHASY and ABBOT, 2015; LEE *et al*., 2018), it is a relevant candidate as input for precipitation forecasting models.

ENSO observation data were collected from NOAA website[4], where monthly data were available for the following variables: Southern Oscillation Index (SOI), Niño 1+2 region SST (NI_1.2), Niño 1+2 SST anomaly (NA_1.2), Niño 3 SST (NI_3), Niño 3 SST anomaly (NA_3), Niño 3.4 SST (NI_3.4), Niño 3.4 SST anomaly (NA_3.4), Niño 4 SST (NI_4) and Niño 4 SST anomaly (NA_4). Observations from Niño regions were available from January 1982 through December 2019, while SOI was available from January 1951 through December 2019.

Dataset Creation

In order to build a weekly dataset, the year was divided into 52 weeks, always starting on the 1st of January of each year, despite the actual day of the week on the calendar. February 29, in case of leap years, and December 31 were inserted as additional days in weeks 9 and 52, respectively. During daily to weekly format transformation, each variable was processed individually, considering only those weeks where data was available for all days of that week.

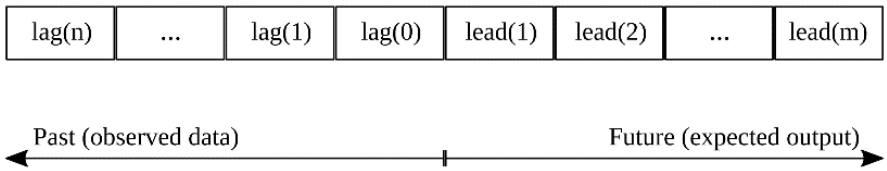
Many gaps have been found in the time series of the weather stations data, which has reduced the actual amount of data available for ANN training. When modeling time series problems with ANNs, data is usually provided to the network as an ordered sequence, where missing steps can invalidate the entire sequence. This structure is illustrated in Figure 2. In order to reduce the problem, estimated data can be used to fill those gaps. As evaluated by de Oliveira *et al*. (2010) and Bier and Ferraz (2017), several methods can be used to fill gaps in time series of precipitation and air temperature. In addition, these methods are also suggested for application in other weather variables (BIER and FERRAZ, 2017). Among those methods, regional weighting was proved to be reasonably simple and effective. Estimation of a missing observation using this method is given by the following equation:

$$D_x = \frac{1}{x}\sum_{i=1}^{n}\frac{M_x}{M_i}D_i \qquad (1)$$

Where $D_x$the weekly estimated value, D_i is the corresponding weekly value from the i^{th} neighbour station, M_x is the weekly mean value from the target station, M_i is weekly mean value from the i^{th} neighbour and n is the number of neighbour stations.

Figure 2 - Input sequence for ANN training

[4] https://www.ncdc.noaa.gov/teleconnections/enso/indicators/soi/



The first step to apply this method is to determine which neighbor stations will provide data for estimation. Tabony (1983), cited by Bier and Ferraz (2017), suggests that neighbors should be selected based on their statistical correlation with the target station. Due to this fact, different neighbors should be selected for each estimated variable. The author also suggests that neighbors should be positioned the most evenly as possible around the target station, increasing weather representation. For this work, neighbor stations were limited to a range of 200 Km around the target station. After that, correlation was calculated for each variable, where those stations with correlation higher than 0.7 were selected. Table 2 lists the forecasting target stations with their respective neighbors for each weather variable.

Table 2 - Target stations with their respective neighbors for each variable.

St. (1)	**Neighbor stations by variable**						
	P	**MAX_T**	**MIN_T**	**SUN, TAR**	**MEAN_T**	**HU**	**WS**
3	4, 9, 8	4, 8, 7, 10, 25, 33, 5, 13, 26, 17, 9, 16, 6, 11	4, 9, 33, 13, 26, 17, 25, 8, 6, 5, 11, 7	7, 25, 10, 16	4, 9, 25, 26, 13, 33, 17, 8, 6	4, 8, 9, 33, 13, 5, 26, 7, 6, 11, 25, 17, 16, 10	n/a
22	34, 18, 23	23, 31, 26, 32, 25, 34, 29, 19, 1, 18, 15, 28, 27	18, 26, 34, 25, 1, 31, 23	27, 31, 25	32, 28, 1, 25, 34, 19, 31, 26, 23	23, 31, 25, 26, 32, 34, 29, 19, 1, 18, 15, 27, 28	34, 32
25	26	9, 11, 3, 4, 19, 22, 23, 10, 31, 6, 7, 8, 26	22, 8, 11, 7, 19, 31, 3, 4, 32, 9, 26	7, 10, 31, 3, 22	10, 23, 7, 8, 3, 32, 22, 4, 19, 6, 9, 31, 26	10, 7, 31, 3, 11, 19, 4, 22, 8, 23, 9, 6, 32, 26	n/a
27	24, 15, 18, 34, 2, 28	28, 15, 34, 24, 14, 2, 31, 30, 1, 12, 18, 22, 23, 29, 35	28, 35, 15, 34, 24, 30, 29, 2, 18, 14, 1, 31, 23	22, 31	15, 28, 2, 24, 18, 35, 34, 29, 1, 30, 31, 12, 14	28, 34, 15, 18, 2, 1, 35, 29, 14, 31, 24, 12, 23, 22, 30	23, 35, 14, 12, 34, 30
31	1, 32, 34, 18	23, 22, 34, 15, 29, 28, 27, 1, 18, 32, 14, 19, 26, 25, 12	14, 15, 26, 23, 28, 22, 25, 27	22, 25, 27	32, 18, 29, 12, 1, 19, 34, 14, 15, 26, 23, 28, 22, 25, 27	32, 18, 29, 12, 1, 19, 34, 14, 15, 26, 23, 28, 22, 25, 27	n/a

36	1	34, 27, 21, 20, 1, 28, 24, 12	1, 12	27	27, 20, 24, 28, 34, 1, 21, 12	24, 34, 28, 1, 20, 27, 21, 12	20, 34, 12

(1) Target station number.

ANNs are supervised learning techniques, which means that the network learns by adjusting itself to the provided data. Thus, the model error is measured by the difference between estimated output and the expected output, in other words, output error. In order to avoid erroneous model evaluation, values estimated using the regional weighting were used only as predictor variables, never as expected output.

As initially proposed, the ANN models are to forecast up to 24 weeks ahead using data from the 52 past weeks. Following this scheme, each dataset row is composed of data from all predictor variables from the last 52 weeks and the forecasted variable (P) from the next 24 weeks. Due to this aspect, the gap filling method had a great impact on the number of rows available for usage. The final dataset numbers are presented in Table 3.

Table 3 - Available data rows for each target weather station.

Station No.	**Area**	**Available rows**	
		After gap filling	**Before gap filling**
3	Araguaína	1180	583
22	Palmas	923	204
25	Pedro Afonso	1215	429
27	Peixe	1492	596
31	Porto Nacional	750	11
36	Taguatinga	924	125

Lastly, the data was divided into training and testing data, respectively 70% and 30% of the available data. In addition, to avoid statistical bias between variables of different scales, all values were normalized between 0 and 1 using the following equation:

$$N_x = \frac{x - M_i}{Ma_x - Mi_x} \quad (2)$$

Where N_x is the normalized value of x, $Mi_x$is x minimum value and Ma_x is x maximum value.

The final datasets were composed of the 17 weather and climatic variables described in Material e Methods Section, and the number of the corresponding week (W).

The ANN Model

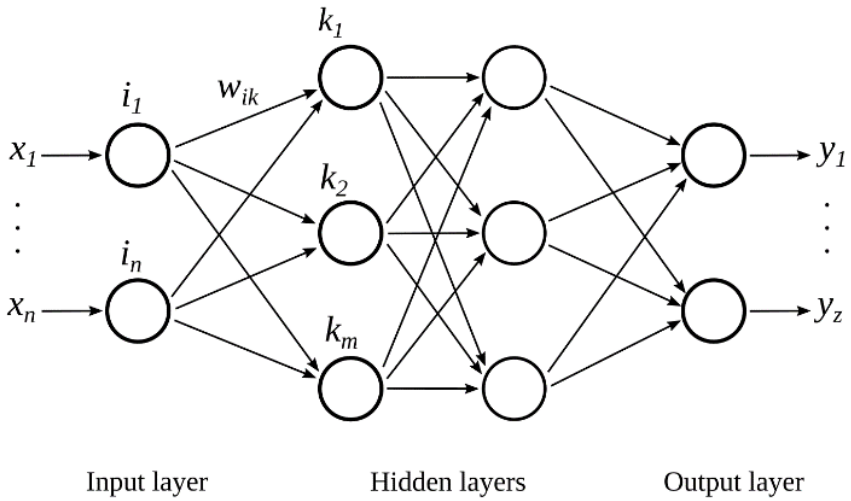
In short, ANNs are computational techniques that can learn linear and nonlinear relations between variables found in each dataset. There are many ANN types, where MLP is among the most common types. These networks are composed of many interconnected nodes arranged in layers, where data flows from input to output layer. Each node is a processing unit which applies

an activation function over the sum of all weighted input values, as denoted by the following equation (VELO *et al.*, 2014):

$$O_i = f\left(w_0 x_0 \sum_{j=1}^{n} w_j x_j\right) \quad (3)$$

Where O_i is the i^{th} node output value, f is the activation function, w_j is the weight value, x_j is the input value, w_0 is a threshold value (usually called bias), x_0 is always 1, and n is the number of input connections. In order to fit the network to the given model, sample data is repeatedly provided to the network, where output error is propagated back through the network for weights adjustment. Thus, if it was properly constructed, the network tends to slowly converge to optimal error. Figure 3 illustrates the general structure of an MLP.

Figure 3 - General structure of a multilayer perceptron network.



Once the ANN type is established, it is required to set the network hyperparameters, which are responsible for defining a great part of the network's behavior. Although being specific to the network mechanisms, their optimal values are often related to the applied model. In this work, MLPs were implemented using the Deep Learning for Java (DL4J)[5] library, and hyperparameters were set after trial and error.

MLP learning rate was set to 0.0001 and learning momentum to 0.90. Weights were randomly initialized, and the number of training epochs was limited to 300. Rectified Linear Units (ReLu) was used as node activation function and Mean Square Error (MSE) as loss function. ReLu is a nonlinear function that, for values below zero, returns zero, otherwise repeats the input value. ReLu has been used for many types of neural networks because a model that uses it is easier to train and often achieves better performance. Root Mean Square Error (RMSE) was the base metric for model analysis and optimization, while mean absolute error (MAE) was only used only to further illustrate results. In RMSE the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE is most useful when large errors are particularly undesirable. And the MAE is a linear score which means that all the individual differences are weighted equally in the average.

[5] https://deeplearning4j.org/

The size of the network proved to be very susceptible to the number of input variables. Thus, it was not optimal to keep a constant network size during model creation. The network depth (number of hidden layers) was kept as 2 until the model adjustment step, where it was optimized, as later described. For the number of nodes per hidden layer (layer width), reasonable results were found when using two heuristic strategies based on the size of the input and the output layers. The first strategy uses the average of those values, while the second uses the sum. On both cases, all hidden layers were set to the same width. As for the output layer, it always had 24 nodes, as each node provided output for each of the forecasted weeks.

As initially proposed, this work evaluated up to 52 weeks of lagged data for each input variable. For this, two strategies were established. The first strategy selects variables with maximum lag and tries to optimize the model by reducing the lag of each variable at a time. The second strategy selects variables with 4 weeks of lag and tries to optimize the model by increasing the lag of each variable.

In order to evaluate all combinations of input lag strategies and layer width strategies, 4 different models were created for each of the 6 target stations as enumerated in Table 4. It should be noted, however, that these model categories are numbered only for later reference, since there is not sequential relationship between them.

Table 4 - The model categories created for each station.

N.º	Initial input lag	ANN width strategy
1	52 weeks	Average
2	52 weeks	Sum
3	4 weeks	Average
4	4 weeks	Sum

After the initial MLP setup, the next step was to select input variables, which was done using the forward selection method. This is a search method in which, for each iteration, a candidate variable is appended to the model and then evaluated, if the model output improves, the candidate is confirmed, otherwise discarded. As suggested by May *et al*. (2011), the forward selection was preceded by a variable ranking that classified variables based on their isolated forecasting strength. It was expected that, with this strategy, the resulting selections would be shorter, since most relevant variables were evaluated, and possibly selected, before the others.

With the input variables selected, the next step was to adjust the lag of input variables and network depth, since both were constant up to this point. At each iteration of the algorithm, it tried to improve lag of each variable individually, reducing (when initially 52) or increasing (when initially 4) by 4 weeks at a time. In addition, for each variable, the algorithm tried to reduce output error by adjusting the network depth. The algorithm stop criterion was to complete an iteration over all variables without any improvement.

After that, an additional step was taken to optimize the number of training epochs using the test dataset error as metric. It was done iteratively, increasing the number of epochs by 200 in each iteration until there was no further improvement. The same process was also attempted using the Leaky ReLu activation function, which is a variation of ReLu that allows negative output values, as denoted by the following equation:

$$f(x) = max(0, x) + \alpha \times min\,(0, x) \tag{4}$$

Where α is a constant, in this case, set to 0.01.

In order to maintain consistent results and reasonable computation times on the forward selection and the two optimization steps, a threshold of what was considered error reduction

was established. This threshold was set to 0.00009 normalized RMSE, approximately 0.03 mm when denormalized.

Results and Discussion

In the variables ranking step, each model produced different results, having no agreement about the variables relevance order. Based on these rankings, the forward selection algorithm selected an average of 5.5 variables per model, having RMSE of approximately 33.53 mm for the training dataset and 35.43 mm for the test dataset. Table 5 lists the algorithm results by model category.

Comparing the selection results by input variables lag, the models with 52 weeks of lag (categories 1 and 2) were only 0.83 mm RMSE more accurate than the models with 4 weeks of lag (categories 3 and 4). Thus, despite having much more data to work with, the larger models were not able to effectively outperform the smaller models. This may indicate that much of the data included in those 52 weeks may have been irrelevant to the models. Comparing the ANN width strategies, results indicate that the strategy that produced larger ANNs performed slightly better (approximately 1.14 mm lower RMSE) for the models with shorter input lag (categories 3 and 4). As for models with longer input lag (categories 1 and 2) and, therefore, larger input vectors, the additional capacity provided by the strategy may have been insufficient, since better results were found only for the training dataset. Regarding the selected variables, the most commonly selected were: W (87.5% of the models), NI_1.2 (70.8% of the models) and NI_3 (58% of the models). Precipitation (P) was selected in only one model, suggesting low relevance as autoregressive variable in the studied scenario.

Table 5 - Forward selection average results by model category.

Model category	Selected variables	Input nodes	Hidden nodes	RMSE (mm)	
				Training dataset	Test dataset
1	5.0	260	284	32.74	35.07
2	5.5	286	620	32.10	36.35
3	3.2	13	37	35.47	35.46
4	8.3	33	115	33.80	34.84

After the model adjustments step, the average RMSE was reduced in approximately 3.23% for the training dataset and 4.31% for the test dataset in relation to the forward selection score. The ANN depth was adjusted to an average of 4.5 hidden layers on categories 1 and 2 models, and 7.4 on categories 3 and 4. The greater depth on the later categories is related to the hidden layers width strategies. As the ANNs were smaller in these categories, more layers were required in order to increase the ANN capacity. As for input variables lag, categories 1 and 2 were kept with an average of 51.4 weeks of lag, while in categories 3 and 4 it was increased to 7.2 average. Table 6 lists the adjusted models, including selected variables, number of hidden layers and the average input lag.

Table 6 - Selected input variables for each created model.

C (1)	S (2)	D (3)	L (4)	Selected variables
1	3	5	51	MAX_T, NA_1.2, NA_3.4, NI_1.2, NI_3, NI_3.4, W
1	22	7	52	MAX_T, NI_1.2, NI_3, NI_3.4, NI_4, W
1	25	2	52	NA_3.4, NA_4, NI_1.2, NI_3, NI_3.4, SOI
1	27	6	52	MIN_T, NI_1.2, W
1	31	8	52	NI_1.2, NI_3
1	36	5	51	MAX_T, NI_1.2, NI_3, NI_3.4, NI_4, W
2	3	4	52	MEAN_T, NI_1.2, SOI, W
2	22	4	51	MAX_T, MEAN_T, NI_1.2, W, NI_3, SOI
2	25	3	51	NI_1.2, W, NI\3, TAR
2	27	3	52	HU, MAX_T, NA_1.2, NI_1.2, SUN, TAR, W
2	31	4	49	NA_1.2, NA_3, NI_1.2, NI_3, NI_3.4, W
2	36	3	51	NI_1.2, NI_3, NI_3.4, NI_4, SUN, W
3	3	11	20	SUN, W
3	22	12	7	NA_4, NI_3.4, NI_4, P, SUN, TAR, W
3	25	5	6	SUN, W
3	27	5	6	SUN, W
3	31	5	6	NI_3.4, SUN, TAR, W
3	36	5	6	SUN, W
4	3	7	4	MEAN_T, NA_3, NA_3.4, NI_1.2, NI_3, NI_3.4, NI_4, SUN, W
4	22	6	5	NA_1.2, NA_3, NA_3.4, NA_4, NI_1.2, NI_3, NI_4, SOI, SUN, TAR, W
4	25	4	7	NA_1.2, NA_3, NI_1.2, NI_3, TAR, W
4	27	6	6	MIN_T, NA_1.2, NA_4, NI_1.2, SOI, SUN, TAR, W
4	31	12	8	HU, MEAN_T, NA_1.2, NI_1.2, NI_3, NI_3.4, SUN, TAR, W
4	36	11	6	HU, NA_1.2, NI_1.2, NI_3, SUN, TAR, W

(1) Model category; (2) station number; (3) ANN hidden layers; (4) average input lag.

The MLP optimization step was able to further reduce RMSE by approximately 3.31% for the training dataset and by 1.56% for the test dataset on top of the previous step score, reaching an average RMSE of 31.35 mm for the training dataset and 33.38 mm for the test dataset. The optimal number of training epochs varied around 1717 epochs. As for activation function, ReLu was kept in 11 models while Leaky ReLu performed better in the remaining 13. Considering the test dataset RMSE by model category, the most accurate was category 3 (32.53 mm), followed by category 4 (32.66 mm), category 1 (33.25 mm) and category 2 (35.07 mm). By station the most accurate was station 27 (30.55 mm), then station 22 (32.97 mm), station 31 (33.18 mm), station 3 (33.28 mm), station 25 (34.11 mm) and station 36 (36.20 mm). Table 7 shows the optimization results for all created models using the test dataset.

When analyzing each model separately, considering the test dataset, the most accurate was in category 3, station 27, with 28.50 mm RMSE, while the worst was in category 2, station 36, with 37.15 mm RMSE. Although station 27 had the largest dataset (Table 3), it was not possible to find a strong correlation between the size of the dataset and better model score. In this regard it is also important to assess the quality of data, considering not only the error introduced by gap filling methods, but also possible errors in the raw data.

Table 7 - MLP optimization step results on all models for the test dataset.

Model Category	Station	Activation Function	Training Epochs	RMSE (mm)	MAE (mm)
1	3	ReLu	700	33.421	-1.880
1	22	Leaky ReLu	300	33.191	8.671
1	25	Leaky ReLu	1500	35.261	-3.907
1	27	ReLu	500	29.424	-1.030
1	31	Leaky ReLu	700	32.466	1.118
1	36	ReLu	700	35.758	-0.064
2	3	Leaky ReLu	1100	33.524	-2.534
2	22	ReLu	300	34.548	7.469
2	25	Leaky ReLu	500	35.629	3.582
2	27	Leaky ReLu	300	35.281	8.123
2	31	Leaky ReLu	700	34.312	1.439
2	36	ReLu	700	37.152	-2.974
3	3	ReLu	1500	32.686	0.496
3	22	Leaky ReLu	4100	32.026	3.047
3	25	ReLu	3500	32.739	4.791
3	27	ReLu	2300	28.502	0.821
3	31	ReLu	4500	33.342	-0.242
3	36	Leaky ReLu	3900	35.895	2.744
4	3	Leaky ReLu	1700	33.488	-3.284
4	22	ReLu	3100	32.096	3.809
4	25	Leaky ReLu	2900	32.825	3.017
4	27	Leaky ReLu	1300	28.975	-0.341
4	31	ReLu	1500	32.591	-0.070
4	36	Leaky ReLu	2900	35.990	0.113

When comparing error for each forecasting interval, a slight downward trend was identified. As illustrated in Figure 4, the error decreased as forecasting interval increased. This may indicate different dynamics for each interval, thus creating separate models for each interval may improve results. Lastly, Figure 5 illustrates all test dataset forecasts of 1 and 24 weeks ahead with the best (category 3, station 27) and worst (category 2, station 36) models. As the illustration shows, both models, in both forecast intervals, were able to reasonably indicate the precipitation seasons. However, they were unable to follow the weekly variations, especially the higher peaks in the time series.

Figure 4 - Mean error by forecasting interval.

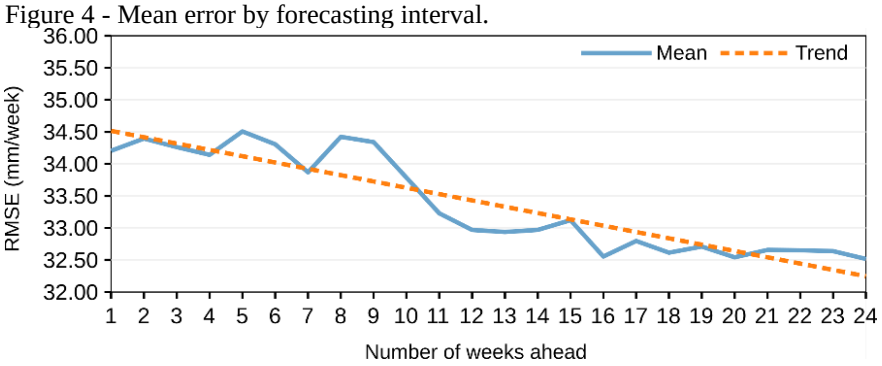
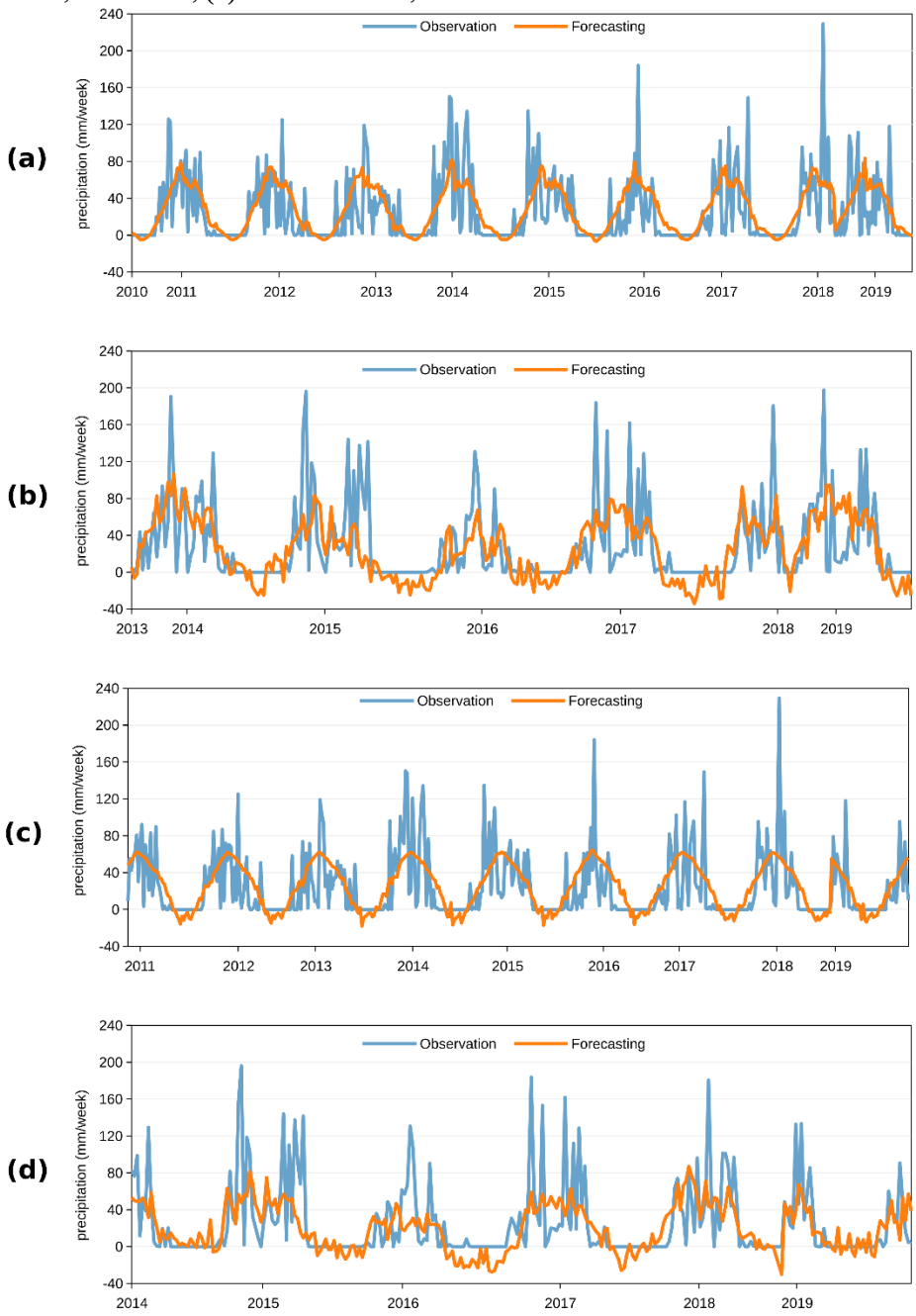


Figure 5 - Test dataset forecasts: (a) best model, 1 week ahead; (b) worst model, 1 week ahead; (c) 24 weeks ahead, best model; (d) 24 weeks ahead, worst model.



Conclusions

This work presented the creation of weekly precipitation forecasting models for locations in the state of Tocantins, Brazil, using multilayer perceptron networks (MLP) and public climatic data. Gaps found in time series data were filled using the regional weighting method. The model creation started by ranking the weather and climatic variables by their forecasting capability, which was measured using the MLPs themselves. Based on these rankings, the model input variables were selected using the forward selection algorithm. After that, two optimization steps were taken. The first optimized the lag of each input variable and the ANN depth, leading to an average RMSE reduction of 3.77%. The second step optimized the number of training epochs and the node activation function, reducing RMSE by an additional 2.43%. The average RMSE of the final models was 31.35 mm for the training dataset and 33.38 mm

for the test dataset. Respectively, these values represent 9.83% and 10.46% of the maximum weekly precipitation found in the work dataset, which was of 319.1 mm.

The results suggest that the created models are capable of reasonably good weekly precipitation forecasts, which can provide valuable information for farming, water resources management, urban planning and other related activies. Although there is possibly room for model improvement. Evaluating other types of ANN may help to achieve greater accuracy, however, the quality of input data tends to be of great relevance in machine learning models. Therefore, a more detailed review of the methods used to fill data gaps may help to produce more accurate training datasets, leading to overall error reduction. Considering the error distribution for each forecasting interval, as illustrated in Figure 4, building separate models for each interval may help to reduce model complexity, leading to lower error. Similarly, creating separate models for each target month, may also help reduce model complexity.

References

ABBOT, J., MAROHASY, J. **Input selection and optimization for monthly rainfall forecasting in Queensland, Australia, using artificial neural networks**. Atmospheric Research, USA, v.138, p.166-178, 2014.

ABBOT, J., MAROHASY, J. **Skilful rainfall forecasts from artificial neural networks with long duration series and single-month optimization**. Atmospheric Research, USA, v. 197, p. 289-299, 2017.

AKRAMI, S.A., EL-SHAFIE, A., JAAFAR, O. **Improving rainfall forecasting efficiency using modified adaptive neuro-fuzzy inference system (MANFIS)**. Water Resources Management, USA, v. 27, p. 3507-3523, 2013.

ANDERSON, W. *et al*. **Life cycles of agriculturally relevant ENSO teleconnections in north and south America**. International Journal of Climatology, USA, v. 37, p. 3297-3318, 2017.

ASSENG, S. *et al*. **Is a 10-day rainfall forecast of value in dry-land wheat cropping?** Agricultural and Forest Meteorology, USA, v. 216, p. 170-176, 2016.

BIER, A.A., FERRAZ, S.E.T. **Comparação de metodologias de preenchimento de falhas em dados meteorológicos para estação no sul do Brasil**. Revista Brasileira de Meteorologia, São Paulo, v. 32, n. 2, p. 215- 226, 2017.

CAO, J. *et al*. **Irrigation scheduling of paddy rice using short-term weather forecast data.** Agricultural Water Management, USA, v. 213, p. 714-723, 2019.

DUTTA, D. *et al*. **An artificial neural network based approach for estimation of rain intensity from spectral moments of a doppler weather radar**. Advances in Space Research, USA, v. 47, n. 11, p. 1949-1957, 2011.

ESTEVES, J.T., DE SOUZA ROLIM, G., FERRAUDO, A.S. **Rainfall prediction methodology with binary multilayer perceptron neural networks**. Climate Dynamics, USA, v. 52, p. 2319-2331, 2019.

INSTITUTO NACIONAL DE METEOROLOGIA, 2011. Nota Técnica No. 001/2011/SEGER/LAIME/CSC/INMET: Rede de Estações Meteorológicas Automáticas do

INMET. Disponível em: http://www.inmet.gov.br/portal/css/content/topo_iframe/pdf/Nota_Tecnica-Rede_estacoes_INMET.pdf. Acesso em: jun. 2018.

LEE, J. *et al.* **Application of artificial neural networks to rainfall forecasting in the geum river basin, Korea**, Water, Switzerland, v. 10, n. 10, p. 1-14, 2018.

MAROHASY, J., ABBOT, J. **Assessing the quality of eight different maximum temperature time series as inputs when using artificial neural networks to forecast monthly rainfall at cape Otway, Australia**. Atmospheric Research, USA, v. 166, p. 141-149, 2015.

MAY, R., DANDY, G., MAIER, H. Review of input variable selection methods for artificial neural networks, *In*: SUZUKI, K. **Artificial Neural Networks**. 1. ed. IntechOpen, Croácia, 2011. p. 19-44.

DE OLIVEIRA, L.F.C. *et al.* **Comparação de metodologias de preenchimento de falhas de séries históricas de precipitação pluvial anual**. Revista Brasileira de Engenharia Agrícola e Ambiental, Paraíba, v.14, n.11, p.1186-1192, 2010.

CARDOSO, A. de O. *et al.* **Extended time weather forecasts contribute to agricultural productivity estimates**. Theoretical and Applied Climatology, USA, v. 102, p. 343-350, 2010.

SCAIFE, A. *et al.* **What is the el niño-southern oscillation?** Weather, USA, v. 74, n. 7, p. 250-251, 2019.

SECRETARIA DO PLANEJAMENTO E ORÇAMENTO DO ESTADO DO TOCANTINS, 2016. Perfil do Agronegócio Tocantinense. Disponível em: https://central3.to.gov.br/arquivo/354694/. Acesso em: jun. 2018.

TABONY, R.C. **The estimation of missing climatological data**. Journal of Climatology, USA, v. 3, n. 3, p. 297-314, 1983.

VELO, R., LÓPEZ, P., MASEDA, F. **Wind speed estimation using multilayer perceptron**. Energy Conversion and Management, USA, v. 81, p. 1-9, 2014.