Development of machine learning models for air temperature estimation using MODIS data

Ovando, G., Sayago S. and Bocco, M.

DOI: 10.31047/1668.298x.v39.n1.33225

SUMMARY

Air temperature is a key variable in a wide range of environmental applications, including land–atmosphere interaction, climate change research and hydrology and crop growth models, among others. The objective of this study was to estimate daily air maximum (Tmax) and minimum (Tmin) temperatures, based on MODIS AQUA/TERRA land surface temperature (LST), NDVI, extraterrestrial solar radiation and precipitation data. Artificial neural networks (ANN) and random forests (RF) models were developed to predict these temperatures covering weather stations in Córdoba (Argentina) for 2018-2020. The results show that RF and ANN machine learning algorithms are capable of modeling non-linear relationships between registered temperatures and LST MODIS data, in a very robust way. The validation of the models confirms that Tmax and Tmin can be accurately estimated using, jointly or separately, AQUA and TERRA LST. The best models present determination coefficients equal to 0.81/0.91 and root mean square error of 2.7/2.1 ºC for Tmax/Tmin, when using AQUA LST day/night satellite overpass time data, respectively. The robustness and confidence of the models developed, and the ease and free accessibility of input data at a global scale, suggest that these methodologies have the potential to be applied to other regions.

Keywords: random forest, artificial neural networks, maximum/minimum air temperature, land surface temperature, AQUA/TERRA satellite.

Ovando, G., Sayago S. and Bocco, M. (2022). Desarrollo de modelos de aprendizaje automático para estimar temperatura del aire utilizando datos MODIS. *Agriscientia 39*: 15-28

RESUMEN

La temperatura del aire es una variable clave en una amplia gama de aplicaciones ambientales, que incluyen interacción tierra-atmósfera, cambio climático, modelos de cultivos e hidrológicos, entre otros. El objetivo de este estudio fue estimar las temperaturas máxima del aire (Tmax) y mínima diaria (Tmin), con datos de temperatura de la superficie terrestre (LST) de MODIS AQUA/TERRA, NDVI, radiación solar extraterrestre y precipitación.

Fecha de recepción: 01/06/2021; fecha de aceptación: 03/05/2022.

Se desarrollaron modelos de redes neuronales artificiales (ANN) y bosques aleatorios (RF) para predecir estas temperaturas considerando estaciones meteorológicas de Córdoba (Argentina) para el período 2018-2020. Los resultados muestran que las metodologías de RF y ANN fueron capaces de modelar relaciones no lineales entre la temperatura registrada y los datos de LST de MODIS, de manera muy robusta. La validación de los modelos confirma que Tmax y Tmin se pueden estimar con precisión utilizando, en conjunto o por separado, AQUA y TERRA LST. Los mejores modelos presentaron coeficientes de determinación iguales a 0,81/0,91 y error cuadrático medio de 2,7/2,1 ºC para Tmax/Tmin, cuando se utilizaron datos de AQUA correspondientes a día/noche, respectivamente. La solidez y el ajuste de los modelos desarrollados, sumado a la libre accesibilidad de datos a escala global, sugieren que esta metodología puede ser aplicada a otras regiones.

Palabras clave: bosques aleatorios, redes neuronales artificiales, temperatura máxima/mínima, temperatura de la superficie terrestre, satélites AQUA/TERRA.

Ovando, G. (ORCID: 0000-0002-6015-404X), Sayago, S. (ORCID: 0000-0003-4723- 6670) and Bocco, M. (ORCID: 0000-0002-0359-7530). Universidad Nacional de Córdoba, Facultad de Ciencias Agropecuarias. Córdoba, Argentina.

Correspondencia a: mbocco@gmail.com

INTRODUCTION

Air temperature (Ta) is a key variable in a wide range of environmental applications, including land–atmosphere interaction, climate change research, and hydrology and crop growth models, among others. Air temperature is measured at specific locations, using thermometer shelters (2 m above the ground) at meteorological ground stations. Generally, weather stations have sparse distribution and they usually present missing data. For use in environmental applications, interpolation techniques were traditionally employed to obtain spatial patterns of air temperature. However, they can present significant errors, particularly in complex landscapes (Oyler et al., 2016; Yang et al., 2017).

On the other hand, land surface temperature (LST) data are actually available from different satellite missions that carry sensors such as Moderate Resolution Imaging Spectroradiometer (MODIS), the Advanced Very High Resolution Radiometer (AVHRR), and the Advanced Along Track Scanning Radiometer (AATSR), among others. These data are available at high temporal resolution over extended regions. (Alfieri et al., 2012; Long et al., 2020).

LST is governed by land–atmosphere interactions that involve the down/upward radiation, latent and sensible heat loss fluxes. LST is a very sensitive parameter to describe the characteristics of surface energy balance, surface thermal inertia, and surface water-heat budget. Several features influence the accuracy of LST retrievals like sensor characteristics, atmospheric conditions, variations in spectral emissivity, surface type heterogeneity, soil moisture, visualization geometry, and assumptions related to the split window method (Benali et al., 2012; Chang et al., 2020).

Recent studies used LST to assess crop water stress (Sayago et al., 2017), characterize urban heat islands (Sobrino and Irakulis, 2020), evaluate joint trends with vegetation index data (Heck et al., 2019), and to estimate near-surface air temperature (Oyler et al., 2016; Yang et al., 2017; Chang et al., 2020).

Although LST and Ta are strongly correlated, they show differences due to several factors. One of them is the high background temperature that rises in regions with a mixture of vegetation and bare soil. Also, surface temperature depends not only on the evapotranspiration rate, but also on varying environmental factors, including the incoming solar radiation and the wind speed (Deery et al., 2014).

There are different types of methods commonly used to estimate Ta based on LST:

i) Statistical approaches, based on regression techniques which can be simpler or more advanced.

ii) Machine learning techniques, which take into account the non-linearity between predictors and LST. The latter can incorporate many explanatory variables in the physical-deterministic modeling.

These techniques include Bayesian-based modeling, support vector regression, ANN and RF (Kamoutsis et al., 2013, Noi et al., 2017, Zhao et al., 2019).

iii) The temperature–vegetation index, which is based on the assumption that for an infinitely thick canopy, the top-of-canopy temperature is the same as within the canopy.

iv) Physically based energy-balance approaches, whose major problem is that they require large amounts of information, among others (Benali et al., 2012; Bartkowiak et al., 2019).

Córdoba province is located in the middle of Argentina and it has one of the highest potentials for crop and livestock production in Argentina. Nowadays it is mainly devoted to soybean and corn production, which is exported as grain, oil, animal feed, and biodiesel or bioethanol. Total cash crop cultivated surface increased over the last two decades, displacing livestock or other traditional crops, and by incorporating new land through deforestation (Wehbe et al., 2018).

The main objective of this study was to estimate maximum (Tmax) and minimum (Tmin) air temperatures based on LST data of MODIS, using ANN and RF. The specific objectives of this study were: i) to explore the relationship between Tmax/ Tmin and NDVI, extraterrestrial solar radiation, precipitation and LST;ii) to develop models to predict these temperatures covering weather stations in Córdoba departments (Argentina); and iii) to assess performance and accuracy of machine learning models considering day and night LST from both the TERRA (MOD) and AQUA (MYD) satellites.

MATERIALS AND METHODS

Study area and meteorological data

The study area is located in the province of Córdoba (Argentina) and includes 26 departments, whose limits are -29.46° to -35° S; -61.7° to -65.8° W, approximately. This province has different environments, such as Mar Chiquita lagoon in the northeast depression, the Salinas Grandes in the northwest, and the pampas plain in the east and southeast. The three hill ranges, separated by valleys, are located to the west, running in almost a north-south direction (Figure 1). The altitude ranges between 69 masl and 2790 masl, with an average of 600 masl.

The climate in the province of Córdoba is temperate characterized as humid to sub-humid (Wehbe et al., 2018), with a good differentiation of the four seasons. The thermal regime is determined by the temperatures of the warmest month (January) and the coldest month (July). The annual average temperature is between 16 °C and 18 °C with average maximum and minimum of 30 °C and 10 °C, respectively. The frost-free period is approximately 300 days. Precipitation shows a decreasing gradient in an east-west direction, reaching annual mean values between 600 and 900 mm, concentrating 80 % of them in the spring-summer semester (Aliaga, 2017; Infraestructura de Datos Espaciales de la Provincia de Córdoba [IDECOR], 2019).

The air temperature data were provided by Ministerio de Agricultura y Ganadería of Córdoba province (MAG, 2021).This institution is responsible for the maintenance of the stations, control of the parameters and calibration of the sensors to ensure quality control.

The records correspond to 24 meteorological ground stations, one for each department in the study area. Only two departments of Córdoba, Sobremonte and Tulumba, do not have official meteorological stations (Figure 1 and Table 1). The measurements included daily Tmin and Tmax. The time period selected for our study ranged from September 2018 to August 2020.

MODIS LST Products

Daily land surface temperature values, with spatial resolution of 1 km, were obtained using MODIS LST V6 products from TERRA and AQUA satellites (MOD11A1 and MYD11A1, respectively). Both satellites have sun-synchronous polar orbits. MODIS TERRA data are available during 9:10–11:00 a.m. and 10:00–11:50 p.m. (day/night) local time, while MODIS AQUA sensor collects the imagery during 1:00–2:50 p.m. and 0:00–2:10 a.m. (day /night). At each pixel corresponding to the ground stations, we extracted the day and night surface temperatures using Google Earth Engine (Gorelick et al., 2017)

The number of data corresponding to day and night MODIS LST products are presented in Table 2, after removing outlier data and unavailable values for each station (when clouds were present).

Auxiliary data

MODIS NDVI Products

Normalized difference vegetation index (NDVI) is the most common remote sensing index used

Table 1. Geographical location for Córdoba departmental stations used in this study

ID	Department	Station name	Latitude (°)	Longitude (°)	Elevation (m)
1	Río Seco	La Rinconada APRHi.*	-30.18	-62.95	75
2	Ischilín	Deán Funes Agr.*	-30.40	-64.35	711
3	Totoral	Villa del Totoral Agr.	-30.71	-64.08	564
4	Cruz del Eje	La Candelaria APRHi.	-30.94	-64.84	887
5	Colón	Colonia Caroya Agr.	-31.02	-64.04	498
6	Minas	San Carlos Minas Agr.	-31.17	-65.11	778
7	Punilla	Cosquín Agr.	-31.21	-64.46	786
8	Río Primero	Río Primero Agr.	-31.32	-63.64	274
9	Pocho	Salsacate APRHi.	-31.34	-65.08	998
10	San Justo	El Tio Agr.	-31.38	-62.81	125
11	Capital	Lab. Hidraulica APRHi.	-31.44	-64.19	475
12	Santa María	Cno 60 Cuadras Agr.	-31.54	-64.13	461
13	Río Segundo	Las Junturas Agr.	-31.83	-63.45	239
14	San Alberto	San Pedro Agr.	-31.93	-65.22	563
15	San Javier	Villa Dolores APRHi.	-31.95	-65.19	584
16	Tercero Arriba	Colonia Almada Agr.	-32.04	-63.87	361
17	Calamuchita	Villa del Dique Agr.	-32.17	-64.45	554
18	Unión	San Antonio de Litín Agr.	-32.21	-62.65	146
19	Gral. San Martín	Ausonia Agr.	-32.65	-63.25	215
20	Marcos Juárez	Marcos Juárez Agr.	-32.72	-62.09	113
21	Río Cuarto	Rio Cuarto Agr.	-33.16	-64.36	451
22	Juárez Celman	Alejandro Roca Agr.	-33.35	-63.70	207
23	Pte. Roque Sáenz Peña	Villa Rossi Agr.	-34.24	-63.27	171
24	Gral. Roca	Buchardo Agr.	-34.71	-63.50	134

* Agr.: Ministerio de Agricultura Córdoba; APRHi.: Administración Provincial de Recursos Hídricos.

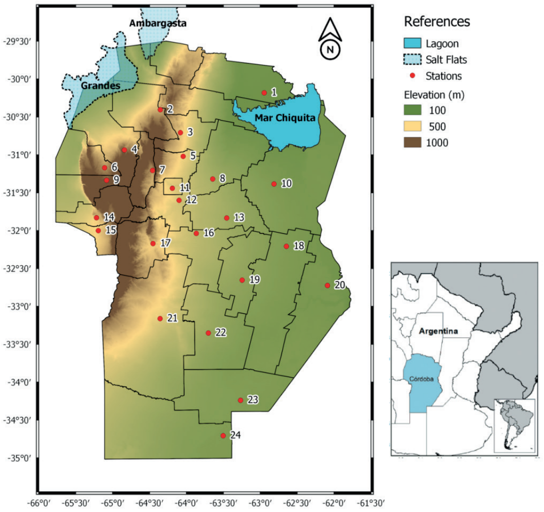


Figure 1. Geographical location of Córdoba departmental stations included in this study (the numbers correspond to the ID of Table 1)

to parameterize vegetation status. Land surface temperature is influenced by vegetation because it selectively reflects and absorbs radiation from the sun and modifies latent and sensible heat exchange. Vegetation abundance reduces LST through latent heat transfer from the surface to atmosphere via the process of evapotranspiration (Alademomi et al., 2020). The combination of NDVI and LST provides information about the condition of the vegetation and surface soil moisture content and therefore successfully monitors the water stress of vegetation (Sayago et al., 2017).

In this paper NDVI values, with spatial resolution of 1 km, were obtained using MODIS 13A2 V6 products from TERRA and AQUA satellites. MOD13A2 and MYD13A2 products generate a single NDVI value to represent the composite period of 16 days. This NDVI is considered as the continuity index to the existing National Oceanic and Atmospheric Administration-Advanced Very High Resolution Radiometer (NOAA-AVHRR). The dates contained in the composite day of the year layer were also considered in order to mix MOD and MYD products and finally interpolate, by cubic splines (SRS1 - Cubic Spline for Excel V2.5), to obtain NDVI daily values.

IMERG - F Product

As surface temperature depends on the

Table 2. Number of valid MODIS LST data (day and night) for each station included in this study

ID	Total	MODIS TERRA		MODIS AQUA	
		Day	Night	Day	Night
1	1449	330	386	384	349
2	1408	333	350	373	352
3	1493	352	380	392	369
4	1520	378	368	402	372
5	1508	352	397	377	382
6	1362	306	375	338	343
7	1392	373	330	358	331
8	1321	296	349	338	338
9	1616	423	377	451	365
10	1376	282	385	345	364
11	1446	344	366	369	367
12	1491	374	364	394	359
13	1364	300	367	345	352
14	1585	402	401	426	356
15	1516	393	386	375	362
16	1360	335	334	366	325
17	1753	433	441	447	432
18	1421	308	395	349	369
19	1392	293	382	345	372
20	1409	315	387	352	355
21	1456	350	382	356	368
22	1436	315	391	355	375
23	1637	397	427	409	404
24	1690	407	434	434	415

evapotranspiration rate, it is important to consider the water availability in the field. In this sense we used precipitation estimation from the Integrated Multi-Satellite Retrievals for GPM (IMERG) products. We used the IMERG Final product, because it is calibrated with the monthly rain gauge information, so the estimates are the most accurate and reliable for research (Huang et al., 2018). IMERG version 5 level 3 products were used in this study. The level 3 products include gridded rainfall and snowfall data, with 0.1°× 0.1° spatial resolution. The IMERG information corresponding to the pixel of each meteorological station was acquired and accumulated in 10 day values (PPacc).

Extraterrestrial solar radiation

In addition to MODIS (LST, NDVI) and IMERG (PPacc) products, extraterrestrial solar radiation (Qa) was used; this auxiliary variable either has an impact on air temperature and LST or influences the relationship between them.

$$Q_a = \frac{1440}{\pi} G_{sc}.d_r \, [\omega_s \sin(\varphi) \sin(\delta) + \cos(\varphi) \cos(\delta) \sin(\omega_s)]$$

$$d_r = 1+0.033 \cos \left(\frac{2\pi}{365} J\right)$$

$$\delta = 0.409 \sin \left(\frac{2\pi}{365} J - 1.39\right)$$

$$\omega_s = \text{arc cos} \, (-\tan(\varphi)\tan(\delta))$$

where Q_a= extraterrestrial radiation [MJ m^{-2} day^{-1}], d_r= inverse relative distance Earth-Sun, ω_s= sunset hour angle [rad], G_{sc}= solar constant = 0.082 MJ m^{-2} min^{-1}, φ= latitude [rad], δ = solar declination [rad] and J = Day of year [1 to 366].

Models

Random Forests

Random Forests (RF) was proposed by Breiman (2001). It is a machine learning technique exploiting statistical nonlinear relationship between variables. RF is a nonparametric and ensemble technique. These models are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The final predicted values are produced by the mean of the results of all the individual trees that make up the forest. The dataset required for RF consists of observations, including predictors, and dependent variables.

In this paper, 75 % of data were used for training and the 25 % remaining for validation purposes, both sets were randomly generated. Figure 2 shows a random forest tree corresponding to the training stage.

Artificial Neural Networks

Artificial Neural Networks models are capable of capturing the most complex relationships among the data and extracting subtle patterns, which is not always possible with multiple linear regression or other deterministic methods. The mathematical relationships describing the modeled process, by ANN, do not need to be known because the network 'learns' these relationships through the association of patterns between independent and dependent variables. In particular, an important feature is that this non-parametric model does not require prior assumption about the statistical behavior (normally distributed data) or about any specific relationship between variables.

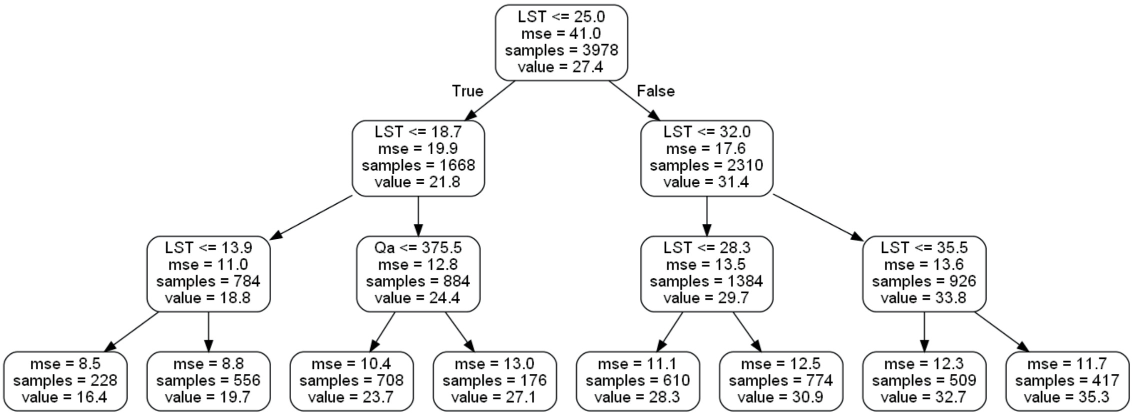


Figure 2. Training stage in the random forest algorithm

ANN are a structure of neurons joined by nodes that transmit information from one neuron to another, which yields a result by means of mathematical functions. The ANN learns from the existing information through a training process by which their parameters (weights) are adjusted, so as to provide an approximate output close to that desired. In this study, multilayer perceptron networks (Figure 3) were designed, including four neuron layers – input (I), two hidden (H), and output (O) layers. These were trained by the back-propagation algorithm to minimize quadratic error. This process was repeated 1000 times. As in RF models, the 75 % of data were used for training and the 25 % remaining for the validation process, both sets were randomly generated.

The general steps that describe the training algorithm were executed according to Sayago and Bocco (2018).

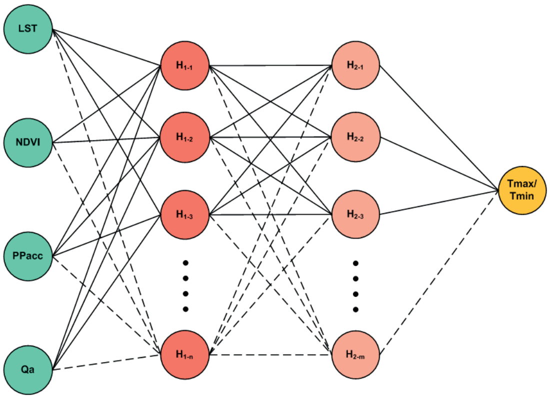


Figure 3. Schema of artificial neural network architecture

Metrics for accuracy assessment

To assess RF and ANN modeling performance, we calculated correlation coefficient *(r)*, determination coefficient (R^2), Root Mean Square Error (RMSE) and bias between registered data and estimated ones from MODIS LST plus auxiliary variables.

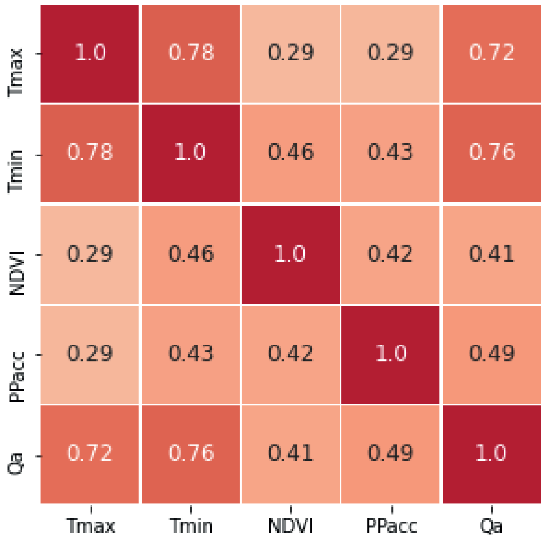
In order to evaluate the model performance, we present a Taylor diagram (Taylor, 2001). This graph shows standard deviation relative to observed standard deviation (RSD), *r* coefficient and the centered root mean square error (RMSEc), which were defined in Kärnä and Baptista (2016).

RESULTS AND DISCUSSION

In order to evaluate the auxiliary variables for Tmax and Tmin estimation, a heat map with the correlations between NDVI, PPacc and Qa, taken in pairs, was built (Figure 4). These predictor variables present only one register by day. The heat map transforms the correlation matrix into color coding.

For all valid data corresponding to MODIS LST and station registers, Figure 4 shows that Qa has the best coefficient of correlation with Tmax and Tmin. NDVI and PPacc are better correlated with Tmin, although both values are low.

The inclusion of NDVI did not substantially improve prediction (r=0.29 and 0.46 for Tmax and Tmin, respectively); this lower correlation was observed in Benali et al. (2012). These authors indicated several reasons that may be behind this: soil moisture and transpiration, land cover structure and the low representation of the station's surrounding environment, among others.



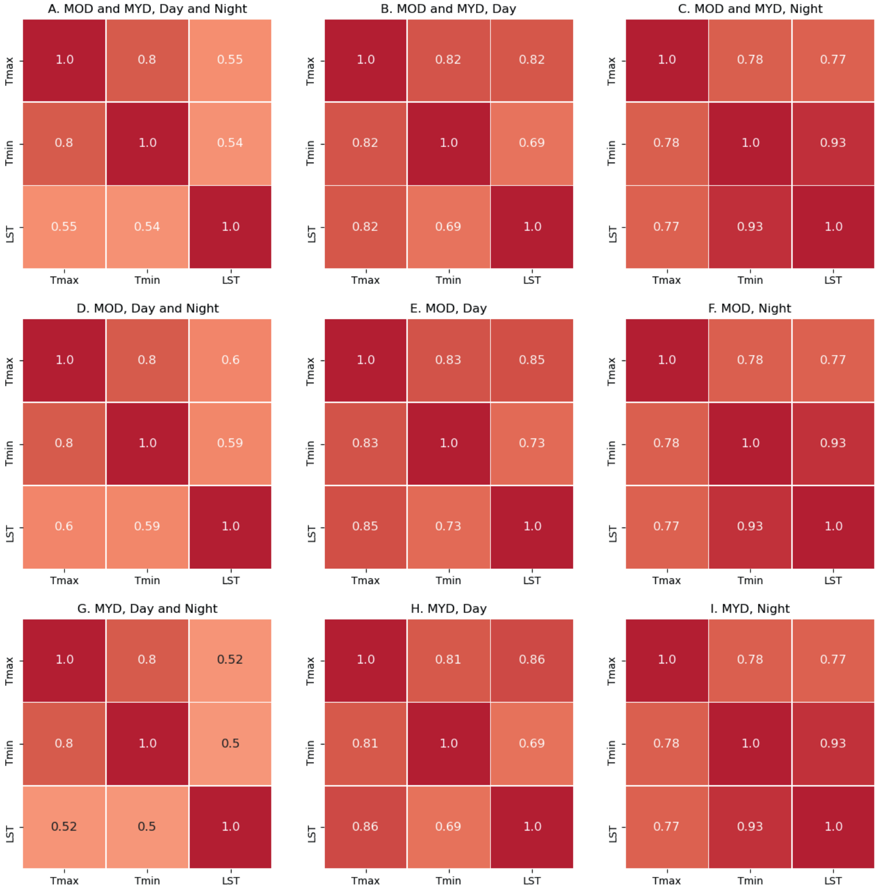
	Tmax	Tmin	NDVI	PPacc	Qa
Tmax	1.0	0.78	0.29	0.29	0.72
Tmin	0.78	1.0	0.46	0.43	0.76
NDVI	0.29	0.46	1.0	0.42	0.41
PPacc	0.29	0.43	0.42	1.0	0.49
Qa	0.72	0.76	0.41	0.49	1.0

Figure 4. Correlation matrix heat map for the Tmax, Tmin, and auxiliary variables considered in this study

In this study, daily Tmax and Tmin had low positive correlation values with PPacc (0.29 and 0.43, respectively). However, Trenberth and Shea (2005) found that negative correlations dominate globally over land; these authors explain that dry conditions favor more sunshine and less evaporative cooling.

The correlation values of Qa with Tmax and Tmin were 0.72 and 0.76, respectively. Janatian et al. (2017) correlated Qa with weekly maximum air temperature and obtained an *r* value of 0.83. Emamifar et al. (2013) used as input variables day and night TERRA-MODIS LST, extraterrestrial solar radiation and Julian day for a model tree to estimate daily mean air temperature. The extraterrestrial solar radiation and Julian day were used to reflect diurnal difference and seasonal variation in variation of surface air temperature, respectively.

Figure 5 (A, B and C) shows that, when MODIS LST products data are analyzed together, Tmax correlates better with LST values corresponding to



A. MOD and MYD, Day and Night

	Tmax	Tmin	LST
Tmax	1.0	0.8	0.55
Tmin	0.8	1.0	0.54
LST	0.55	0.54	1.0

B. MOD and MYD, Day

	Tmax	Tmin	LST
Tmax	1.0	0.82	0.82
Tmin	0.82	1.0	0.69
LST	0.82	0.69	1.0

C. MOD and MYD, Night

	Tmax	Tmin	LST
Tmax	1.0	0.78	0.77
Tmin	0.78	1.0	0.93
LST	0.77	0.93	1.0

D. MOD, Day and Night

	Tmax	Tmin	LST
Tmax	1.0	0.8	0.6
Tmin	0.8	1.0	0.59
LST	0.6	0.59	1.0

E. MOD, Day

	Tmax	Tmin	LST
Tmax	1.0	0.83	0.85
Tmin	0.83	1.0	0.73
LST	0.85	0.73	1.0

F. MOD, Night

	Tmax	Tmin	LST
Tmax	1.0	0.78	0.77
Tmin	0.78	1.0	0.93
LST	0.77	0.93	1.0

G. MYD, Day and Night

	Tmax	Tmin	LST
Tmax	1.0	0.8	0.52
Tmin	0.8	1.0	0.5
LST	0.52	0.5	1.0

H. MYD, Day

	Tmax	Tmin	LST
Tmax	1.0	0.81	0.86
Tmin	0.81	1.0	0.69
LST	0.86	0.69	1.0

I. MYD, Night

	Tmax	Tmin	LST
Tmax	1.0	0.78	0.77
Tmin	0.78	1.0	0.93
LST	0.77	0.93	1.0

Figure 5. Correlation matrix heat map for the Tmax, Tmin, and LST, considering dayand night of TERRA and AQUA satellites

day and Tmin with LST night values (*r* coefficients equal to 0.82 and 0.93, respectively).

LST night and Tmin show identical correlation (*r*=0.93), independently of the satellite (Figure 5, F and I) considered, while the relationships between Tmax versus LST day are similar, although slightly lower (correlation coefficients equal to 0.85 for MOD data and 0.86 for MYD data, Figure 5, E and H, respectively). Janatian et al. (2017) found correlation values of 0.85 and 0.90 between weekly maximum air temperature and day and night MOD LST, respectively. In this work, using daily values, we found very similar correlation values for day MOD and MYD LST with Tmax, however for night MOD and MYD LST the correlation values were lower (0.77 in both cases). When LST day and night data are considered at the same time, the coefficients of correlation for Tmax or Tmin decrease to values lower or equal to 0.6 (Figure 5, A, D and G).

Values of correlation coefficients presented in this paper were comparable to those found by Emamifar et al. (2013), who, with data obtained from 17 meteorological ground stations distributed throughout Iran and using only MOD data for one year (2007), presented *r* values from 0.91 to 0.96, when they considered Tmax and day LST and between 0.91 and 0.97 for Tmin and night LST. In this paper the best behavior was observed for the relationship between Tmin and night LST for both satellites. Zeng et al. (2015) reported similar results for Tmin estimation (*r*=0.93 for MOD and *r*=0.95 for MYD), and the lowest correlation coefficients values were found for Tmax *vs.* day LST relationship (*r*=0.46, for both satellites).

Several random forest models were trained and tested, using satellite LST dataset and daily auxiliary variables to estimate Tmax and Tmin. The determination coefficient values (Table 3), agree with the correlation values obtained using only MODIS data with temperature (Figure 5). All models present statistical bias values approximately equal to 0 (between -0.1 to 0.1), which means that there are no significant over/under estimations.

The maximum temperature estimation, according to the R^2 and RMSE values (Table 3), is obtained with greater precision when using day MYD LST data than both satellites together or MOD data. Similarly, minimum temperature is accurately estimated with night LST data, considering one or both satellites, reaching the best statistics with MYD. Analyzing statistic values, the estimates of Tmin have the best R^2, however RMSE values, considering their relative importance (Tmin observed range is 9.6-10.1°C and 26.6-27.4°C for Tmax), indicate that Tmax estimates are better.

Table 3. Metric of RF models for Tmax and Tmin estimation (validation stage) using MOD and MYD MODIS day, and night LST and auxiliary variables

LST data - Estimated temperature	R^2	BIAS (°C)	RMSE (°C)
MOD and MYD: day and night - Tmax	0.79	0.0	3.0
MOD and MYD: day and night - Tmin	0.84	0.0	2.7
MOD and MYD: day - Tmax	0.78	0.0	3.0
MOD and MYD: day - Tmin	0.79	0.0	3.2
MOD and MYD: night - Tmax	0.78	0.0	3.1
MOD and MYD: night - Tmin	0.90	0.0	2.2
MOD: day and night- Tmax	0.72	0.0	3.4
MOD: day and night - Tmin	0.80	0.1	3.1
MYD: day and night - Tmax	0.76	0.0	3.2
MYD: day and night - Tmin	0.81	0.1	3.0
MOD: day - Tmax	0.78	0.1	3.0
MOD: day - Tmin	0.74	-0.1	3.4
MOD: night - Tmax	0.70	0.0	3.6
MOD: night - Tmin	0.89	0.1	2.3
MYD: day - Tmax	0.81	-0.1	2.7
MYD: day - Tmin	0.73	-0.1	3.4
MYD: night - Tmax	0.74	0.0	3.4
MYD: night - Tmin	0.91	0.0	2.1

Figure 6 presents scatter plots for Tmax and Tmin best estimation, using TERRA and AQUA MODIS data with RF algorithms, for the validation process. These figures (Figure 6 A to D) indicate a little difference when day or night MOD or MYD data were used; it can be observed a very small underestimation for Tmax when the temperature is over 35°C.

Figure 7 shows the importance ranking of the input variables calculated by the RF. This importance score gives a relative ranking regarding the contribution of the input variables, but it is not equivalent to the correlation coefficient. The contribution of LST is the most important to estimate Tmax and Tmin, using the four best RF models; its importance is greater when estimating Tmin (0.89-0.91) than for Tmax (0.77-0.79). MYD LST has little difference in importance when compared with MOD LST, for both Tmaxand Tmin. The second predictive variable for estimating Tmax is NDVI (Figure 7 A and C), instead this place is occupied by Qa for Tmin forecast (Figure 7 B and D).

ANN models were trained and tested, using LST dataset and daily auxiliary variables to estimate Tmax and Tmin. The determination coefficient, BIAS and RMSE values (Table 4), were similar to the values obtained using RF models (Table 3). All models presented bias values between -0.8 to 0.9°C and good RMSE, considering their values relative to temperature range, principally for

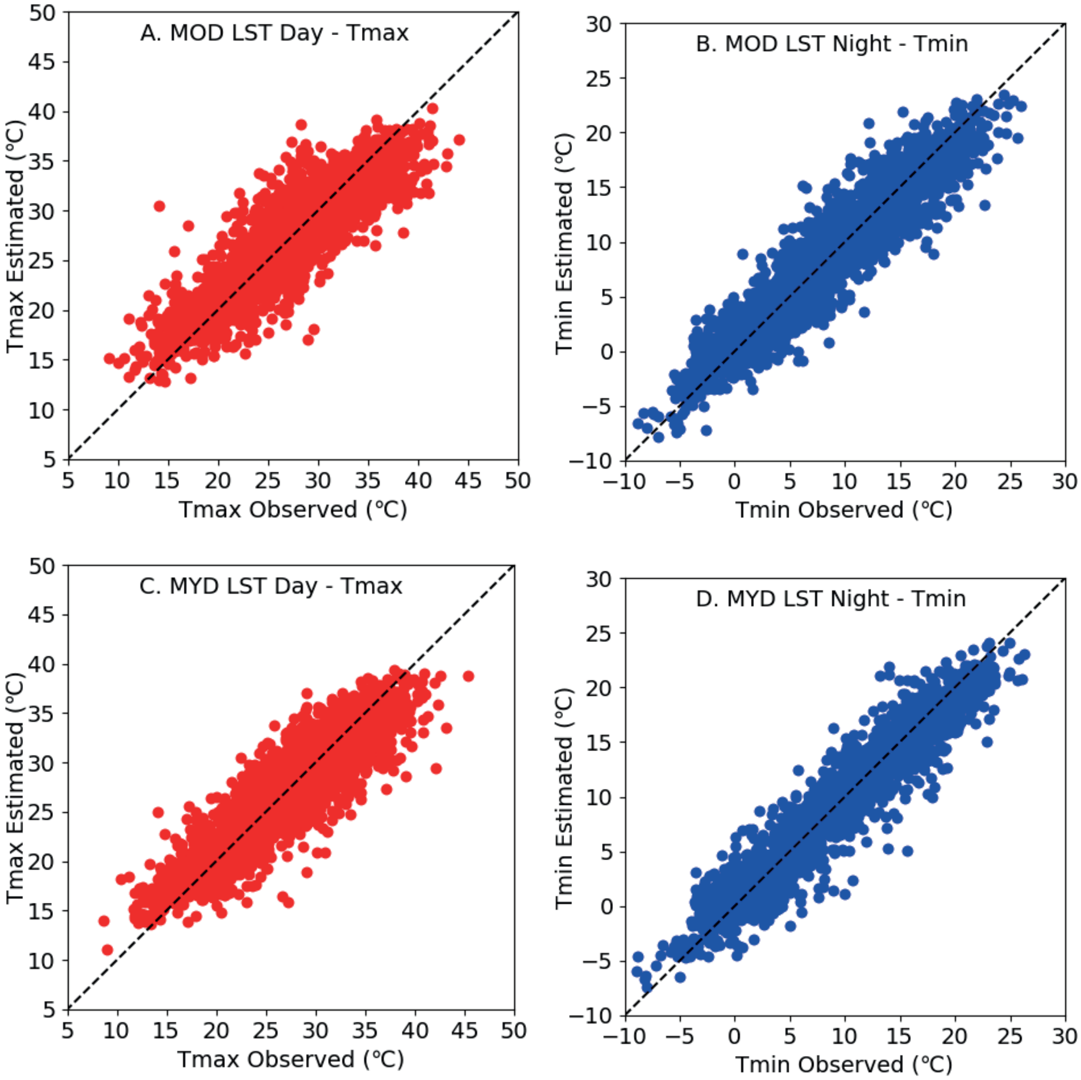


Figure 6. Scatter plots of the best RF models to estimate daily Tmax and Tmin temperature using TERRA and AQUA MODIS data, and auxiliary variables (validation stage); dashed lines represent 1:1 line

maximum temperature. For this variable root mean square error obtained was from 3.1°C to 4.2°C.

The ANN models with MYD day or night LST data presented the best behavior to estimate Tmax and Tmin respectively, as in RF. If MYD data are not available, due to cloudiness, the use of the MOD data is a satisfactory alternative, because their adjustment statistics also present good values.

Scatter plots for Tmax and Tmin best estimations using TERRA and AQUA MODIS data with ANN algorithms, for validation process, are presented in Figure 8. These figures (Figure 8 A to D) show, for day or night MODIS data, over and under estimation for small and elevated Tmax and Tmin values. As Sayago and Bocco (2018) observed, ANN overestimate low values and underestimate high values.

Figure 9 shows a Taylor diagram based on the r, RSD and RMSEc statistics, for the best RF and ANN models. This diagram provides a way of graphically summarizing how closely a set of patterns matches observations. Each point in the two-dimensional Taylor diagram represents, simultaneously, the three statistics for Tmax and Tmin (Taylor, 2001).

Taylor diagram (Figure 9) shows that all models for Tmin estimation are, in relative position, closer to the reference dot than the ones for Tmax. Considering RF and ANN models for Tmax or Tmin,

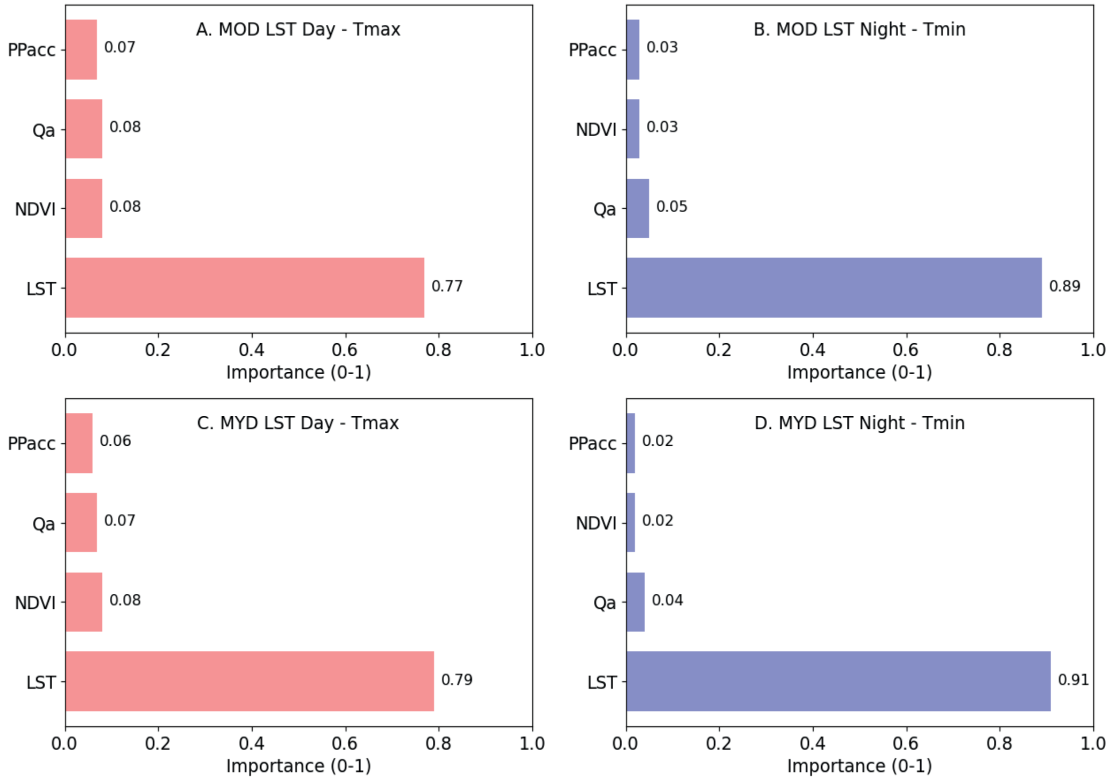


Figure 7. Relative importance of variables used by the best random forest models to estimating daily Tmax and Tmin temperature with TERRA and AQUA MODIS data

Table 4. Metric of ANN models for Tmax and Tmin estimation using MOD and MYD MODIS day, and night LST and auxiliary variables

LST data - Estimated temperature	R^2	BIAS (°C)	RMSE (°C)
MOD and MYD: day and night - Tmax	0.59	0.5	4.2
MOD and MYD: day and night - Tmin	0.70	-0.1	3.8
MOD and MYD: day - Tmax	0.71	0.7	3.5
MOD and MYD: day - Tmin	0.65	-0.2	3.9
MOD and MYD: night - Tmax	0.66	0.4	3.8
MOD and MYD: night - Tmin	0.88	-0.5	2.4
MOD: day and night- Tmax	0.60	-0.5	4.2
MOD: day and night - Tmin	0.73	0.2	3.5
MYD: day and night - Tmax	0.58	0.1	4.2
MYD: day and night - Tmin	0.75	0.6	3.4
MOD: day - Tmax	0.74	-0.8	3.3
MOD: day - Tmin	0.67	0.6	3.9
MOD: night - Tmax	0.64	0.2	3.9
MOD: night - Tmin	0.87	0.3	2.5
MYD: day - Tmax	0.77	0.9	3.1
MYD: day - Tmin	0.66	0.4	3.8
MYD: night - Tmax	0.66	-0.2	3.9
MYD: night - Tmin	0.91	0.0	2.1

separately, we observe that the statistics included in the Taylor diagram achieve close values. All models had an RMSEc, correlation coefficients and RSD near 0.3, 0.95 and 1.0 for Tmin and 0.45, 0.9 and 0.9 for Tmax, respectively.

Finally, it is important to discuss the statistical values obtained by applying different models developed in this paper, to estimate Tmax and Tmin, considering LST values provided by MODIS images and the auxiliary variables.

In this work the best fits were obtained using MYD data, these results are supported by Zeng et al. (2015) who expressed that the difference between satellite overpass had little impact on the temperature estimation accuracy. The AQUA LST images were also chosen by Yang et al. (2017) because their acquisition time is closer to the actual occurrence of observed maximum and minimum temperatures. Benali et al. (2012) also stated that the use of AQUA (day and night) could improve the estimation of Tmax and Tmin, respectively, due to the fact that overpass time of this satellite is closer to the occurrence time of Tmax and Tmin than TERRA´s.

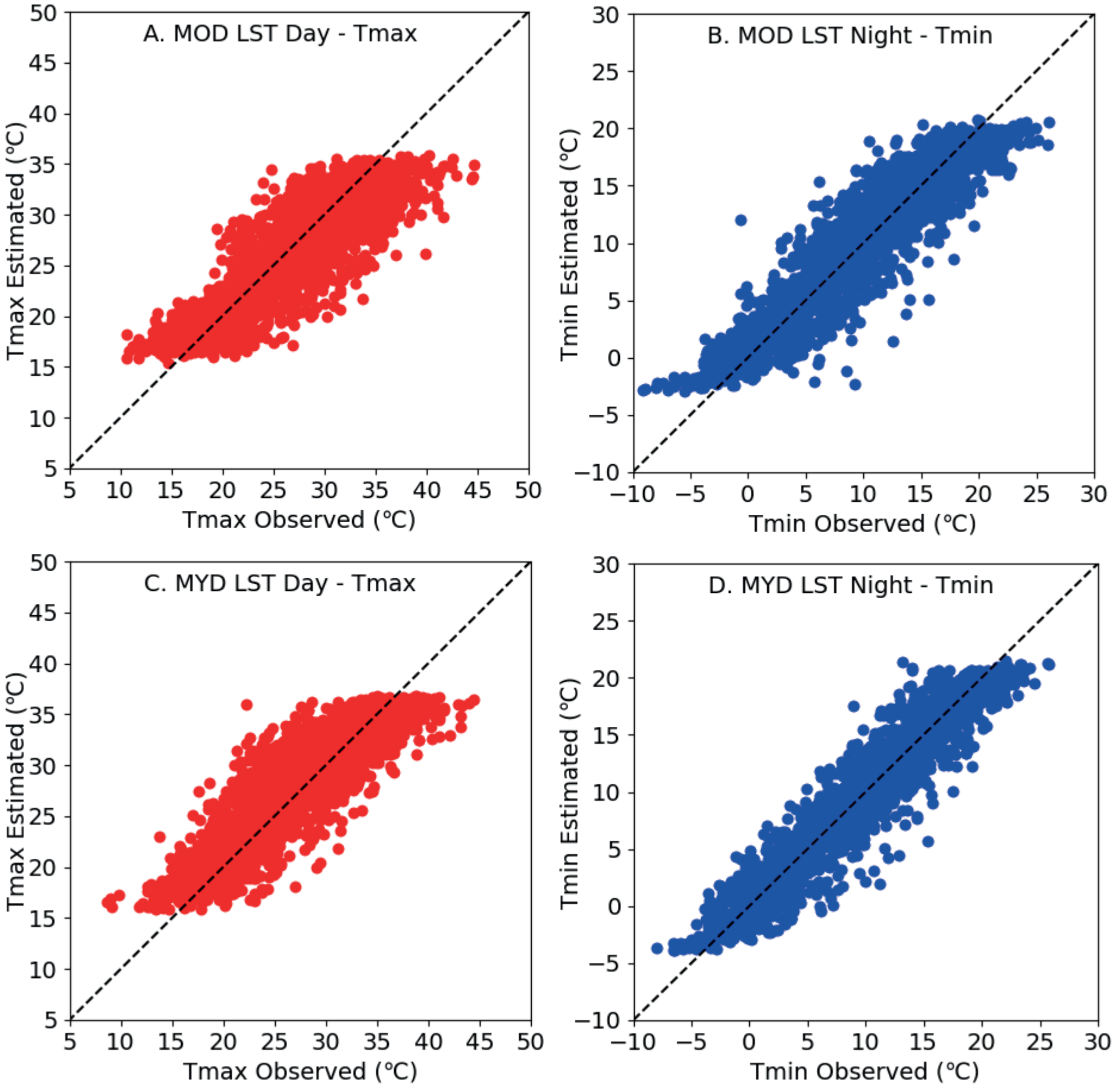


Figure 8. Scatter plots of the best artificial neural network models to estimate daily Tmax and Tmin temperature using TERRA and AQUA MODIS data, and auxiliary variables; dashed lines represent 1:1 line

The statistics obtained when RF models were applied (Table 3) are similar to the results observed by Xu et al. (2014) who estimated daily Tmax in Canada, using random forest models with nine environmental variables including AQUA LST, NDVI and solar radiation. They found $R^2 = 0.74$ a lower value than $R^2 = 0.81$ calculated in this paper. Likewise, Zhao et al. (2019), states that the view time difference can potentially introduce significant differences in surface temperature during the daytime observation because of the big impact of incoming solar radiation on LST evolution.

Noi et al. (2017) used algorithms of random forests for daily Tmax and Tmin estimation; they considered combinations of MOD and MYD LST data with two additional auxiliary datasets (elevation and Julian day), in the mountainous area of Vietnam, for a period of five years. These authors found that models for Tmax/Tmin estimation produced R^2 and RMSE values equal to 0.87/0.80 and 2.1/2.1 °C, respectively.

The ANN models present coefficients of determination (Table 4) in the range of those found by various authors. Marzban et al. (2018) estimated Ta (registered on the MODIS overpass times) with LST from MODIS TERRA and AQUA and eleven environmental variables, for a period of six years in Berlin. These authors used an ANN model and

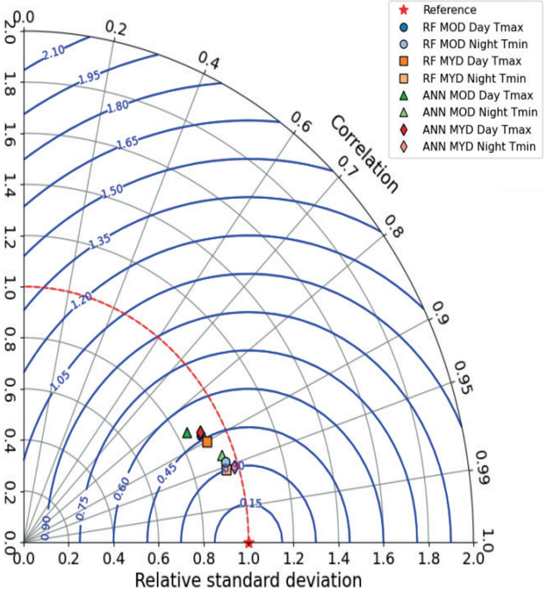


Figure 9. Taylor diagrams based on the correlation coefficient *(r)*, relative standard deviation (RSD), and centered root mean square error (RMSEc) (blue lines) of registered versus estimated temperatures derived from MODIS product, for random forest and artificial neural network best models (the red star symbol represents the reference dot)

obtained R^2 and RMSE values of 0.95 and 2.2 °C, respectively. Aher et al. (2011) applied LST derived from thermal bands of Landsat-TM/ETM+ images (1998-2002, India), in combination with ground measurements of meteorological data, as inputs to ANN models to estimate Ta.

The results of the proposed model showed that the best coefficient of determination and Root Mean Squared Error were 0.98, and 0.4 °C, respectively. Kamoutsis et al. (2013) estimated mean temperature in Greece (urban and adjacent mountain regions) using data of reference stations, with ANN, for one year (December 2009–November 2010). They presented results with determination coefficients between 0.74 and 0.93. Chronopoulos et al. (2008) also using ANN for mean Ta estimation in different meteorological stations, considering mean temperature and relative humidity, found validation R^2 between 0.61 and 0.91.

CONCLUSIONS

This paper demonstrates that RF and ANN machine learning methodologies were capable of modeling the relationships between registered temperature and MODIS LST data and auxiliary variables, in a very robust way in Córdoba (Argentina).

The results of this study confirm that Tmax and Tmin can be accurately estimated using AQUA and TERRA LST, and auxiliary data: solar radiation, NDVI and precipitation. The performance of models was similar using either of the two satellites. The LST from night/day satellite overpass time was the best to estimate Tmin/Tmax, respectively.

The robustness and confidence of the models developed and the easy and free accessibility of input data at a global scale suggest that the methodology presented has the potential to be applied in other regions. Because the models used in this paper employ only satellital and calculated information as input data, they can be applied to estimate Tmin/Tmax in regions where there is a small number of operational meteorological stations or to fill the missing data in a record.

Future studies will focus on the introduction of geographical information (e.g. latitude and longitude), temporal information (e.g. Julian days), and other surface status information (e.g. soil moisture, wind velocity) to further improve the generalization ability of our models.

REFERENCES

Aher, P. D., Adinarayana, J., and Gorantiwar, S. D. (2011). Remote Sensing and Artificial Neural Network in Spatial Assessment of Air Temperature in a Semi-arid Watershed. *International Journal of Earth Sciences and Engineering*, *4*(6), 351-354. https://d1wqtxts1xzle7.cloudfront.net/30680999/020410229-with-cover-page-v2.pdf?Expires=1654361918&Signature=egT15fyoyGBtbHxgeRoCZ~s5fqWPV-IJcmIhwV1fDZAOgWLM26nL2rUG~ZRiBLAeDOdZL6ID62X29lASrktIX0I-maORsYqtzgwGivHgNBAaKa63-aXTEFog14zM9h8oOPUbjvwyda4PQDqUzGT2a-jV76dgMyRYok4G-8EqaQZznFfgO0LiooiacFe9Z-ZMbqem8tpUx4~Qut5H5GOnk4BOk-st5~BzlYArzvscmmGnsa52ReeieHsrFj5chh~6YHy5z5Vj-9ALU80ShlRuHZLAj5zFqgdBS9ZU8sfyb9BXsI5aUlsQNklEcAEdDM~MN5BgSLTf6lYTEK58u4l4DzwTg__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

Alademomi, A. S., Okolie, C. J., Daramola, O. E., Agboola, R. O., and Salami, T. J. (2020). Assessing the Relationship of LST, NDVI and EVI with Land Cover Changes in the Lagos Lagoon Environment. *QuaestionesGeographicae*, *39*(3), 87-109. https://doi.org/10.2478/quageo-2020-0025

Alfieri, S. M., De Lorenzi, F., Bonfante, A., Basile, A., and Menenti, M. (2012). Mapping Air Temperature by Fourier Analysis of Land Surface Temperature Time Series Observed by TERRA/MODIS. In *Proceedings of the 1st EARSeL Workshop on Temporal Analy-*

sis of Satellite Images (21-24). Mykonos, Greece. https://www.researchgate.net/profile/Francesca-De-Lorenzi/publication/260979297_Mapping_Air_Temperature_by_Fourier_Analysis_of_Land_Surface_Temperature_time_series_observed_by_TERRAMODIS/links/5cdc2718a6fdccc9ddaebe19/Mapping-Air-Temperature-by-Fourier-Analysis-of-Land-Surface-Temperature-time-series-observed-by-TERRA-MODIS.pdf

Aliaga, V., Ferrelli, F., and Piccolo, M. C. (2017). Regionalization of climate over the Argentine Pampas. *International Journal of Climatology, 37* (Suppl.1), 1237-1247. https://doi.org/10.1002/joc.5079

Bartkowiak, P., Castelli, M., and Notarnicola, C. (2019). Downscaling Land Surface Temperature from MODIS Dataset with Random Forest Approach over Alpine Vegetated Areas. *Remote Sensing*, *11*(11), 1319. https://doi.org/10.3390/rs11111319

Benali, A., Carvalho, A. C., Nunes, J. P., Carvalhais, N., and Santos, A. (2012). Estimating air surface temperature in Portugal using MODIS LST data. *Remote Sensing of Environment*, *124*, 108-121. https://doi.org/10.1016/j.rse.2012.04.024

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Chang, Y., Ding, Y., Zhao, Q., and Zhang, S. (2020). A Comprehensive Evaluation of 4-Parameter Diurnal Temperature Cycle Models with In Situ and MODIS LST over Alpine Meadows in the Tibetan Plateau. *Remote Sensing*, *12*(1), 103. https://doi.org/10.3390/rs12010103

Chronopoulos, K. I., Tsiros, I. X., Dimopoulos, I. F., and Alvertos, N. (2008). An application of artificial neural network models to estimate air temperature data in areas with sparse network of meteorological stations. *Journal of Environmental Science and Health Part A*, *43*(14), 1752-1757. https://doi.org/10.1080/10934520802507621

Deery, D., Jimenez-Berni, J., Jones, H., Sirault, X., and Furbank, R. (2014). Proximal Remote Sensing Buggies and Potential Applications for Field-Based Phenotyping. *Agronomy*, *4*(3), 349-379. https://doi.org/10.3390/agronomy4030349

Emamifar, S., Rahimikhoob, A., and Noroozi, A. A. (2013). Daily mean air temperature estimation from MODIS land surface temperature products based on M5 model tree. *International Journal of Climatology*, *33*(15), 3174-3181. https://doi.org/10.1002/joc.3655

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D. and Moore, R. (2017). *Google Earth Engine*: *Planetary-scale geospatial analysis for everyone.* Remote Sensing of Environment. Google. https://earthengine.google.com/

Heck, E., de Beurs, K. M., Owsley, B. C., and Henebry, G. M. (2019). Evaluation of the MODIS collections 5 and 6 for change analysis of vegetation and land surface temperature dynamics in North and South America. *ISPRS Journal of Photogrammetry and Remote Sensing*, *156*, 121-134. https://doi.org/10.1016/j.isprsjprs.2019.07.011

Huang, W. R., Chang, Y. H., and Liu, P. Y. (2018). Assessment of IMERG precipitation over Taiwan at multiple timescales. *Atmospheric Research*, *214*, 239-249. https://doi.org/10.1016/j.atmosres.2018.08.004

Infraestructura de Datos Espaciales de la Provincia de Córdoba. (2019). Mapas Córdoba - Geoportal IDE de la Provincia de Córdoba. Ministerio de Finanzas, Provincia de Córdoba. https://www.mapascordoba.gob.ar/Documentos/METADATOS_LIMITANTES_DE_SUELO.pdf

Janatian, N., Sadeghi, M., Sanaeinejad, S. H., Bakhshian, E., Farid, A., Hasheminia, S. M., and Ghazanfari, S. (2017). A statistical framework for estimating air temperature using MODIS land surface temperature data. *International Journal of Climatology*, *37*(3), 1181-1194. https://doi.org/10.1002/joc.4766

Kamoutsis, A. P., Matsoukis, A. S., and Chronopoulos, K. I. (2013). Air temperature estimation by using artificial neural network models in the greater Athens area, Greece. *International Scholarly Research Notices*, *2013,* 489350. https://doi.org/10.1155/2013/489350

Kärnä, T. and Baptista, A. M. (2016). Evaluation of a long-term hindcast simulation for the Columbia River estuary. *Ocean Modelling*, *99*, 1-14. https://doi.org/10.1016/j.ocemod.2015.12.007

Long, D., Yan, L., Bai, L., Zhang, C., Li, X., Lei, H., Yang, H., Tian, F., Zeng, C., Meng, X., and Shi, C. (2020). Generation of MODIS-like land surface temperatures under all-weather conditions based on a data fusion approach. *Remote Sensing of Environment*, *246*, 111863. https://doi.org/10.1016/j.rse.2020.111863

Ministerio de Agricultura y Ganadería. (2021). Gestión de Estaciones Meteorológicas Clima. Gobierno de la Provincia de Córdoba. https://newmagya.omixom.com/accounts/login/?next=/

Marzban, F., Conrad, T., Marzban, P., and Sodoudi, S. (2018). Estimation of the Near-Surface Air Temperature during the Day and Nighttime from MODIS in Berlin, Germany. *International Journal of Advanced Remote Sensing and GIS*, *7*(1), 2478-2517. https://doi.org/10.23953/cloud.ijarsg.337

Noi, P. T., Degener, J., and Kappas, M. (2017). Comparison of Multiple Linear Regression, Cubist Regression, and Random Forest Algorithms to Estimate Daily Air Surface Temperature from Dynamic Combinations of MODIS LST Data. *Remote Sensing*, *9*(5), 398. https://doi.org/10.3390/rs9050398

Oyler, J. W., Dobrowski, S. Z., Holden, Z. A., and Running, S. W. (2016). Remotely Sensed Land Skin Temperature as a Spatial Predictor of Air Temperature across the Conterminous United States. *Journal of Applied Meteorology and Climatology*, *55*(7), 1441-1457. https://doi.org/10.1175/JAMC-D-15-0276.1

Sayago, S. and Bocco, M. (2018). Crop yield estimation using satellite images: comparison of linear and non-linear models. *AgriScientia*, *35*(1), 1-9. https://doi.org/10.31047/1668.298x.v1.n35.20447

Sayago, S., Ovando, G., and Bocco, M. (2017). Landsat images and crop model for evaluating water stress of rainfed soybean. *Remote Sensing of Environment*, *198*, 30-39. https://doi.org/10.1016/j.rse.2017.05.008

Sobrino, J. A. and Irakulis, I. (2020). A Methodology for Comparing the Surface Urban Heat Island in Selected Urban Agglomerations Around the World from Sentinel-3 SLSTR Data. *Remote Sensing*, *12*(12), 2052. https://doi.org/10.3390/rs12122052

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, *106*(D7), 7183-7192. https://doi.org/10.1029/2000JD900719

Trenberth, K. E. and Shea, D. J. (2005). Relationships between precipitation and surface temperature. *Geophysical Research Letters*, *32*(14). https://doi.org/10.1029/2005GL022760

Wehbe, M. B., Seiler, R. A., Vinocur, M. G., and Tarasconi, I. E. (2018). Is It Possible to Completely Adapt Agriculture Production to the Effects of Climate Variability and Change in Central Argentina? New Approaches in Face of New Challenges. In F. Alves et al. (Eds.), *Theory and Practice of Climate Adaptation* (443-463). Springer, Cham.

Xu, Y., Knudby, A., and Ho, H. C. (2014). Estimating daily maximum air temperature from MODIS in British Columbia, Canada. *International Journal of Remote Sensing*, *35*(24), 8108-8121. https://doi.org/10.1080/01431161.2014.978957

Yang, Y. Z., Cai, W. H., and Yang, J. (2017). Evaluation of MODIS Land Surface Temperature Data to Estimate Near-Surface Air Temperature in Northeast China. *Remote Sensing*, *9*(5), 410. https://doi.org/10.3390/rs9050410

Zeng, L., Wardlow, B. D., Tadesse, T., Shan, J., Hayes, M. J., Li, D., and Xiang, D. (2015). Estimation of Daily Air Temperature Based on MODIS Land Surface Temperature Products over the Corn Belt in the US. *Remote Sensing*, *7*(1), 951-970. https://doi.org/10.3390/rs70100951

Zhao, W., Wu, H., Yin, G., and Duan, S. B. (2019). Normalization of the temporal effect on the MODIS land surface temperature product using random forest regression. *ISPRS Journal of Photogrammetry and Remote Sensing*, *152*, 109-118. https://doi.org/10.1016/j.isprsjprs.2019.04.008