ESTIMATION OF RELATIVE HUMIDITY IN MESOAMERICA AND CARIBBEAN REGION USING SATELLITE DATA

By

CESAR M. SALAZAR AQUINO

A thesis submitted in partial fulfillment of the requirements for the degree of

MASTERS OF SCIENCE

In

INDUSTRIAL ENGINEERING

UNIVERSITY OF PUERTO RICO
MAYAGÜEZ CAMPUS

2017

Approved by

_____	_____
Nazario D. Ramírez Beltran, PhD. President, Graduate Committee	Date
_____	_____
Mauricio Cabrera Ríos, PhD. Member, Graduate Committee	Date
_____	_____
Jonathan Muñoz Barreto, PhD. Member, Graduate Committee	Date
_____	_____
Víctor Huérfano Moreno, PhD. Representative of Graduate Studies	Date
_____	_____
Viviana Cesaní Vázquez, PhD. Industrial Engineering Department Head	Date

ABSTRACT

This work addresses the estimation of the relative humidity (RH), which is related to the way the body reacts under extreme hot events, and in consequence has influence in human health. RH is commonly obtained from ground stations, however in areas with few stations data is not enough to obtain a reliable spatial estimation of RH. Therefore, an algorithm is proposed to retrieve RH based on satellite data from two instruments: Moderate Resolution Imaging Spectroradiometer (MODIS), installed on Terra and Aqua satellites, and Imagery, installed on a Geostationary Operational Environmental Satellite (GOES). The proposed models estimate RH in hourly basis and at 4 km over the Mesoamerica and Caribbean region. Results show the coefficient of multiple determination (R^2) range from 0.49 to 0.70; and the root mean squared error and mean absolute error vary from 7.71% to 9.64% and 5.76% to 7.38% respectively.

RESUMEN

Este trabajo propone la estimación de la humead relativa (RH), la cual es una variable relacionada con la forma en la que el cuerpo humano reacciona ante eventos de calor extremos, y está, en consecuencia, relacionada con la salud de las personas. RH es usualmente obtenida de estaciones localizadas en la superficie, sin embargo, en zonas donde existen pocas estaciones resulta insuficiente la cantidad de datos para obtener una estimación espacial confiable de la RH. Se propone un algoritmo para obtener estimaciones de RH basado en data de satélite la cual es extraída de dos instrumentos: MODIS, el cual está instalado en los satélites Aqua y Terra, y en Imagery, instalado en los satélites GOES. Los modelos propuestos estiman la RH en forma horaria y a una resolución espacial de 4 km sobre la región de Mesoamérica y el Caribe. Los resultados muestran que el coeficiente de múltiple determinación (R^2) varía entre 0.49 y 0.70. Además, la raíz cuadrada del error cuadrático medio y un error promedio absoluto desde 7.71% a 9.64% y 5.76% a 7.38% respectivamente.

This work is dedicated to God and especially to Elva Valenzuela Acosta,

Carla Aquino Valenzuela and Manuel Salazar Cerreño, to my family,

and to the memory of Cesar Aquino Quiroz.

ACKNOWLEDGMENTS

To God for all his mercy and blessings, for give me the opportunity to pursue this degree. To my family, especially to my parents and grandparents. Thanks for being with me during the entire process, even at the distance your advice, support and love motivated me to achieve this goal. This journey would not be being feasible without you.

To my advisor Dr. Nazario N. Ramírez Beltran, thanks for your trust and help and financial support, this research work will not be possible without your advice. Thanks also for being my mentor in this research process, I am learned a lot from you and I enjoyed my time working with you. Thanks for give me the opportunity to being part of your research team, I have grown thanks to that experience. Also, thanks to my Friend and research team partner Joan M. Castro Sanchez, for your help and time during my research time.

To the Industrial Engineering school at the UPRM and to their professor for guide me thought my MS degree, I am proud to be part of this community. Thanks for their economic support during my time as TA.

To my friends in Puerto Rico, for being like a family for all the afternoons talking about everything, for all the jokes and lunches that we share, for all the travels together, and in summary for the countless time that we spend together. As Edna Buchanan says “Friend are the family that one chooses for ourselves”. In summary, to every single person that I had known in Puerto Rico, for teach me that you can be at home if they are people that you can count as family.

To NSF/ENG Environmental Sustainability Program with award number: CBET – 1438324, for the economical supported.

TABLE OF CONTENT

FIGURE LIST

TABLE LIST

GLOSSARY OF TERMS

ANN - Artificial neural network.

AVHRR - Advanced Very High Resolution Radiometer.

BP - Backpropagation algorithm.

BT - Brightness temperature.

DEM – Digital elevation model.

FFT - Fast Fourier Transform.

GOES – Geostationary Operational Environmental Satellite.

LST – Land surface temperature.

MAC – Mesoamerica and Caribbean.

MAE – Mean absolute error.

MODIS - Moderate Resolution Imaging Spectroradiometer.

NCEP - The National Center for Environmental Prediction.

NDVI – Normalized difference vegetation index.

NIR – Near infrared.

PW – Precipitable water.

RH – Relative humidity.

RMSE – Root mean squared error.

VIF - Variance inflation factor.

TABLE OF DEFINITIONS

Variable	Definition
RH	Relative humidity is defined by NASA (2016) and Ahrens (2013) as a ratio of the water vapor present in the air to the water vapor that is necessary for saturation. RH is also defined as the ratio of the actual vapor pressure versus the saturation vapor pressure, in percentage units (Ahrens, 2013).
LST	LST is an atmospheric variable that describes the value of temperature captured or sensed over the surface of the earth. This product is focused to only the land covered areas of the planet. Akhoondzade and Saradjian (2008) defines LST as the portion of radiation, that the land surface emits, perceived by MODIS in a certain angle.
PW	American Meteorology Society defines PW, in its Glossary of meteorology, as: the amount of water vapor that is inside a column of a specified area. This area can be defined as an ideal segment of a projected point, or is defined by the spatial resolution of the instrument used. This amount of vapor is usually limited by two different altitude levels. PW is the relation between the volumes of condensed water that is occupied divided by the specified defined area (AMETSOC, 2016). Marin et al. (2015) defines PW as the amount or level of liquid water that results when all the vapor over a determined area is condensed and precipitated.
NDVI	Australian bureau of meteorology defines NDVI as an index to measure the level of vegetation and is based on the difference of two bands, usually visible and infrared. It is usually related to how dense the canopy, or the fraction of land that is covered with vegetation (BOM, 2016).
BT	It is defined as the temperature value that will be assigned to a black body (body with a surface emissivity equal to 1) to emit, on the same wavelength, the same value of radiation (GES DISC, 2016).

1. INTRODUCTION

Climate changes has become an important research topic. It is somehow motivated by the consequences on human health. One of the main drivers of the climate change is the elevated greenhouse gas concentration, where the increment of air temperature is one of the major consequences. However, the way that humans feel temperature is influenced by the RH; it increases the perceived temperature when both variables are high. The prolonged extreme hot events over several days is called a heat wave and impacts human health, causes agricultural devastation, etc. For instance, in Chicago during July 1995 a heat wave killed 522 people (Levy et al., 2011). The Caribbean usually shows high levels of temperature and RH and consequently people are exposed to those heat waves. However, it does not only affect the MAC region.

To observe the influence of this variables it is necessary to track them on real time for the entire area of study, to understand the changes and to identify its effects. This study is focused on the opportunity to estimate RH in real time. It will be estimated using information from Terra, Aqua and GOES satellites. Terra and Aqua have a sensor called MODIS and provide three parameters that has been shown to have a physical relationship with RH. However, information from MODIS is limited to only two observations per day, it does not provide enough information to obtain hourly estimations. To derive hourly estimation, information from GOES satellite data will be used, it provides visible and infrared information every 30 minutes. It will be necessary to use information from each of the sources to do a good estimation of RH. This project presents the opportunity to apply industrial engineering tools, as: regression analysis, quality control techniques, artificial neural networks, optimization algorithms, as well as designing computational algorithms to organize and process big data sets.

Computational algorithms are important to organize data in time and space domains, and to identify the physical characteristics such as daytime, nighttime, clear sky, land ocean areas. Hence, if data is not properly organized the regression and mathematical tools will become useless. Quality control techniques are especially useful to remove inconsistent data. Regression methods will be important to develop the empirical models to estimate RH. Optimization algorithms are used to choose the variables or group of variables that better

explain the variability of RH. Artificial neural network was used to model the nonlinear behavior between the physical parameters and infrared brightness temperature with RH.

2. OBJECTIVES

The objective of this investigation is to derive a group of models to estimate RH from different satellite data, especially GOES infrared channels and MODIS physical parameters. RH estimations will be limited to land covered areas and under clear sky conditions. These models will be obtained based on two different techniques: Regressions and feed-forward artificial neural network techniques. This study is being developed for the Mesoamerica and Caribbean (MAC) region.

The following specific objectives were proposed as the axis to complete this project:

- To develop a model for estimating RH based on physical parameters and surface characteristics. The physical parameters are: PW, LST and the normalized difference vegetation index (NDVI) and the surface characteristics are: latitude, longitude, elevation and time.
- To estimate hourly LST and PW parameters for the MAC region.
- To develop a new set of models to estimate RH. These models are based on satellite data to derive hourly estimates of RH over the land of the MAC region.

3. JUSTIFICATION

RH is an atmospheric variable that affects the air temperature human perception and is closely related to the Heat Index, which is a function of air temperature and RH and is used to measure the human's perception temperature. The human body usually adapts to hot temperatures by perspiration, when heat is removed from our body by sweat evaporation. However, high values of RH reduce the evaporation rate causing lower heat removal from the body and hence the sensation of being overheated. A prolonged period of excessively hot weather may cause heart strokes, human death, and sever economic impacts (Kunkel et al. 1999). When the values of temperate and RH humidity became larger than average, human body feels temperature higher than its real value. In consequence, it is expected to observe an increase in the frequency of the use of air conditioning systems, which in consequence causes an increase in energy consumption.

RH is traditionally obtained from ground stations that are in charge of observing and storing weather data. However, in some regions, the number of stations that provide this product are scarce. For example, the MAC region has a low number of stations that provides RH compared to the continental U.S. Moreover, the time gaps in some of those stations decrease the quantity of available information. This problem limits the applicability of this data in scientific studies. The National Center for Environmental Prediction (NCEP) offer global grids of RH based on reanalysis. However, these data provide poor time resolution (4 times a day) and spatial resolution (2.5 º). Consequently, the MAC region is poorly covered by 234 grids (NCEP, 2016).

Satellite data appears to be a good solution to estimate RH, since it increases the time and spatial resolution and cover the MAC region. However, satellites do not offer RH observations. Nevertheless, the information and products from those satellites can be used to derive a real time estimate of RH for the MAC region. There are a couple of empirical models that are described in the literature, to estimate RH but those cannot be implemented since they were developed for some specific sensors and for a very limited area. Therefore, there is a need to derive an algorithm to retrieve RH. We expect to contribute to this area providing an hourly estimation model of RH that provides information for the land covered areas for the MAC region and under clear sky conditions and at 4 km spatial resolution.

RH is a key variable to improve the performance of the atmospheric models. The models that are used to estimate the weather and climate use data from the surface and the atmosphere. These models usually have issues in terms of precision motivated by the data that is also imprecise. The proposed algorithm will help to improve the performance of those models, since it will provide hourly estimation of RH at 4 km of spatial resolution.

4. LITERATURE REVIEW

The literature review is organized in four sections. The first one describes the literature that supports the physical parameters to estimate RH that are based on regression. The second section describes the artificial neural networks, which are helpful especially when relations between variables follows a nonlinear behavior. A third section describes the basic characteristics of the satellites. Finally, the fourth section describes the characteristics of the products and bands that will be used for estimating RH.

4.1 STATISTICAL MODELS TO ESTIMATE ATMOSPHERIC VARIABLES

In this section, different models studied in literature will be discussed. A particular focus will be given to the estimation of RH based on satellite data. It is proper to start this section by defining RH. It is defined by NASA (2016) and Ahrens (2013) as a ratio of the water vapor present in the air to the water vapor that is necessary for saturation. This means that RH does not have physical units. Furthermore, it is concluded that RH is always in the range from 0 to 1 or 0 to 100%. RH is also defined as the ratio of the actual vapor pressure versus the saturation vapor pressure, in percentage units (Ahrens, 2013).

The literature suggests that it is feasible to estimate surface RH based on satellite data. The research done by Peng et al. (2006), describes a model to estimate RH from satellite data. It is based on Moderate Resolution Imaging Spectroradiometer (MODIS) information against RH from ground observations. MODIS level 1 and level 2 products were included as the input variables. Those products were used to derive a set of variables related with RH and those are: specific humidity and air temperature. Specific humidity is closely related with PW, which is calculated from 5 different MODIS bands. Three of them are absorption bands: 17 (0.905 μm), 18 (0.936 μm) and 19 (0.940 μm); and the other two are atmospheric window bands: 2 (0.865 μm) and 5 (1.24 μm). The 5 bands are in the near infrared (NIR) spectrum. Specific humidity is estimated, as explained before, from PW using regression techniques, and obtaining a quadratic expression with a R^2 value of 0.922 and Root Mean Squared Error (RMSE) less than about 0.00032. Air temperature is available from MODIS as a product. Another important variable is the air pressure. This is a MODIS product; however, this product is quite inexact. Air pressure could be derived as a function of the elevation, which is obtained

from the Digital Elevation Model (DEM). DEM can be obtained from: http://www.ngdc.noaa.gov/mgg/dem/demportal.html (NGDC, 2016). From those variables, and using some mathematical equations RH is calculated obtaining a good approximation when comparing with observed values. Their method provides an estimation of RH but it is an 8 day-average. This model cannot be replicated in this study because it is required to obtain hourly estimations. Also, the results may be different, because training the model using 8 day-averages reduces the variability compared to hourly observations provided by the proposed model.

K.S. Han et al. (2005) studied the feasibility to derive surface level RH based on satellite data. This study was performed over the area of Quebec, at the west side of Canada. The time interval includes the first ten days of June and July 1997. It used two instruments mounted in different satellites: Advanced Very High Resolution Radiometer (AVHRR), mounted on NOAA number 12 and 14, and imagery, mounted on GOES 8. Both instruments have 5 bands, also called channels. GOES has one visible and 4 infrared bands and AVHRR has 1 visible, one near infrared (NIR) and 3 Infrared bands. The wavelength of bands 4 and 5 are similar. These bands sense radiation that is centered on 10.7 and 12.0 μm respectively. Bands 4 and 5 from both satellites and regression techniques were applied to derive RH. Those bands were used to derive emissivity that will work as input data to estimate surface temperature, PW and NDVI. These products are the basis of the calculation. However, some other variables were included as input variables, such as: elevation, local time and Julian day. On the other hand, the response variable came from observations of RH gathered from stations. An average error of 10.6% was obtained. However, in some stations, especially those located around mountains and over forest areas the error had a higher value.

Temperature and sometimes even RH can be obtained from ground stations. However, this information is by definition punctual, it means that station information is only valid for the position of the station.

Satellites, on the other hand, are a greater option for estimating atmospheric variables over larger areas. A group of satellite images can provide enough amounts of information to perform the spatial estimation task, for example, estimation of physical parameters over the entire Mesoamerica and Caribbean region. This however creates a new problem to solve, and it is the spatial and temporal resolution of those satellites. This spatial resolution is related with the pixel size and the minimum amount of area covered by it. Spatial resolution is

defined as the minimum amount of resolution that an image is capable to differentiate. This spatial resolution is different even on the same satellite and it depends on the chosen band or product. However, Satellites with a good spatial resolution may sometimes suffer from low temporary resolution.

4.2 ARTIFICIAL NEURAL NETWORKS

Other alternative to estimate RH is the use of the artificial neural network technique. For instance, Kuligowski and Barros (2001) applied the artificial neural networks techniques as well as data analysis to estimate vertical profiles of temperature and dew point based on satellite data. Akbari et al. (2008) also applied neural networks techniques to estimate Temperature and humidity based on MODIS satellite data. The challenge of the current work is use satellite data and ANN technique to estimate RH at the surface level.

When is suspected that there is a nonlinear relationship between a group of variables, neural networks appear as a good alternative to work with them. As a matter of fact, this group of techniques have been developed to work with this kind of data. From different types of neural network implementations, the backpropagation algorithm (BP) has been chosen in many studies. For example, Li et al. (2015) describes backpropagation as an algorithm commonly applied to approximate a model, however it is commanded by a local optimum vision rather than offering a global optimum vision.

The algorithm described in this section was presented by Hagan et al. (2014). Backpropagation is an algorithm based on neural networks. It is used over different applications, one of them being to approximate different functions. It is based on multilayer network which means that it contains a set of different layers between inputs and outputs. The common notation for those networks, is:

$$R\text{-}S^1\text{-}\ S^2\ldots\text{-}S^n \tag{1}$$

Where R is the number of inputs, S^i is the number of neurons inside the layer i, and n is the total number of layers proposed. It is common to find in literature that a good approximate of the number of layers is 2 or 3 layers, where the last one is the output layers and the others are hidden layers. To decide the number of neurons, is necessary to study the characteristics of the data to be estimated. The number of neurons is closely related in a direct way the number of inflection points. Each neuron is defined as a combination of a weight (w) multiplied by a transfer function and with a bias (b) value added.

$$n = w * f(x) + b \tag{2}$$

The transfer functions of the last layer depend on the data and the scale expected to obtain, it can be: Linear, log-sigmoidal, hyperbolic tangent, Step function, as others. The transfer function in the hidden layer is obtained by trial and error. The weight and bias are randomly obtained to create the starting point. Usually the number of neuron in the last layer is closely associated to the number of variables that are wanted to approximate. Each layer has its own inputs and outputs, and the outputs from one layer will be the input to the following layer. The outputs of the last layer are associated with the approximate values for the studied variables. The output of each layer is calculated on the following way:

$$a^{i+1} = f^{i+1}\left(w^{i+1} * a^{i} + b^{i+1}\right) \tag{3}$$

Equation 3 is defined as a matrix operation if one or more layers has more than one neurons. Where a^i is the output of the previous layer and acts as the input of the current layer and a^{i+1} is the output of the current layer. On the last layer, the output is called "a".

Initial weights and bias have to be trained and changed to obtain the best approximation, adapting its values during the process. These changes will be determined as a function of the error between observations and estimations from the training data and Function respectively. This stage is usually called performance index, and the error defined as:

$$F(x) = E[e^2] = E[(t-a)^2] \tag{4}$$

Where t correspond to the array of observed values, and x the vector that contain both bias and weights. If it is more than one output, the function will be expressed as a matrix operation, see equation 5:

$$F(x) = E[e^T * e] = E[(t-a)^T * (t-a)] \tag{5}$$

This algorithm is called backpropagation because the error will be used as an indicator to change the weight and bias for each layer, starting from the last layer in a recursive form until arriving to the initial layer in order to minimize the sum of squared errors. This iteration will be repeated until the sum of squared errors are minimized, this is usually when the algorithm accomplished convergence.

The gradient algorithm is used to perform the actualization of weights and bias, as follows:

$$w_{i,j}^{m}(k+1) = w_{i,j}^{m}(k) - \alpha * s_i^m * a_j^{m-1} \tag{6}$$

$$b_i^m(k+1) = b_i^m(k) - \alpha * s_i^m \quad (7)$$

Where alpha (α) is the learning rate and is a value selected between 0 through 1. Usually low values are the most common. The indicator "i" corresponds to the neuron in the "m" layer, "j" correspond to the neuron in the previous layer where the input comes from, and "k" is related to the number of iteration. Variable "s" is the sensitivity of the layer and it is the retrospective value that determines the change from the error calculated. The indicator "m" corresponds to the analyzed layer. If weights or bias are determined as matrix the BP equations will be:

$$W^m(k+1) = W^m(k) - \alpha * s^m * (a^{m-1})^T \quad (8)$$

$$b^m(k+1) = b^m(k) - \alpha * s^m \quad (9)$$

Sensitivity is the direction of the optimization search and it corresponds to the derivative of the squared errors with respect to the weights or bias. This algorithm is called backpropagation due to the nature of its calculations that start with the last layer and ends up with the first one. In consequence, the first sensitivity to be calculated corresponds to the last layer, using the following equation that is expressed in a matrix form:

$$s^M = -2 * F^M(n^M) * (t - a) \quad (10)$$

Where M is the last layer on the network, and $F^M(n^M)$ is a matrix that express the derivatives of the transfer function$f^M(n)$, with respect to the net input to the neuron, "n". The propagation of the sensitivity throughout the layers is expressed by the following recursive equation:

$$s^m = F^m(n^m)(W^{m+1})^T s^{m+1}, \text{ for m=M-1… 2, 1} \quad (11)$$

In this case, to calculate the sensitivity for a previous layer, it will be necessary to include in the calculations the sensitivity from the next layer.

The equations (8) and (9) are applied at each iteration and the algorithm stops once the convergence is accomplished; i.e., when the changes on the objective function are very small. In this case, the objective function it is defined as the sum of squared errors.

Literature suggest that this method has been applied before to approximate functions on different areas. Cheng et al. (2016) show in their work the utility of using BP compared with the regression method. Authors wanted to estimate and predict the ignition temperature and activation energy using a sample of 64 different Chinese coals and their blends. Using the Pearson correlation, they founded that the most relevant factors to estimate them were: moisture, volatile matter, calorific value and oxygen of coals. These factors were used into two

methodologies one is a nonlinear regression using a quadratic polynomial regressing and the other is BP with 3 layers: input layer, hidden layer and output layer. Results with BP algorithm, gave a better output in terms of the relative mean when compared to the regression model, obtaining error values of 1.22% and 3.89% respectively. However, this methodology results complex in application for the multiple unknown elements involved. For example, in the determination of the number of neurons included in the hidden layer specifically the first and the second layers. Another complication is the difficulty of the mathematical explanation of the model, because at the end it is a black box. However, results obtained with BP are promising for authors.

4.3 ENVIRONMENTAL SATELLITES

There are two different satellites that appear to be important when modeling RH. Both are included in the present investigation. Low Earth orbit or polar satellites such as Terra and Aqua and geosynchronous satellites such as GOES.

4.3.1 GENERAL CHARACTERISTICS OF GOES

Geostationary Operational Environmental Satellites (GOES) is the name received by a group of geosynchronous satellites, launched over different generations that are orbiting the Earth. Air University (2009) defines geosynchronous satellites as flying rockets that orbit the earth at about 35786 km, following a circular orbit and an inclination of zero degrees. They orbital period is the same when compared to the earth rotation (1 day or 24 hours). Rumerman (2009) also includes that these satellites are aligned with the equator, with an orbital speed that is the same when compared to the speed on the surface of the earth that is below. This generates the illusion of constant floating over a static position (Air University, 2009), which means, that at any time the satellite will be located over the same geographic area. In consequence, it is feasible to obtain observations over a specific area in a good temporal resolution. GOES-13 for example corresponds to the current satellite that covers the Caribbean area and offers images every 30 minutes.

To maintain this geosynchronous orbit is necessary to be at high altitudes, which has a negative effect over the data that leads to the loss of detail on the observations. It has been shown when analyzing the spatial resolution that it is considerably low, of about 4 km for the infrared channels. GOES-13 was mentioned as an example before because it is the latest generation of

satellites that is constantly offering images over the east side of the United States, and in consequence offering information for the MAC region.

GOES satellites have different instruments installed, for example, this investigation is based on the data gathered from the Imagery instrument. Imagery in its different generations throughout time has improved some of the channel installed on it. For example, the version installed on GOES 8 to 11 has 5 channels from 1 to 5, but since GOES 12, channel 5 has been replaced for channel 6. Table 1 show the different Imagery channels configurations per satellite:

Table 1: Wavelength and common application per GOES' imagery channel, extracted from Hillger and Schmit (2010).

GOES Imager Band	Wavelength Range (µm)	Central Wavelength (µm)	Meteorological Objective	Spatial resolution
1	0.53 to 0.75	0.65 (GOES-8/12) 0.63 (GOES-13/15)	Cloud cover and surface features during the day.	Up to 1km
2	3.8 to 4	3.9	Low cloud/fog and fire detection.	Up to 4km
3	6.5 to 7.0 5.8 to 7.3	6.75 (GOES-8/11) 6.48 (GOES-12/15)	Upper-level water vapor.	Up to 4km
4	10.2 to 11.2	10.7	Surface or cloud-top temperature.	Up to 4km
5	11.5 to 12.5	12.0 (GOES-8/11)	Surface or cloud-top temperature and low-level water vapor.	Up to 4km
6	12.9 to 13.7	13.3 (GOES-12/15)	CO_2 band: Cloud detection.	Up to 8km (GOES 12/13) Up to 4km (GOES 14/15)

GOES does not provide products by itself, data downloaded from it are usually data from their different channels that are called Counts. These values are divided in different datasets depending on which band was used to collect the measurements. Those can be either values of 8 or 16 bits. Channel 1 are in 16 bits and channels 2 to 6 are in 8 bits. (Hillger and Schmit, 2010).

Data from GOES is available online in the following web page:

http://www.class.ngdc.noaa.gov/saa/products/welcome (NOAA, 2016)

4.3.2 GENERAL CHARACTERISTICS OF POLAR SATELLITES TERRA AND AQUA.

Moderate resolution Imaging Spectroradiometer (MODIS) is an instrument to collect environmental data and is mounted in two different polar satellites Terra and Aqua. A polar satellite, different from the geosynchronous one, flown at lower altitudes, of about 705 km. It flies around both poles of the earth, either way from north to south or vice versa (MODIS, 2015).

In consequence of both characteristics, the spatial resolution is higher when compared to geosynchronous satellites, and the entire planet surface is covered for each satellite. However, the temporal resolution became poor. The number of daily observations is related to the number of times that the satellites pass over one specific area, and MODIS Terra and MODIS Aqua have the capacity to offer two passes per day.

In contrast from GOES, the amount of information, in terms of products that can be obtained from MODIS is higher, due to the existence of 36 bands. The corresponding spatial resolution for those are: 250 m. for bands numbers 1 and 2, 500 m. for bands 3 to 7 and 1 km for bands 8 to 36 (MODIS, 2016). However, bands are useful to estimate even more products. The data offers products in different levels, the raw data is given by level 1, and products by levels 2 to 4. Some of those products are: LST, PW, cloud mask, sea ice cover or NDVI. However, NDVI has a lower temporal resolution when compared to the others because it just offers 1 observation every 16 days. Usually most of the products have a spatial resolution of 1km however NDVI has a spatial resolution that increase up to 250m.

Data from MODIS is available online in the following web page:

https://ladsweb.nascom.nasa.gov/data/ (LAADS WEB, 2016)

It is proposed to estimate RH using satellite variables as predictors and observations from stations as an input variable. It is expected that a linear regression can be enough to express the relationship between the described variables. However, it is also possible that a linear regression could not be enough to obtain good results. To solve this issue, it is proposed to implement a neural networks technique, using a feedforward multilayer structure with the Levengberg-Marquardt Backpropagation as the learning algorithm.

4.4 DESCRIPTION OF SATELLITE PRODUCTS

4.4.1 PRECIPITABLE WATER

American Meteorology Society defines PW, in its Glossary of meteorology, as: the amount of water vapor that is inside a column of a specified area. This area can be defined as an ideal segment of a projected point, or is defined by the spatial resolution of the instrument used. This amount of vapor is usually limited by two different altitude levels. PW is the relation between the volumes of condensed water that is occupied divided by the specified defined area (AMETSOC, 2016). Marin et al. (2015) defines PW as the amount or level of liquid water that results when all the vapor over a determined area is condensed and precipitated.

PW is one of the products available to download from MODIS. However, as it was explained when defining MODIS, this product has some characteristics that limit its application as a predictor variable for the RH estimation model. From the different problems detected the most important is its availability, twice a day and only for areas captured by the satellite. Most of the time figures do not bring information about the entire studied area. This effect is observed in figure 01.

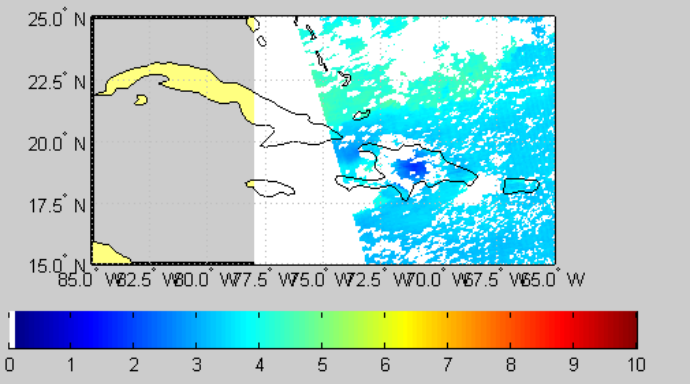


Figure 01: The portion of a MODIS PW image that fall inside the studied area – June 12 2012 at 02 30 UTC.

This product is calculated only over clear sky conditions areas, areas with no cloud interference. Pixels covered by clouds will be assigned a missing value code.

To process PW, the products downloaded from MODIS are MOD05_L2 (MODIS Terra) and MYD05_L2 (MODIS Aqua). Both have a spatial resolution of 5km.

4.4.2 NORMALIZED DIFFERENCE VEGETATION INDEX (NDVI)

NDVI is a normalized index that helps determine the distribution and the healthy conditions of the vegetation over the land. It is usually calculated based on the ratio of different bands that

are associated with the light absorbance effect that can be observed on the plant's surface. This scale depends on the author or the methodology. It usually falls between -1 to 1 or 0 to 1. Values are distributed as: zero if it is related to the presence of water, values below 0.3 for dry or low foliage areas, and values over 0.3 for areas with presence of vegetation. Australian bureau of meteorology defines NDVI as an index to measure the level of vegetation and is based on the difference of two bands, usually visible and infrared. It is usually related to how dense the canopy, or the fraction of land that is covered with vegetation (BOM, 2016).

It has been shown that NDVI is a factor related to humidity (Ulivieri et al., 1994), and could be considered important in the analysis.

Different from PW this product is not available for every satellite pass. Instead, it is a pre-processed product. Pre-processed images as in this case, offer the chance to correct some of the problems with data, for example the quantity of available information offering data for every single pixel over the observed area. A negative effect is the temporal resolution decrease. NDVI is only available every sixteen days.

MODIS NDVI, is composed by a group of images, each of those related to a portion of earth. For example, MAC region requires 23 images to obtain NDVI over the area. Figure 02 shows an example of the area covered by a single file, and Figure 03 shows the NDVI for a portion of the MAC region that can be represented using 10 images.

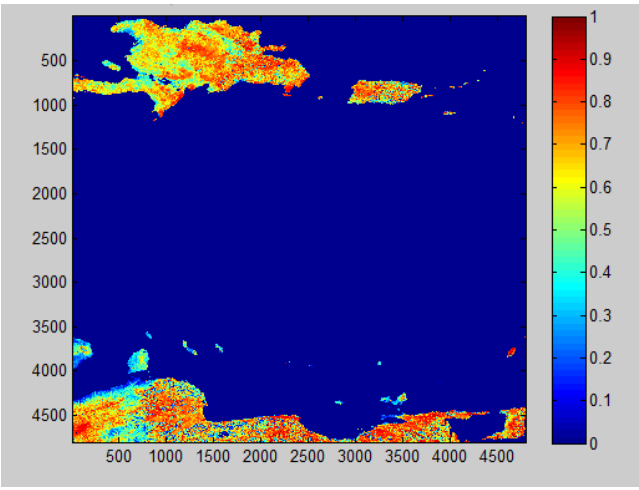


Figure 02: Downloaded NDVI observation from MODIS. Date: June 06 2012.

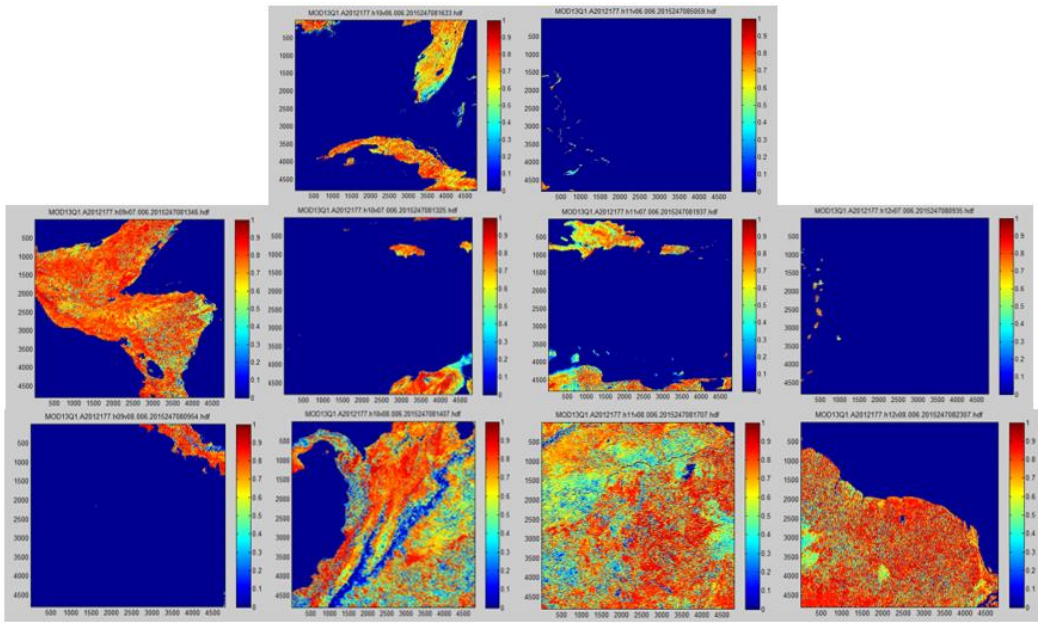


Figure 03: Composition of NDVI observations for Caribbean area. Date: June 06 2012.

To process NDVI, the products downloaded from MODIS are MOD13A2 (MODIS Terra) and MOD13A2 (MODIS Aqua). Both have a spatial resolution of 1km and a time resolution of 16 days.

4.4.3 LAND SURFACE TEMPERATURE

LST is an atmospheric variable that describes the value of temperature captured or sensed over the surface of the earth. This product is focused to only the land covered areas of the planet. Akhoondzade and Saradjian (2008) defines LST as the portion of radiation, that the land surface emits, perceived by MODIS in a certain angle.

In this study two LST files will be downloaded: MOD11_L2 for MODIS Terra and MYD11_L2 for MODIS Aqua, available on NetCDF format. LST captures information for the entire earth including the poles, in both day and night. (LPDAAC, 2016).

Wenhui Wang et al. (2008) provides a table that contains the most important information about the different LST products that are available from MODIS Terra. From that, MOD11_L2 in particular, has a spatial resolution of 1 km, and an accuracy of 1° Celsius, and the time interval between every observation captured by the sensor over its observed position is 5 minutes approximately. The area captured per observation is about 2,330 km of latitude by 2,000 km of longitude, which is equivalent to approximately a matrix of pixels with a dimension of 2,030 row by 1,354 Column (LPDAAC, 2016). These characteristics are also valid for MODIS Aqua.

LST, and PW data are available twice a day. In addition, PW is available over clear sky conditions (Ji Zhou et al., 2014). Also, every pixel that does not correspond to a land covered areas is assigned to missing value (NaN value). Figure 04 shows an instantaneous image of LST in Kelvin degrees. This figure corresponds to a portion of the studied area captured in January 1, 2011 at 3:10 UTC.

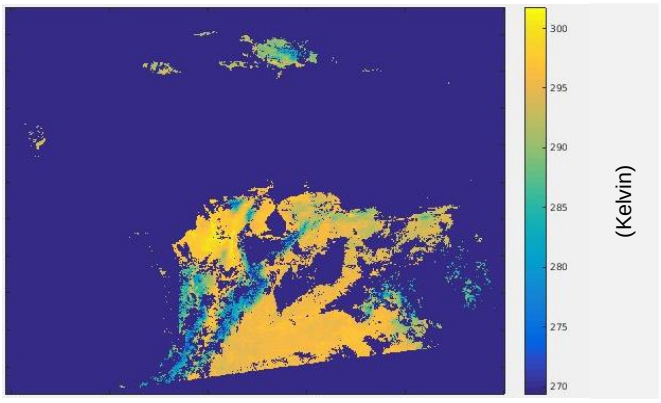


Figure 04: Preprocessed LST observation from MODIS Terra. Date: January 01 2011 at 03:10 UTC.

Different from the other MODIS PW, this one does not include a georeference component into the file. It is necessary to download a georeference component that should be combined with the observations. Geolocation files are downloaded as a product itself and it receives the codification MOD03 for MODIS Terra and MYD03 for MODIS Aqua.

5. METHODOLOGY: SATELLITE DATA PREPROCESSING

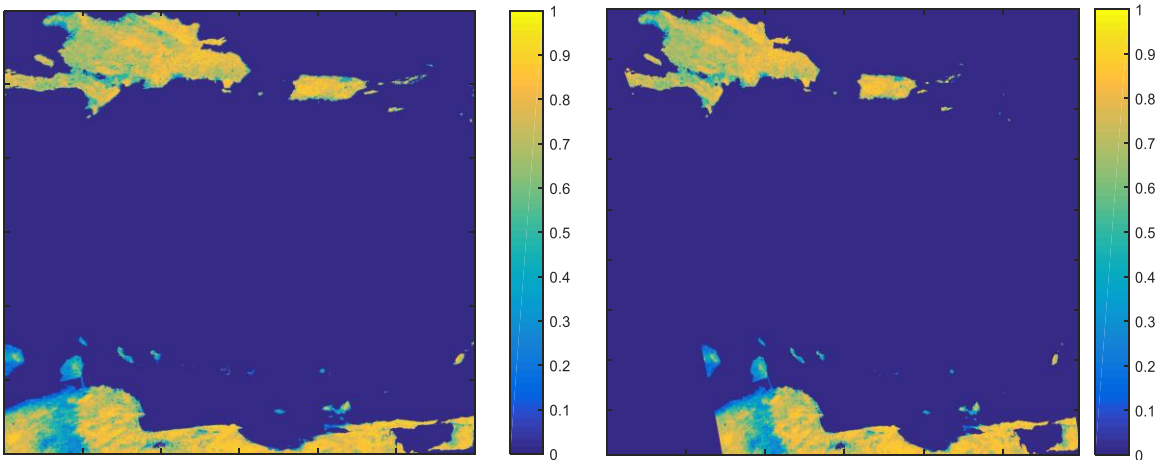
This chapter corresponds to the discussion of all the algorithms necessary to adequate the data obtained from the satellites to be introduced into the regression models. It starts defining a general procedure that will be necessary for every satellite dataset that is called georeferencing. Subsequently, the algorithms to preprocess each of the satellites variables will be described in detail.

5.1 GEOREFERENCING

This algorithm will be commonly employed in the preprocessing stage because most of the products extracted from satellites requires to be correctly processed. Data from satellites are presented as a matrix, where each element is associated to a grid or a pixel, however the original projection causes each pixel to appear bigger and smaller depending on their position from the equator. The georeferencing is a change in the size of the pixel transforming the original information into squares equally distributed with the corresponding geographical location, and when it is necessary, a resampling is performed to convert into a uniform spatial distribution.

This process start reading the coordinates of the original image, looking for their corners in terms of both latitude and longitude, and also reading the number of pixels that are found on the image in each direction. The corners and the number of pixels define the latitude and longitude for each pixel and they will be assigned to look for a uniform distribution and square pixels. The new latitude and longitude values for each pixel will be saved and they replace the previous values.

Sometimes the georeferencing process will be accompanied by an interpolation algorithm (or resampling) to change the resolution of the image and in consequence the number of pixels necessary to cover the same geographical area. An example of the georeferencing process over the data is shown in figure 05.



a. b.

Figure 05: NDVI 01 Sept. 2009 panel a: observation previous to the georeferencing process (as downloaded from the servers). Panel b: observation after the georeferencing process.

5.2 PW PREPROCESSING

PW is the easiest variable to be preprocessed from the three physical parameters. This process starts opening each of the PW images and performing the georeferencing process for each of them. This process also includes the interpolation of the product from its original 5 km to a 4 km resolution.

Furthermore, this image is cleaned by looking for missing values. Original missing values are linked to a numerical value of -9999 and should be transformed to a standard convention (NaN). Finally, a new image is saved which has a group of matrices that contains both the new latitude and longitude for each pixel and the corresponding PW values. An example of a preprocessed PW file is showed on figure 06:

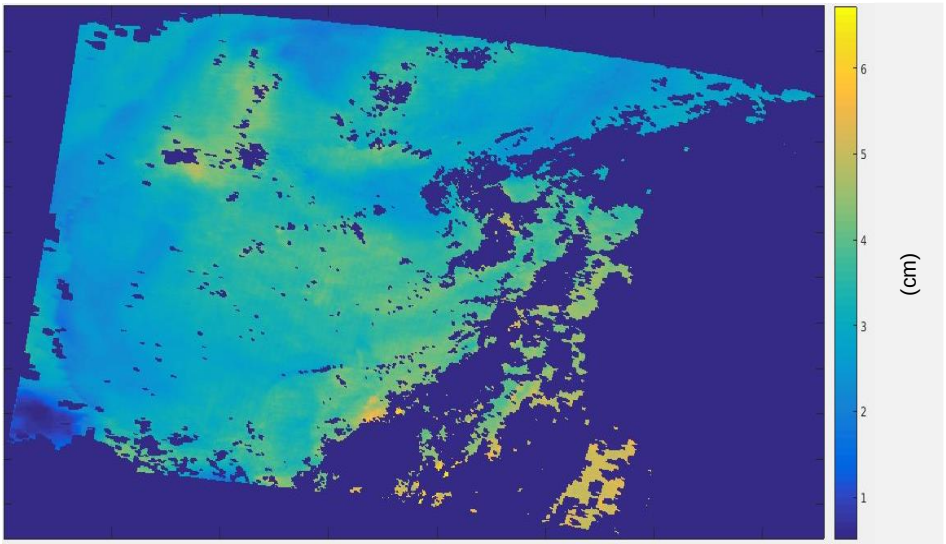


Figure 06: PW Aqua Date: June 10 2011 at 07:35 am

5.3 LST PREPROCESSING

In terms of the LST, to preprocess the data and to do the georeferencing process, it is necessary to include a different product called Geolocation which provides the corresponding set of coordinates necessary for those files.

This process starts opening each of the LST files and searching for the corresponding Geolocation file for the same time. If no file is found, then the LST file will be skipped. Once a match is found, the two files are linked and opened creating the first set of matrix that contains the coordinates from the Geolocation files and the temperature for the LST file. Then, the georeferencing process will be performed over these matrices. It becomes necessary to interpolate the files changing its original 1 km resolution to a 4 km resolution.

Similarly, the image is cleaned looking for missing values. Finally, the new image is saved. It has a group of matrices that contains both the new latitude and longitude for each pixel and also the corresponding values of LST. An example of a preprocessed LST file is showed on figure 07:

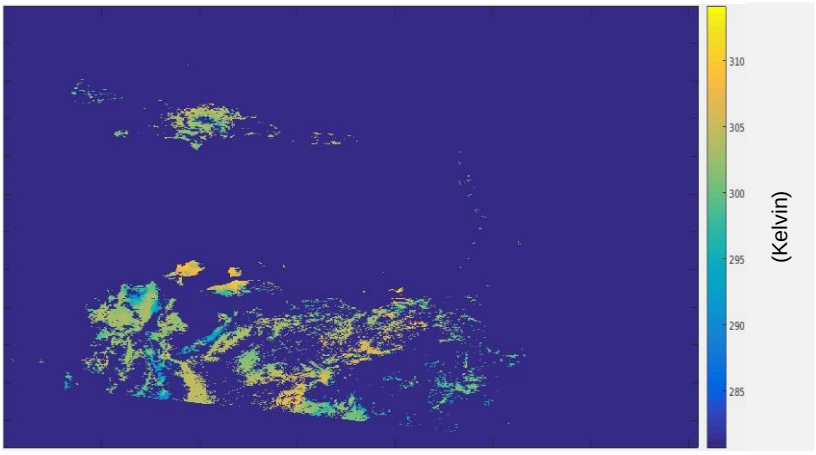


Figure 07: LST Terra Date: July 16 2011 at 03:00 pm

5.4 NDVI PREPROCESSING

NDVI requires a deeper level of preprocessing due to the characteristics of the product, which needs to be changed:

- The first step is to open each of the NDVI files and apply the georeferencing process. Also, it is important to interpolate them from their original 1 km to 4 km.
- The second step is to incorporate small NDVI files for the same time period into a single file that contains to the MAC region. To accomplish this task, it is necessary to join 23

NDVI images together based on their position (Latitude and Longitude). It is essential to be especially careful on the borders of the individual images in order to avoid errors in the process. A new image will be saved and it contains different matrices: Latitude, Longitude and NDVI.

- Finally, the time resolution of the NDVI files will be changed, from their original 16 days into hourly basis. This algorithm start analyzing the difference in time between each combination of two different NDVI images, and this time gap will be filled with a copy of the oldest of both NDVI files repeated every hour.

When all those three steps and their corresponding algorithms have been performed now the NDVI files are ready to be matched with the other products that serves as input variables. This algorithm will be repeated for both, MODIS Terra and MODIS Aqua observations. An example of a preprocessed NDVI file is showed on figure 08:

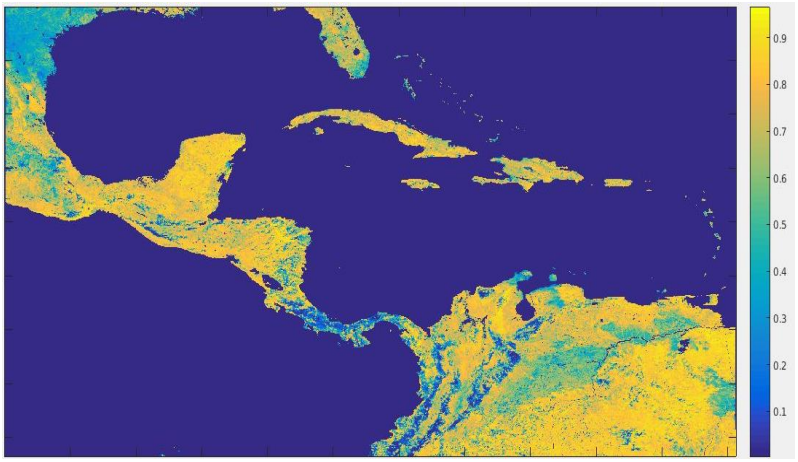


Figure 08: NDVI Aqua Date: July 30 2011 at 04:00 pm

5.5 STATION DATA PREPROCESSING

The stations dataset is a complex group of data. It is available online (NCDC, 2016). Information comes in two different types of datasets both on .txt format. The first one is a compendium of data, for either one station or a group of stations. Thus, if the file has more than one station they will be accommodated, one bellows the other, with the variables divided in different columns. The second dataset has information related to the station: the name, the country, the state, a reference code, the latitude, longitude, elevation and time.

Preprocessing downloaded station data requires implementing the task described in figure 09:

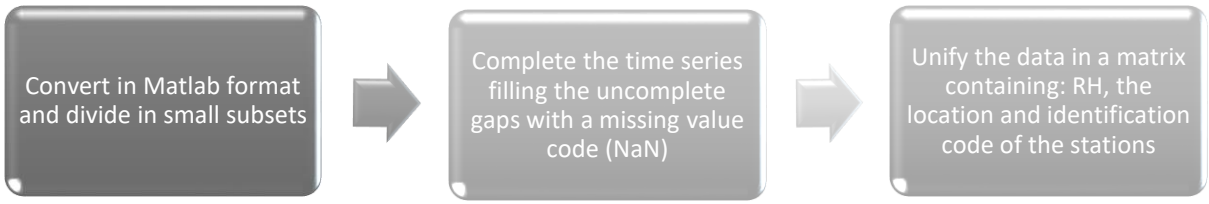


Figure 09: Algorithm framework for the station preprocessing process.

5.6 GOES DATA PREPROCESSING

GOES servers (NOAA, 2016) offer the option to download datasets that contain count values. Usually GOES provides one dataset per channel for every 30 minutes. To preprocess GOES data, it will be reduced to an hourly scale, which is the time interval for the estimation of RH. The process consists on finding the nearest observation to every hour, and eliminate the other observations for that particular hour, and repeat the process for every hour.

Then it is necessary to extract the information for each GOES channel. GOES imagery offers 5 channels (for details, see Table 1), however only the 4 infrared channels are included in the model. The visible channel was not included due to computational time in downloading and process. GOES imagery provide a set of variables called counts, however it is necessary to transform the count values into a more useful variable called the brightness temperature (BT).

BT is a measure of radiation, and it is defined as the temperature value that will be assigned to a black body (body with a surface emissivity equal to 1) to emit, on the same wavelength, the same value of radiation (GES DISC, 2016).

Two different equations were employed to transform count values to brightness temperature, the formulas are similar to each one of the channels (2 to 6) and that are divided in two different groups depending on the count value (OSPO, 2016). When the count value is lower than 176, the brightness temperature is calculated using the equation (12), in other cases it is calculated using the equation (13):

$$Bt_n = 330 - 0.5 * Counts_n \quad (12)$$

$$Bt_n = 418 - Counts_n \quad (13)$$

Where:

- Bt = Brightness temperature (Kelvin degrees).
- n = channel number.

It should be noted that, before saving the new brightness temperature from each channel it is necessary to georeferencing the information and to change their original resolution to a uniform 4 km for every channel. This product will be saved and for each hour the file is formed by a group of tables, and one table for each variable: latitude longitude and the brightness temperature from each channel. An example of a preprocessed GOES file, in terms of their brightness temperature is shown on figure 10:

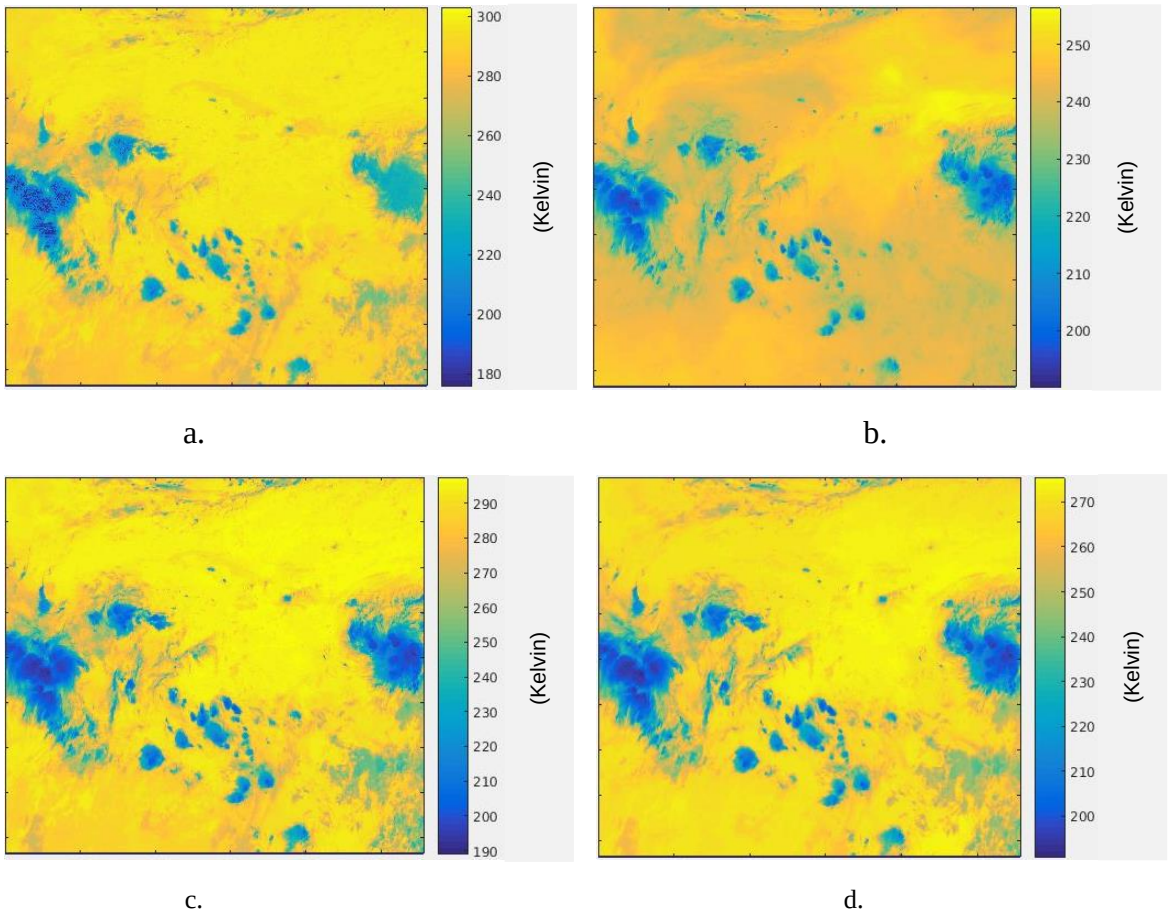


Figure 10: GOES BT Date: August 02 2011 at 10:00 am. Panel a.: channel 2 BT. Panel b.: channel 3 BT Panel c.: channel 4 BT Panel d.: channel 6 BT

6. ESTIMATION OF RELATIVE HUMIDITY BASED ON MODIS PHYSICAL PARAMETERS – STATION DATA, AND REGRESSION TECHNIQUES

The literature shows that, the physical parameters that are related to RH are: LST, PW and NDVI. However, to include them as important factors on the final model it is necessary to analyze what is the actual contribution of this parameter to explain the variability of RH. To measure the contribution of the physical parameters a preliminary model is developed in this chapter. This model is developed based on MODIS data, which are available twice a day; and therefore, all the variables involved in the model must match on time and space with MODIS data.

6.1 DATA DESCRIPTION

A regression model includes two types of variables regressors or predictors and the dependent variable or response variable (Montgomery et al.,2012).

The response variable is related to the variable or variables that will be estimated, in this case it is the RH. It is necessary to obtain observations of this variable to train the model, and these observations were obtained from weather stations, which are located across MAC region, and the hourly RH data were provided by NCDC (2016).

There are many stations that are located in the MAC region, but from those only 584 stations were selected. These stations were chosen based on two characteristics: those that offer hourly observations of RH, and the one that have information during the period 2011-2015. However, it is expected to find some missing values in those observations, and those have a missing value code which is 999. Figure 11 shows the location for the selected stations. Although 5 years of data are available, preliminary models include information only for 2011.

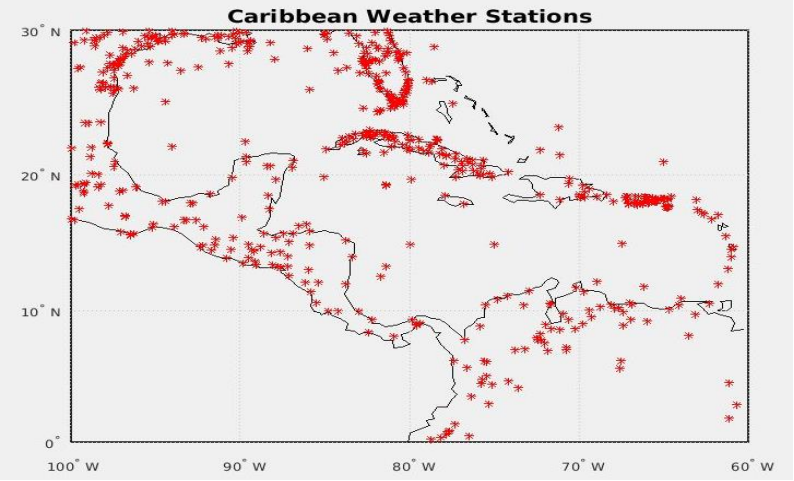


Figure 11: Selected weather stations in the MAC region.

The second group of variables correspond to the predictors or inputs of the model, which as explained on the introduction of this chapter is commanded by three physical parameters: LST, PW and NDVI. These products are obtained from MODIS instrument, mounted on Terra and Aqua satellites. These products have been preprocessed to modify some of their characteristics as the spatial and temporal resolution, those are described in table 2. Every satellite image is represented as a group of matrices that contains the product, the position and some other information for every pixel observed in the studied area. But there are some pixels that do not present information in the image, those will be coded with a missing value.

The model also includes variable that exhibit the geographical characteristics such as elevation. This variable is obtained from a digital elevation model (NGDC, 2016), and it has a spatial resolution of 4 km. It is presented as a group of matrices representing the pixel position and the product itself. This product is represented in figure 12:

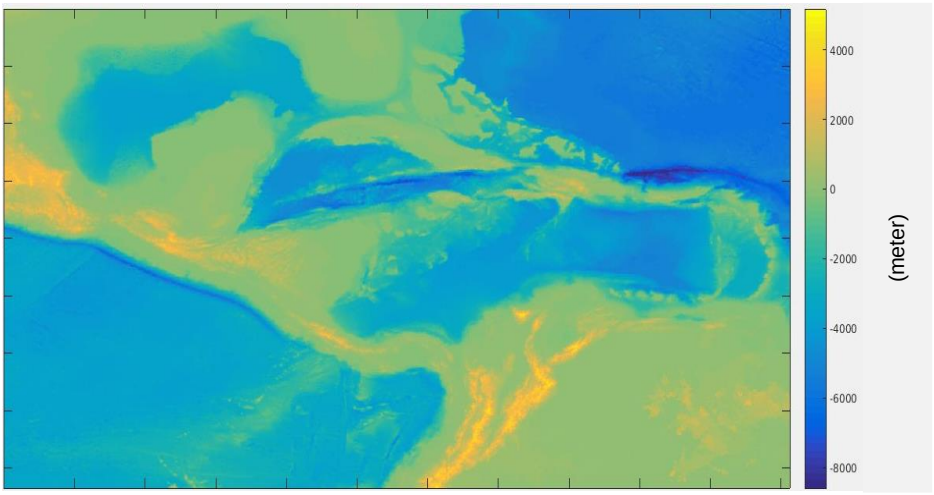


Figure 12: Elevation map 4km. Unit: meters.

Elevation values are not expected to change in a short period of time as the one presented here, it means that the same elevation file will be linked to every satellite image. Elevation is important to include as part of the components corresponding to the position. The characteristics of the elevation product will be found in table 2.

Table 2: Characteristics of the data

Product	Instrument	Spatial resolution	Time resolution
Land Surface Temperature	MODIS	4 km	Twice a day
Precipitable water	MODIS	4 km	Twice a day
NDVI	MODIS	4 km	Hourly
Relative Humidity	Station	N.A*	Hourly
Elevation	DEM	4 km	N.A.*

* Does not apply for this product.

The data described before correspond to the products that are directly gathered from different satellites, stations or models (Elevation). However, there exists some other variables that are included in the model but that are indirectly obtained from the previous products, those are: the location of each pixel (observation) expressed in terms of latitude and longitude and the time when each image was captured that are expressed in terms of Month Day and Hour. The images also expressed the time in terms of minutes, however this level of information is not necessary because the objective is to estimate RH every hour.

6.2 METHODOLOGY

Previously, a chapter was advocated to define the methodology in general terms; however, the methodology algorithms, change depending on the product that will be estimated and in the characteristics of the data.

6.2.1 MATCH ALGORITHM

This algorithm, is the start in the methodology to estimate the RH product. It has the function to do the match between the products from satellite and station into a single file, and keeping only the information necessary to be included into the model. This model has the objective to estimate RH based on the physical parameters obtained from MODIS. It means that it is necessary in this algorithm to do the match between stations and MODIS products.

The match process will be based on the stations, because they provide the location of each of the points that has information and the hourly time where it can be found. The first step is to

load the data from stations, saving the RH product, as well as the time related to each observation and the position of each selected station. Based on these data the match process is developed:

- Using the time information, the station data is trimmed, saving only the information corresponding to the year to train the model (2011). Every other year will be discarded from the dataset.
- Based on the time information a first match with the MODIS files will be performed. Every hour that exists in the station file is analyzed searching for the images from MODIS that correspond to this hour, using a window of 30 minutes up and 30 minutes down. If no images are found for this specific hour, the corresponding variables related to each station are filled with missing values. If a group of images are found, it is necessary to study each pixel searching for the closest to each station.
- The group of images corresponding to the hour are opened and the position values are studied searching for: the stations that are inside the group of images, and the nearest pixel to each station, each station is linked to the nearest pixel and the value for the specific variable in that hour for that station is assigned. For all the stations that do not have a pixel near to them in the studied hour, then the value of the variable related to this station will be a missing value. On the other hand, during the same hour a pixel near to them in more than one image corresponding to that hour, in this case the average of the values for this pixel in all the image is going to be used. This process is repeated for two of the physical parameters, the LST and the PW.
- However, this match process is a little different for the NDVI. Both variables are presented as hourly observations with constant information for the studied area. In this case the process is resumed to analyze the images for each hour, search for the nearest pixel to each station and save the values for both variables and repeated for every hour. If an hour does not have a corresponding image of NDVI, the corresponding values of this variable for all the station is a missing value. Also, if the corresponding pixel near to each station does not have information, it will be filled with a missing value. For the elevation, it is only necessary to search for the pixel nearest to each station and save the value corresponding to it.

The result of this algorithm is a group of matrices, one for each variable. They have the following structure: the number of columns corresponds to the number of stations, and the number of rows correspond to the observed RH value, or to the value of the observed physical parameter in the corresponding pixel to each station and for each of the hours during the studied year. Also, the table of time and position extracted from the stations and used to the match process and save results as matrices. For the time table exist 3 columns (month, day and hour) and one observation for every hour on the year. For the position table exist 3 columns, two for latitude and longitude and one for elevation. This table includes 584 rows, one for each station.

6.2.2 STRUCTURE AND CLEANING ALGORITHM

The objective of this algorithm is to generate the regression table that include both the input variables and the response. It uses as inputs the different products, which were imported as matrices that is how they were saved. But, to create a new table it is necessary to reshape each of those matrices to be saved as part of one bigger table where each column represent a specific variable and each row one observation. This process is different from each group of variables and will be explained bellow:

- The LST, PW, NDVI and RH were imported as matrices and each matrix has: one column for every station (584) and one row for every hour during the entire training year (8760). They will be converted and regrouped as vectors, one for variable. Those have as many rows as observation for station for each time that exists. The process is simple and consists in rearrange one station (column) bellow the other. The size of the new vector, corresponding to each variable, is 5115840×1
- The position matrix was imported as a matrix that has: 3 columns (latitude, longitude and elevation) and 584 rows (stations). It will be transform but keeping the 3 columns, one for each variable. Each row of the matrix represents one of the station and has 8760 observations related to it. Each of the values of the rows is the same for each of the 8760 observation, and it will be copied this number of times creating new rows, later in the next row the same process will be repeated for the next value and so on until all the original row values will be included. The size of the new matrix is 5115840×3
- The time matrix contains: 3 columns (month, day and hour) and 8760 observations (hours in the training year). It will be transform but keeping the 3 columns, one for a variable. This number of hours is the same for each of the 584 stations, in consequence

to transform it is necessary to copy this entire matrix one bellow the other 584 times. The size of the new matrix is 5115840×3

All those vectors and matrices, were aggregated one left to the other as columns in a new table. It has the required structure and products necessary to be introduced into the regression. This new array has included, per every station, all the variables included that correspond to a position near to it in every time that an observation is captured. But, sometimes some of the variables were not been captured in that image or in that time, it generates a missing value that must be eliminated, estimations will not be possible if any of the variables have at least one missing value.

To eliminate those missing values is necessary to analyze each observation at each time (row), and if any row has at least one variable that contains a missing value the entire observation (row) is eliminated for that specific time. After finishing this process this table is saved.

Also, it is necessary to eliminate every observation that presents values that are outside the normal limits of temperature for the specific area. In terms of RH every value that is above 100 or below 0 should be eliminated. In terms of the PW values should be at least 0. Same procedure will be employed for every variable. The rule to eliminate those observations is like the one implemented for missing values.

6.2.3 DIVISION AND DEVELOPMENT OF THE MODEL ALGORITHM

The tables saved before, one set for Terra products and one for Aqua, already provides the structure necessary to perform the estimation techniques. It contains, only as inputs the variables obtained from the satellites and the DEM. However, those may be insufficient to obtain a good estimation model. To improve the performance of this model, it is proposed to include some new variables based on combination of products, as well as transformation obtained from a statistical software. The entire list of variables included all the original variables, their transformation and combinations as well as the response variables are described in table 3:

Table 3: Description of the variables

Variable	Description	Variable	Description
X_1	Latitude	X_{27}	$NDVI^2$
X_2	Longitude	X_{28}	Sin(NDVI)
X_3	Elevation	X_{29}	Cos(NDVI)
X_4	Month	X_{30}	$NDVI^{-1/2}$
X_5	Day	X_{31}	PW+LST
X_6	Hour	X_{32}	PW-LST
X_7	$e^{(4.23113-0.28227/PW)}$	X_{33}	PW*LST
X_8	$(24485.5-0.224678*LST^2)^{1/2}$	X_{34}	PW/LST
X_9	$e^{(4.34226-0.0983105/NDVI)}$	X_{35}	LST/PW
X_{10}	PW	X_{36}	Ln(PW/LST)
X_{11}	Ln(PW)	X_{37}	Ln(LST/PW)
X_{12}	$e^{(PW)}$	X_{38}	PW+NDVI
X_{13}	PW^2	X_{39}	PW-NDVI
X_{14}	Sen(PW)	X_{40}	PW*NDVI
X_{15}	Cos(PW)	X_{41}	PW/NDVI
X_{16}	$PW^{-1/2}$	X_{42}	NDVI/PW
X_{17}	LST	X_{43}	Ln(PW/NDVI)
X_{18}	Ln(LST)	X_{44}	Ln(NDVI/PW)
X_{19}	$e^{(LST)}$	X_{45}	LST+NDVI
X_{20}	LST^2	X_{46}	LST-NDVI
X_{21}	Sen(LST)	X_{47}	LST*NDVI
X_{22}	Cos(LST)	X_{48}	LST/NDVI
X_{23}	$LST^{-1/2}$	X_{49}	NDVI/LST
X_{24}	NDVI	X_{50}	Ln(LST/NDVI)
X_{25}	Ln(NDVI)	X_{51}	Ln(NDVI/LST)
X_{26}	$e^{(NDVI)}$	**Y**	Station RH

This table now will be divided first in two different ones: the first one contains all the input variables (X) and the second one the response variable (Y). Two different techniques will be implemented to estimate RH those are: Forward estimation technique and group variable selection technique.

To obtain the best estimation of RH, two approximations will be adopted to work with the dataset. One is to estimate regression using the data for the MAC region. However, this big amount of data may result difficult to estimate based on the different variations in the data. Also, the large amounts of data increase considerably the processing time. To improve the model results a different approach is being considered. This is to divide the MAC region in homogenous zones.

6.2.3.1 Division in homogeneous zones

To explain the variability of a dataset sometimes became problematic, especially when these changes are barely related to a group of variables, and that might reduce the capability of a regression model. It can be noticeable when they are irregular groups of data with complex characteristics that may subtle the information from a specific and important dataset. To

segregate the data in small groups with homogeneous characteristics results in a useful methodology to reduce the variability inside of each group. With the delimitation of homogeneous zones, the regression can track the variability in a more efficient way without the misleading effect of data with different behavior. This solution also indirectly solves the problem with processing time, which grows considerably when the number of observations introduced in the regression increases.

The election and definition of this homogenous regions are based on the discoveries from our investigation group. And their selection is based on the geographic and climatic characteristics of the area, from the tropical Antillean islands to the desert areas of Texas to name a few examples. In this exercise the division generates 4 homogenous climatic zones: South America, Center America, North America and the Antillean islands. However, a particular case is observed on the region of Florida, which has conditions of humidity and temperature similar to the observed in the Antillean islands. For this reason, Florida is included in the Antillean group rather than in USA. Figure 13 shows the 4 areas defined before: North America (green), Center America (black), South America (Red) and Caribbean (Blue).

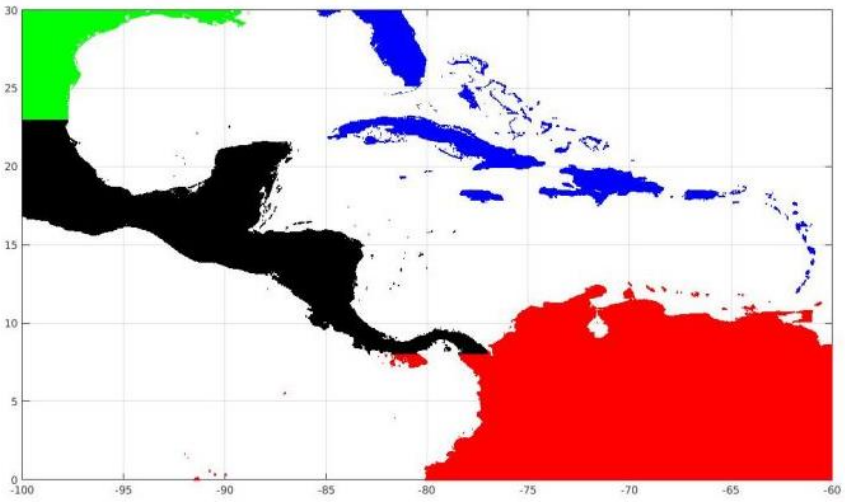


Figure 13: Representation of the different zones.

The algorithm necessary to generate these homogeneous zones is based on the latitude and longitude positions. They are defined different limits for each zone and every observation will be analyzed and compared with those values. When a pixel is defined as part of a specific region, it will be aggregated to the tables corresponding to this region, one for the inputs and the other for the response. The process will be repeated with each of the following observation in the original table. Also, a copy of the table that contains all the observations for the MAC

region is created, and this will also be conserved to compare the models between the entire region and the homogeneous zones and decide which the best approximation is.

After the data are organized in 5 groups, the model identification and parameter estimation are conducted. Two different regression techniques are being involved in this process. Those are:

- Forward Selection Algorithm: This algorithm is one of the stepwise regression methods. Following the definition from Montgomery et al. (2012), this algorithm starts assuming that does not exist any regressor in the model. Then, it looks for the regressor with the largest correlation with the response variable and that also produces the largest F statistic value. If this F value is larger than a previously defined F_{in} value, then the variable is added into the model. now the algorithm identifies the next variable that contributes to best explain the response variable by calculating the corresponding F statistic. Tus, if the statistic is larger than the F_{in} the variable is included into the model. This process will be repeated until the F statistics associated to a given variable does not surpass the defined F_{in} value then, this variable will be discarded and the process finished. Also, this process ends if all the variables have been added to the model.
- Group variable selection algorithm: This technique has been proposed by Ramírez-Beltran el al. (2007) and included by Castro (2007) in his thesis. This methodology starts dividing the number of variables in small groups, usually 5 variables in each group. A regression model is fitted on each group and the t-statistic test is used to identify the significant variables that are selected and named important variables. Then, the important variables from each group will be regrouped and the process starts once again. This process will be repeated until it is reached a number of variables less than or equal of the group size.

It is necessary to test if the models suffer from Multicollinearity problems, referred to input variables that are almost linearly dependent. This usually affect the performance of the model adding unnecessary variable with large coefficients. (Gunst and Webster, 1975 and Montgomery et al., 2012). To solved the multicollinearity problems, a routine based on the variance inflation factor (VIF) was implemented (Montgomery et al. 2012). This routine eliminates values that are large in terms of the VIF. The formula to calculate VIF is showed in the equation (14):

$$VIF = abs(diag((X'X)^{-1})) \tag{14}$$

Where X correspond to a matrix which contains the set of standardized observations for any of the important variables included in the regression model.

Values of VIF will be contrasted to a delimited threshold, Montgomery et al. (2012) recommend that the VIF value should not be higher than 5. When VIF is larger than 5 but smaller than 10, the multicollinearity problem is moderate and sometimes the involved model is acceptable. In this study the threshold has been defined as a value of 5. The largest VIF value is searched and if the value is above the threshold then this variable is discarded because it causes problems of multicollinearity, and the VIF are recalculated without it. The process is repeated until any of the remaining variables exhibit a VIF value above the threshold. When this condition is met, the regression model has been identified and the regression coefficients will be recalculated for this new group of variables. This Methodology was successfully implemented in (Ramírez-Beltran et al. 2007, Castro 2007).

The complete methodology is summarized in the diagram presented in figure 14:

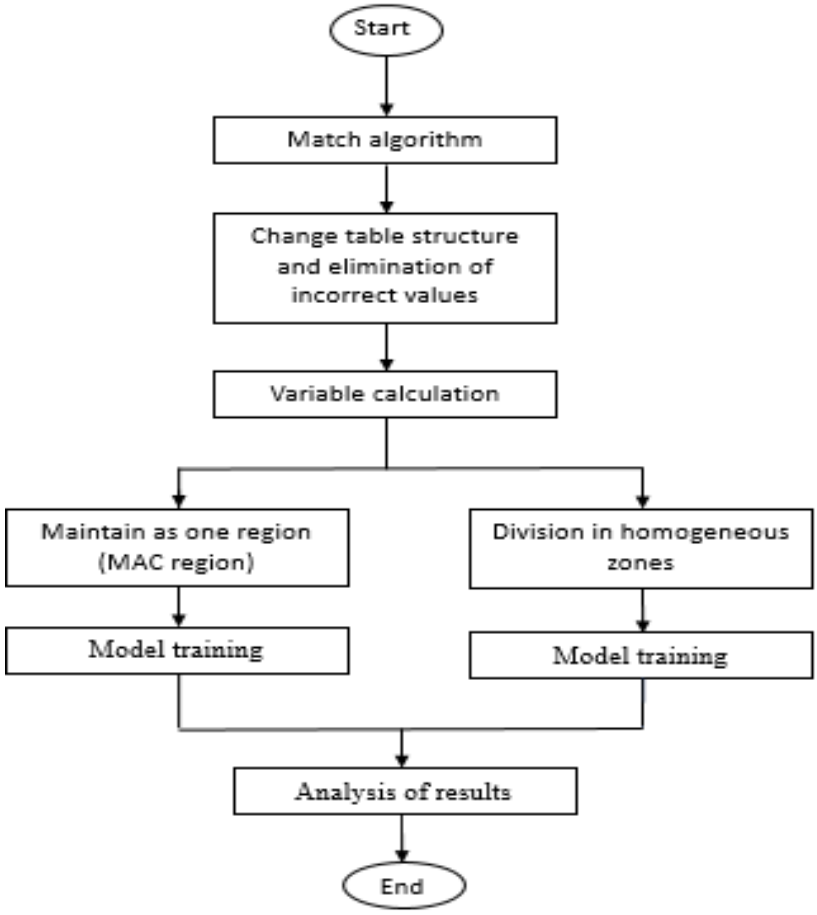


Figure 14: Methodology diagram. Estimation of RH.

6.3 RESULTS

Regression techniques were implemented to derive different models estimate RH based on satellite data, specifically physical parameters gathered from MODIS instrument.

Results are divided in different categories, the methodology employed to build the model and the set of data used to train the models. Results are presented and evaluated in terms of the important variables in the model, the coefficient of determination R^2 and the mean absolute error (MAE). As it was mentioned before models included in this chapter were trained using data from year 2011.

Since the MODIS instrument is installed in two different satellites, each of those provides data with different characteristics. In consequence, a different set of models were developed for each of both, providing different RH products.

6.3.1 ESTIMATION BASED ON MODIS TERRA

In this section, it is presented the results corresponding to the models obtained based on the data gathered from the instrument installed in Terra satellite. Results show the two criteria developed to work with the dataset: one model trained with data from MAC region or a set of independent models trained with data from each of the homogeneous zones previously described. The results for the MAC region are presented in italics to differentiate them from the other results.

Results are presented in two different tables. Each of those correspond to the results from on each of the proposed regression methodologies. Table 4.a contains the results corresponding to the group variable selection technique and Table 4.b the results corresponding to the forward selection technique. Each of those results were obtained after the application of the methodology to eliminate multicollinearity.

Table 4.a: Results from Group variable selection technique

Area	**R^2**	**MAE (%)**	**Important variables**
MAC region	*0.6021*	*9.5188*	*X6, X50, X42, X8, X1, X5, X14*
Antilles	0.7057	7.4584	X6, X51, X35, X32, X2, X4, X5, X1
South America	0.5178	7.2260	X32, X6, X2, X1, X36, X51, X4
Center America	0.6170	8.7334	X6, X7, X32, X3, X48, X40
USA	0.5847	10.8123	X6, X3, X7, X2, X1, X41, X48

Table 4.b: Results from Forward selection technique

Area	R^2	MAE (%)	Important variables
MAC region	*0.6039*	*9.5034*	*X6, X42, X8, X1, X5, X2, X15, X12, X35, X14, X41, X4*
Antilles	0.6957	7.6279	X6, X40, X8, X2, X42, X4, X5, X1, X21, X19
South America	0.5236	7.1816	X32, X6, X2, X1, X15, X21, X14, X27, X4
Center America	0.6190	8.6823	X6, X32, X3, X48, X49, X1, X22, X14, X42
USA	0.5985	10.5968	X6, X3, X5, X8, X4, X2, X1, X14, X27, X48, X21, X12, X15

Results show that the performance of these techniques are very similar in terms of the model fitting. However, the confidence and prediction intervals is narrower for the group variable selection method since the number of variables involved in the models are smaller in this method.

It can be found and average error of about nine percent in the estimation of RH based on this dataset and an R^2 value of nearly 0.6 which can be explained by the low amount of information and the disparity in time and location. Also, it appears that the division in homogeneous zones may result important to obtain better estimates, as it can be appreciated in the values of R^2 and MAE.

In terms of the most significant variables it can be concluded that some variables related to the three physical parameters or their combination, appear as significant. In addition, the time component and the variables related to the position, result also significant in most of the model independent of the methodology to be applied. This indicate the climate associated to the surface characteristics of a given location and seasonal variations help in explaining the behavior of the RH. A particular variable that appear as important is the elevation (X3) that consistently appear as significant, for the areas of Center America and USA areas with complicated topography. It is important to look for the South America region, this region has also a complex topography but with low number of stations, usually distributed in coastal areas, diminishes the influence of the elevation in the regression exercise. It is interesting to notice that the hour (X6) is the most important variable in most of the exercises, it denotes that the time change of the RH is important. The constantly inclusion of the hour as a variable to estimate RH, because it reinforces the importance to implement an hourly model in order to capture this change in a more correct way.

6.3.2 ESTIMATION BASED ON MODIS AQUA

This section presents the results corresponding to the estimation of RH using data gathered MODIS instrument installed in Aqua satellite. The results table provides similar information

that was explained before and are divided in the following structure: Table 5.a provides the results from the group variable selection technique and table 5.b the results from the forward selection techniques.

Once again, the results presented here correspond to the models after the application of the algorithm to solve the multicollinearity problem.

Table 5.a: Results from Group variable selection technique

Area	**R^2**	**MAE (%)**	**Important variables**
MAC region	0.7745	8.8694	X6, X20, X35, X30, X10, X5, X2, X3
Antilles	0.8168	7.2316	X6, X32, X35, X2, X27, X3, X10
South America	0.7029	7.8310	X6, X3, X2, X8, X51
Center America	0.7364	8.9804	X6, X48, X8, X3, X2, X43
USA	0.7721	9.6304	X6, X35, X2, X23, X50, X13, X19, X21

Table 5.b: Results from Forward selection technique

Area	**R^2**	**MAE (%)**	**Important variables**
MAC region	0.7783	8.7925	X6, X8, X15, X1, X27, X14, X5, X4, X35, X2, X3, X12
Antilles	0.8135	7.3385	X6, X15, X2, X3, X42, X8, X5, X47, X14
South America	0.6801	8.2015	X40, X32, X1, X2, X3, X15
Center America	0.7372	8.9385	X6, X8, X3, X48, X33, X2, X1, X42
USA	0.7886	9.1987	X6, X3, X8, X15, X4, X5, X12, X2, X1, X48, X14, X42, X22

It is observed a slightly better result in this group of models compared to the results from MODIS Terra dataset. In average, the R^2 coefficient has augmented in almost 0.1 and the error have been reduced in almost 1%. This result is constant except for the region of South America where the error appears to augments instead to decrease.

The difference in the results from the models trained with MODIS Aqua against the trained with MODIS Terra might be related to the time when satellite cross over the studied area. However, it is difficult to decide whether this dataset or MODIS Terra performs better to estimate RH.

Results in terms of the important variables keeps constant from the observed in MODIS Terra. The physical parameters as well as their combination are still considered as important and also it is observed the importance of the position. Also, elevation and hour (X3, X6) are the variables that appears constantly in most of the models. This effect now it is easy to appreciate for the elevation variable that appear as significant in almost every model developed. Results looks promising and some conclusions are obtained from this exercise.

7. ESTIMATION OF LAND SURFACE TEMPERATURE AND PRECIPITABLE WATER, FROM GOES DATA, USING REGRESSION TECHNIQUES

Based on the conclusions from the previous chapter, it has been determined that is possible to estimate RH from a group of physical parameters LST, PW and NDVI which can be obtained by remote sensing. However, it was also concluded that by the limitations from MODIS the developed models are not capable to offer hourly estimations covering always the studied area.

However, MODIS physical parameters (LST, PW and NDVI) came from a product that, by definition, is obtained in an indirect way, it means that those products are calculated from validated equations and models that uses different channels, in fact in the description of most of MODIS products are included the retrieved algorithm based on raw data (MODIS, 2017). From this logic, it is proposed to develop a new set of equations or models to estimate in hourly basis the physical parameters LST and PW. GOES Imagery data was the selected instrument, because it increases the time resolution and always cover the studied area. This chapter presents a new set of models to estimate LST and PW in hourly basis and further to be able to estimate RH.

7.1 DATA DESCRIPTION

The data required for this group of exercises is mainly focused on the variables necessary to estimate LST and PW variables that are mainly obtained from remote sensing. Specifically, data comes from two different satellite instruments, MODIS and GOES imagery. Like the previous chapter, these variables are divided into two different types: the response variable and the regressor variables. (Montgomery et al., 2012). One year and three months of data are used to perform model training and validation. The training period that goes from December 2010 to November 2011, and the validation period is one month for each rain season and corresponds to December 2011, July 2012 and August 2012.

The response variables are the physical parameters from MODIS: LST and PW. These variables provide the observations that are used to train and validate the model. Those where already preprocessed and have the following characteristics: a spatial resolution of 4 km and a temporal resolution of two observations per day. These observations are disaggregated in a set of images necessary to cover, in portions, the entire MAC region

The regressor variables were mostly obtained from GOES imagery instrument. In addition, some other variables as the elevation and MODIS NDVI were also included.

GOES imagery instrument provides a different set of variables. These variables came from the preprocessed files and they are the brightness temperature (BT) from the channels 2 through 6, all of them are in the infrared domain (see Table 1). The BT have been prepared to have a temporal resolution of one observation per hour and a spatial resolution of 4 km. It is also necessary to include the NDVI from MODIS as a predictor. This product has already been preprocessed and has a spatial resolution of 4 km and it is scaled to hourly estimations. It is proposed to include this product because it is expected that the vegetation chlorophyll level exhibit an indirect measurement of temperature, rain and soil moisture; and therefore, the NDVI is an important predictor to estimate RH.

It has been shown in the previous chapter that the elevation is a relevant regressor variable, and therefore it will be included in the model.

Satellite variables usually provides information for land and oceans areas. However, this study is limited to land cover areas. Therefore, the ocean pixels were eliminated. To achieve this, it has been derived a Mask file based on the elevation model. This mask discriminates between oceans and land covered areas.

To create this mask, it is necessary to evaluate the elevation value for each pixel, and assign a value of 0, if the pixel has an elevation value equal or below 0 m.; otherwise, it is assigned a value of 1. The result is a binary file where values of 1 are linked to land covered areas. A representation of this mask is shown in figure 15.

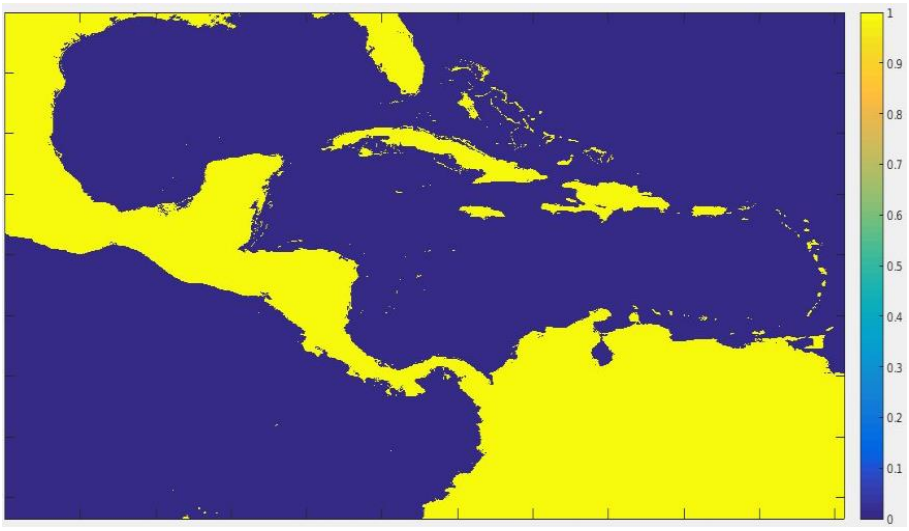


Figure 15: Land covered area Mask. 4 km resolution.

Table 6 shows a summary of the characteristics of the described variables. There are other variables associated to the location and observation time that are going to be included into the model: Latitude, Longitude, Month, Day, and Hour.

Table 6: Characteristics of the data

Product	Source	Spatial resolution	Time resolution
LST (Preprocessed)	MODIS instrument.	4 km	2 times per day
PW (Preprocessed)	MODIS instrument.	4 km	2 times per day
NDVI (preprocessed)	MODIS instrument.	4 km	Hourly
GOES Channels 2 to 6 Brightness temperature (Preprocessed)	GOES imagery.	4 km	Hourly
Elevation	DEM.	4km	N.A.*

* Does not apply for this product.

Datasets presented in those models correspond one year of observations. The time necessary to preprocess the data and to develop the time series is one of the issues that requires to be optimized. It has already been proposed that data might be segregated in homogenous groups to improve the regression and reduce the processing time, it has already been implemented dividing the set of observations based on their geographic position. Now, a second division is developed, complementing the already established one, this division now divides the dataset in terms of climatological periods. This division is based on the work from this investigation group that has observed that, in general, during the year the climatic behavior of the MAC region could be divided in three rain seasons and those are:

- Dry season: that correspond to the months of December, January, February and March.
- Early rain season: that correspond to the months of April, May, June and July.
- Late rain Season: That correspond to the months of August, September, October and November.

7.2 METHODOLOGY

This methodology describes the main steps that are necessary to develop the estimation models. These steps are mainly focused on the data preprocessing and processing, and the development and validation of the models. There is necessary to develop a group of models to estimate both PW and LST and depending on which will be estimated the algorithms may have some minor changes. These algorithms will be explained in details in the following lines.

7.2.1 MATCH ALGORITHM

This code represents the beginning of the estimation process for the two studied physical parameters, LST and PW. It is responsible for the matching process between MODIS physical parameters that provides the observations to train the model and the files of input variables: BT, NDVI and elevation. All those files have already been preprocessed. There are 4 variations of this code, one for each physical parameter that will be estimated (LST and PW) and one for the satellite where they come from (Aqua or Terra). However, each of those follow the same procedure.

This algorithm starts opening the set of observations from the response variables and saving the corresponding time where they were captured. The same will be done for the set of observations from GOES brightness temperature and for MODIS NDVI. And based on this information the match algorithms are developed:

- Based on the time from each MODIS physical parameter observation, the nearest GOES and MODIS NDVI observations will be searched. If no file from MODIS NDVI or GOES is found, then the physical parameter observation is discarded and the next one is examined. On the other hand, if both corresponding images are found, then both and the physical parameter observation are matched in terms of time and they will be saved and analyzed.

- GOES is a geostationary satellite and their images have usually the same limits, that were defined when the file was downloaded. The limits are latitude: 0 to 30, and longitude: -100 to -60. With this information, the MODIS files will be evaluated and only the pixels that are inside these limits will be kept, every other pixel will be eliminated. This searching has an extra step for the PW algorithm. This product does not discriminate between land and ocean and it is necessary to eliminate sea pixels. For PW, it is necessary to open the mask, and search the nearest pixel to every pixel in the PW observation. If the nearest pixel in the mask has a value related to a land covered area, then this pixel will be kept otherwise it will be eliminated. All those pixels that were not eliminated will be then compared to the observation from GOES and NDVI searching for the nearest pixels in those images. In addition, it is necessary to look for the nearest pixel in the elevation file.

- Once the nearest pixels were found, they are saved in a table that contains the following columns: the time for the observation where this pixel came from, their Latitude,

Longitude, and the value for the physical parameter, BT for each individual channel, NDVI observation and their corresponding elevation. If the pixel does not have a value in any of the variables it will be filled with a missing value. This process will be repeated for all the pixels in the observation and for every MODIS observation in the dataset.

This table will be saved to be introduced in the next algorithm. There are 4 different tables, each one came from the variations of this algorithm, which was explained above.

7.2.2 STRUCTURE AND CLEANING ALGORITHM

This algorithm is the responsible for the structure and arrange of the table data and for cleaning the variables. It starts reading the table previously saved. This table has already been saved with the necessary structure for the estimation process, dividing the variables, which are in the columns and the observations in the rows.

However, the data previously saved have some missing and incorrect values that needs to be fixed to improve the results of the regressions. The algorithm to eliminate missing values and to clean the data are explained bellow:

- The table is loaded and every row is analyzed. If a row has at least one missing value in any of their variables (columns), then the entire row is eliminated from the table. Any value that do not correspond to clear condition. Were in this work clear sky condition was based on BT from channel 4. A pixel values bellow 280 Kelvin degrees was considered a cloudy or rainy event and the entire row was eliminated from the file.
- A similar procedure was developed to eliminate incorrect values. To define what an incorrect value is, thresholds have been defined for each variable based on the dataset. In consequence, every pixel that have at least one variable with an associated value that is outside their correspondent threshold will be eliminated. The limits defined for each variable were: 260 to 330 Kelvin degrees for the BT of channel 2, 235 to 270 Kelvin degrees for the BT of channel 3, 280 to 330 Kelvin degrees for the BT of channel 4, 240 to 290 Kelvin degrees for the BT of the channel 6, 0 to 320 Kelvin degrees for the LST, 0 to 1 for NDVI.

A last arrangement process is performed over the dataset. To obtain homogeneous dataset and to reduce the processing time in the regression, the entire year of data have been divided in the three seasons previously described: Dry, Early rain and Late rain season. It is necessary to

separate the original 4 datasets and split each of them into three different ones that correspond to each of the rainy seasons.

7.2.3 DATA PROCESSING

The data processing algorithm is focused on: the organization of the matrix variables, the inclusion of some variables that are based on the combinations of variables, the division of dataset, which includes the homogeneous regions and the application of the estimation techniques. It starts opening the table saved on the previous algorithm that corresponds to either the dry, early or late rain season. This table has 4 different variations in terms of the response variable: LST or PW from MODIS Aqua or Terra.

Then the dataset is divided in homogeneous zones. The original table is divided into 4 matrices one per zone and the complete table is kept to test also the entire MAC region as a single mode. The division criteria base their selection from the latitude and longitude values. After that, each table will be divided in two different tables "y" that contains only the response variable (called Y) and "x" that contain all the predictors (called X's) created so far. The predictors and response variables are described in the table 7.a. The X value correspond to the value in the regression variables and not to their position in the table.

Table 7.a: Description of the variables

Variable	Description	Variable	Description
Y	LST or PW	X_6	BT4
X_1	Month	X_7	BT6
X_2	Day	X_{14}	NDVI
X_3	Hour	X_{15}	Elevation
X_4	BT2	X_{16}	Latitude
X_5	BT3	X_{17}	Longitude

The next step is to generate some other variables that will be included as input variables. Those variables are related to the difference between two different BT's. These variables are included to explain a higher level of variability and because the effect of some channels can be enhanced when they are combined with another channel. Thus, there is a total of 17 predictors and Table 7.b shows the used differences of BT as input variables. The literature shows the application of the differences of GOES BT. For instance, Ba and Gruber (2001) and Kuligowski (2002) used the difference of GOES BT to estimate rainfall. Ramírez-Beltran et al. (2009) also used difference of GOES BT to detect rainy clouds.

Table 7.b: Description of the variables

Variable	Description
X_8	BT2-BT3
X_9	BT2-BT4
X_{10}	BT2-BT6
X_{11}	BT3-BT4
X_{12}	BT3-BT4
X_{13}	BT4-BT6

The group variable selection and forward selection techniques were used to identify the variables that best explain the response variable. Those model identification techniques were employed to generate a widely groups of models: one model for the MAC region and a model for each individual homogeneous zone. All those 5 models will be independently created for each of the 3 season delimited during the training period, and Furthermore, all the models will be generated with each individual variation of the dataset: LST and PW from MODIS Terra or Aqua. The identified models and their performances are discussed in section of results

7.2.4 MODEL EVALUATION

It should be noted that the models were developed using PW and LST observed by MODIS only twice a day. Now the regression equations will be evaluated for every hour and for all the pixels included in the MAC region, with the purpose of estimate the LST and PW in places where there were not MODIS data, and it is referred this process as the model evaluation. The output will be maps filled with the estimations of PW and LST and these values will be used as the input variables to estimate RH. This task will be described in chapter 8.

The algorithm for model evaluation is as follow:

- The algorithm start loading the set of observations corresponding to each of the input variables and for the entire studied period. Also, the coefficients and the parameters from the best model for each region and for each season are loaded.
- Then, every image will be studied and the time will be extracted from the corresponding file. The images from each of the input variables that correspond to the same time will be linked. If at the selected hour there is one or more missing value in at least one input variables, the information of this particular hour is discarded.
- If all the images corresponding to the inputs for a specific hour are found, it is necessary to open the files and to study every pixel inside of them. The position variables are extracted and the nearest pixel from each pixel is searched, once they are found it is necessary to matched them with each other. Once they are matched, the regression

equation is evaluated. In contrast if at least one pixel has a NaN value linked to it, then the value of their corresponding physical parameter (LST and PW) for this pixel at this time will be a NaN. The regression equation to estimate LST or PW are different depending on the region and the season, it is necessary to study those parameters to select the appropriate model. This step will be repeated for all the pixels included in the dataset.

- Once all the pixel for a specific hour have the corresponding estimated of LST and PW then the corresponding map are created for these products. This map file is a combination of three tables one correspond to the Latitude a second one to the Longitude and a third one has the estimated values of LST or PW for each pixel. This process will be repeated for every hour during the studied period.

This evaluation process was performed for the training period, December 2010 to November 2011. In addition, the algorithm generates the required image files for the months assigned to validation: December 2011, July 2012 and August 2012.

7.2.5 VALIDATION

The dataset is divided into data for training and for validation. The data for training is used to develop the regression equations and the validation set is used to test how well or bad is the estimation. In this exercise 4 months (from 2011) were used for training in each of the rainy season and one month (from 2012) for validation.

A new algorithm is going to be developed to implement the validations for each of the physical parameters. This algorithm requires two set of images: the images of the observed physical parameters (extracted from MODIS), and the estimations images developed by model evaluation. The algorithm is described as follows:

- First each observed physical parameter (MODIS) image is analyzed and its corresponding time is studied, linking this image with the nearest image from the estimation set. The estimations are generated every hour, but the observations (twice a day) may come in intervals inside of an hour it means that if it exists an image of MODIS at the 4:15 pm, it will be linked to the estimated at 4 pm.
- Once both images are linked it is necessary to eliminate from the observation image every pixel that is outside of the limits of GOES (estimation limits), then for every pixel inside the limits is necessary to search for the nearest one in the image of estimated

products. The difference between observations and estimations are the validation error and is calculated for every pixel that is found in both images. This process will be repeated for every image in the studied period.

- Finally, the validation accuracy scores are computed and they are: the RMSE, the MAE, the error rate and the R^2 coefficient.

In figure 16, it is presented a diagram that summarizes the methodology employed in this stage:

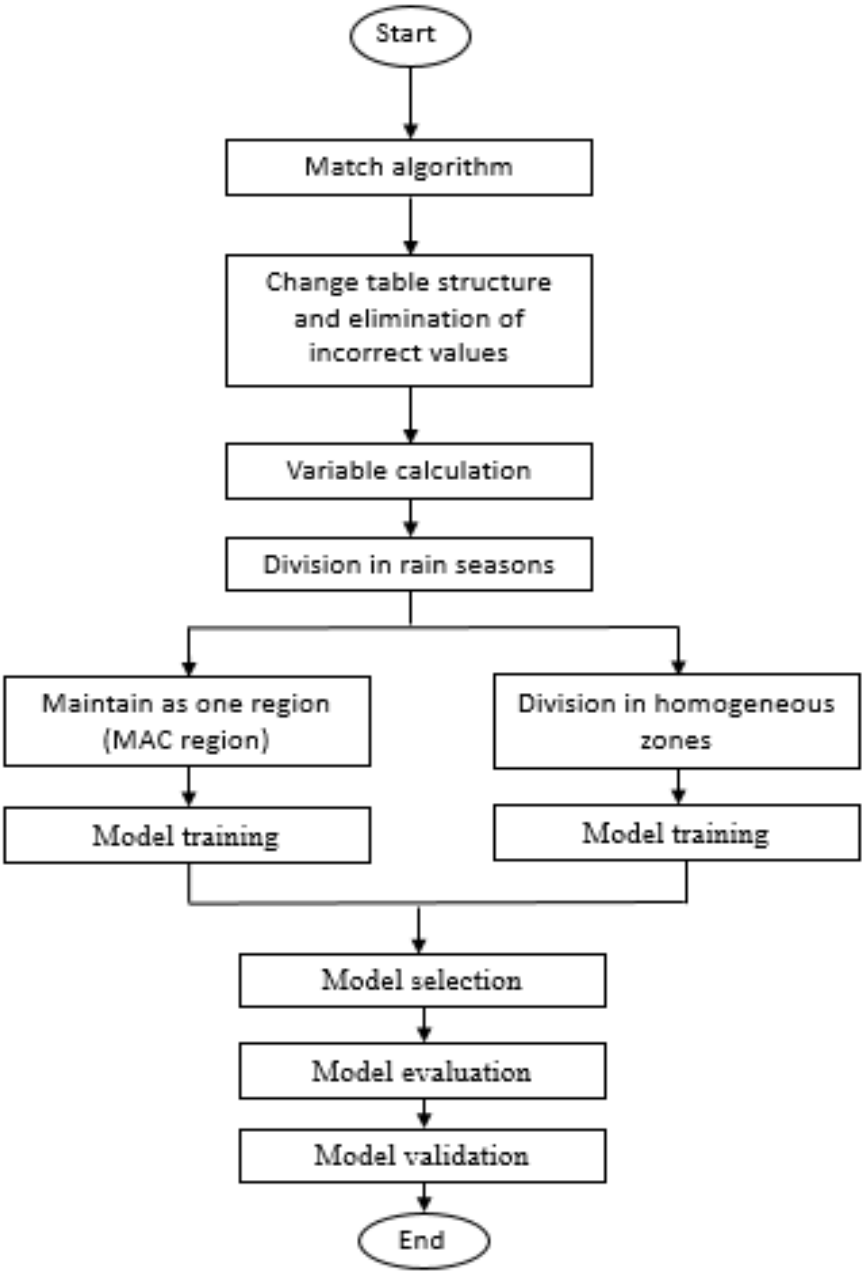


Figure 16: Methodology diagram. Estimation of LST and PW.

7.3 RESULTS

Results were organized by parameter (LST and PW) and by satellite (Terra, Aqua) and are presented in this four sections the next section includes model evaluation results and the last section describes the model validation results.

7.3.1 PW-MODIS AQUA

Results presented here correspond to the models created to estimate PW. The model was trained using MODIS data of the Aqua satellite. Results were obtained from the application of two different regression techniques. Five models were developed one associated with the MAC region which are given in italic letters, and four models that correspond to homogeneous regions.

Results are presented in 3 different tables: table 8 corresponds to the results for the dry season model, table for the early rain season model and table 10 for the late rain season model. Each of those tables have 3 sub tables: the results for the Group variable selection technique correspond to the sub table a., the results for the Forward selection technique corresponds to the sub table b. and the results corresponding to the error rates of both methodologies, are given in sub table c. The definition of the error value presented in the tables, is the ratio of the mean absolute error divided by the magnitude of the observed value, see appendix 2.

Table 8.a: Results Dry season from Group variable selection technique - PW

Area	R^2	MAE (cm)	Important variables
MAC region	0.6792	0.5669	X16, X15, X5, X9, X3, X6, X17, X2, X1
Antilles	0.4666	0.4239	X16, X8, X4, X15, X1, X3, X9, X17, X2
South America	0.5920	0.6719	X5, X15, X3, X16, X7, X1, X2, X17, X9
Center America	0.6044	0.4555	X15, X17, X5, X3, X1, X7, X2, X16, X9
USA	0.4353	0.3658	X5, X15, X16, X9, X3, X6, X17, X1, X2

Table 8.b: Results Dry season from Forward selection technique - PW

Area	R^2	MAE (cm)	Important variables
MAC region	0.6841	0.5608	X16, X15, X5, X9, X14, X3, X12, X17, X1, X2
Antilles	0.4467	0.4301	X16, X6, X15, X1, X3, X12, X17, X14, X2
South America	0.5929	0.6715	X5, X15, X3, X16, X12, X1, X2, X14, X17, X9
Center America	0.6070	0.4538	X15, X17, X5, X3, X1, X7, X2, X16, X14, X9
USA	0.4385	0.3649	X5, X15, X16, X9, X3, X6, X17, X1, X2, X14

Table 8.c: Dry season: Error rate - PW

Area	Error rate (Group variable selection) (%)	Error rate (Forward selection) (%)
MAC region	6.09	6.03
Antilles	6.49	6.58
South America	7.36	7.35
Center America	6.02	6.00
USA	6.37	6.36

Table 9.a: Results Early rain season from Group variable Selection technique - PW

Area	R^2	MAE (cm)	Important variables
MAC region	0.6981	0.5828	X17, X5, X15, X9, X1, X16, X2, X4, X3
Antilles	0.6576	0.5121	X1, X3, X7, X16, X2, X5, X15, X9, X17
South America	0.4228	0.6287	X15, X5, X3, X16, X1, X7, X17, X2, X8
Center America	0.6569	0.5126	X15, X1, X5, X17, X9, X16, X2, X3, X6
USA	0.6495	0.4205	X1, X5, X9, X15, X2, X17, X16, X6

Table 9.b: Results Early rain season from Forward selection technique - PW

Area	R^2	MAE (cm)	Important variables
MAC region	0.6991	0.5816	X17, X5, X15, X9, X1, X16, X2, X7, X3, X14, X13
Antilles	0.6476	0.5227	X1, X3, X7, X16, X2, X15, X17, X13, X9, X14
South America	0.4237	0.6282	X15, X5, X3, X16, X1, X12, X17, X2, X14, X13
Center America	0.6590	0.5116	X15, X1, X5, X17, X9, X14, X16, X2, X3, X6
USA	0.6497	0.4204	X1, X5, X9, X15, X2, X17, X16, X8, X14

Table 9.c: Early rain season: Error rate - PW

Area	Error rate (Group variable selection) (%)	Error rate (Forward selection) (%)
MAC region	6.13	6.11
Antilles	5.75	5.87
South America	6.98	6.97
Center America	5.49	5.48
USA	4.96	4.96

Table 10.a: Results Late rain season from Group variable Selection technique - PW

Area	R^2	MAE (cm)	Important variables
MAC region	0.7100	0.6296	X5, X17, X15, X9, X16, X1, X3, X6, X2
Antilles	0.7247	0.5459	X8, X16, X1, X6, X15, X17, X2, X9, X3
South America	0.4823	0.5727	X15, X5, X3, X1, X4, X16, X17, X2, X9
Center America	0.7452	0.5867	X5, X15, X9, X17, X2, X6, X3, X1, X16
USA	0.7165	0.4187	X1, X5, X9, X15, X16, X17, X2, X7

Table 10.b: Results Late rain season from Forward selection technique - PW

Area	R^2	MAE (cm)	Important variables
MAC region	0.7172	0.6219	X5, X17, X15, X9, X16, X14, X1, X3, X2, X7
Antilles	0.7285	0.5408	X8, X16, X1, X7, X15, X17, X2, X14, X13, X3
South America	0.4860	0.5705	X15, X5, X3, X1, X12, X14, X16, X17, X2, X13
Center America	0.7466	0.5848	X5, X15, X9, X17, X13, X2, X3, X1, X16, X14
USA	0.7115	0.4230	X1, X5, X9, X15, X16, X17, X2, X11, X14

Table 10.c: Late rain season: Error rate - PW

Area	Error rate (Group variable selection) (%)	Error rate (Forward selection) (%)
MAC region	6.39	6.31
Antilles	6.32	6.26
South America	6.00	5.98
Center America	6.54	6.52
USA	4.82	4.87

Results obtained from both techniques are pretty comparable, the difference between them is not too big. It is difficult to determine which technique offer the best estimation, but to decide for the model that will be used to estimate the product in this exercise it has been decided to select the alternative that provides the best results in terms of the performance metrics. This alternative will be selected individually for every single region and could be different from region to region. For example, the selected estimation techniques for the dry season are: forward selection techniques (South America, Antilles, Center America) group variable selection technique (USA).

In terms of the important variables, it is appreciated the importance of the time and position variables. Also, The GOES variables specially the difference between BT2-BT4 appear as important in every single model except for some models on the area of South America during the Early and late rain season and the Antilleans in the Late rain seasons, where they were replaced by some other differences. NDVI also became important in explain the variability of PW.

South America exhibits the hardest region to estimate PW, offering the lowest value in terms of their R^2 coefficients for most of the seasons, and shows one of the biggest errors from the four regions.

In terms of the R^2, it appears that the best model fit of PW was obtained for the late rain season; however, in terms of the error the best estimation was obtained for the dry season which even that provide the lowest R^2 values also have the lowest errors, it might be explained because the PW values in this seasons are constantly lower compared to the other seasons.

To decide whether estimate PW using one model for the MAC region or to select for the homogeneous regions, it was found that the best estimation was obtained using models for each region. In terms of the R^2 values is hard to obtain a clear decision; however, it becomes clearer when looking the error values. Error values are constantly lower for each of the homogeneous zones compared with the model for the entire MAC region, except for the area of South

America, but this area is problematic for estimating PW. Thus, PW is best estimated by using a different model for each of the 4 homogeneous zones. Similar results were observed in terms of the error rate, it was noted that the estimation errors became smaller when divided in regions compared to the total MAC region, one exception is noticed in dry season where MAC region appear to be with less variability in terms of the error rate.

7.3.2 PW-MODIS TERRA

Similar results are presented based on observation of MODIS Terra. The statistical models were evaluated with the same performance metrics. Results are organized and presented in three tables.

Table 11 shows results of dry season, table 12 of Early rain season and table 13 of Late rain season. These tables are divided in three sub tables: a. corresponds to results for the models based on group variable selection technique, b. presents results for the models based on forward estimation techniques and c. shows the error rate for both techniques.

Table 11.a: Results Dry season from Group variable Selection technique - PW

Area	R^2	MAE (cm)	Important variables
MAC region	0.6759	0.4848	X16, X15, X5, X9, X17, X7, X2, X1, X3
Antilles	0.4012	0.3859	X16, X8, X15, X1, X17, X2, X3, X7
South America	0.5964	0.5911	X5, X15, X16, X1, X2, X9, X17, X3
Center America	0.6283	0.3777	X15, X17, X5, X1, X3, X7, X2, X16
USA	0.4973	0.3053	X5, X15, X16, X7, X3, X17, X9, X2, X1

Table 11.b: Results Dry season from Forward selection technique - PW

Area	R^2	MAE (cm)	Important variables
MAC region	0.6751	0.4844	X16, X15, X5, X9, X17, X8, X14, X2, X1, X3
Antilles	0.4251	0.3742	X16, X12, X15, X9, X1, X7, X17, X3, X2, X14
South America	0.5843	0.6027	X5, X15, X17, X1, X2, X14, X13, X17, X3
Center America	0.6314	0.3748	X15, X17, X5, X1, X3, X7, X2, X16, X13, X14
USA	0.5071	0.3028	X15, X16, X3, X9, X17, X11, X5, X2, X1, X14

Table 11.c: Dry season: Error rate - PW

Area	Error rate (Group variable selection) (%)	Error rate (Forward selection) (%)
MAC region	6.27	6.27
Antilles	6.62	6.42
South America	8.15	8.31
Center America	4.89	4.85
USA	6.35	6.30

Table 12.a: Results Early rain season from Group variable selection technique - PW

Area	R^2	MAE (cm)	Important variables
MAC region	0.6852	0.5283	X17, X1, X15, X5, X9, X16, X2, X3, X7
Antilles	0.6894	0.4869	X1, X3, X2, X7, X16, X15, X5, X9, X17
South America	0.4670	0.5881	X15, X5, X3, X16, X7, X1, X17, X2, X9
Center America	0.6738	0.4838	X15, X5, X1, X16, X3, X7, X2, X9, X16
USA	0.7146	0.4035	X1, X9, X15, X5, X2, X7, X16, X17

Table 12.b: Results Early season from Forward technique - PW

Area	R^2	MAE (cm)	Important variables
MAC region	0.6856	0.5276	X17, X1, X15, X5, X9, X16, X2, X3, X12, X14
Antilles	0.6887	0.4866	X1, X3, X2, X7, X16, X15, X17, X14, X11
South America	0.4707	0.5863	X15, X5, X3, X16, X12, X10, X1, X17, X14, X2
Center America	0.6738	0.4837	X15, X5, X1, X17, X3, X7, X2, X9, X15, X14
USA	0.7158	0.4028	X1, X9, X15, X5, X2, X7, X16, X17, X14

Table 12.c: Early rain season: Error rate - PW

Area	Error rate (Group variable selection) (%)	Error rate (Forward selection) (%)
MAC region	5.88	5.87
Antilles	6.05	6.05
South America	8.59	8.57
Center America	5.93	5.93
USA	4.49	4.48

Table 13.a: Results Late rain season from Group variable selection technique -PW

Area	R^2	MAE (cm)	Important variables
MAC region	0.7120	0.5991	X5, X15, X16, X9, X1, X17, X6, X2, X3
Antilles	0.6658	0.6324	X1, X16, X15, X6, X17, X10, X2, X3
South America	0.5329	0.5376	X15, X8, X7, X3, X1, X16, X9, X2, X17
Center America	0.7510	0.5304	X5, X15, X17, X7, X2, X1, X13, X14, X3
USA	0.7695	0.4212	X1, X5, X9, X15, X3, X16, X17, X2

Table 13.b: Results Late rain season from Forward selection technique - PW

Area	R^2	MAE (cm)	Important variables
MAC region	0.7161	0.5946	X5, X15, X16, X9, X1, X14, X17, X2, X7, X3
Antilles	0.7387	0.5460	X8, X1, X16, X7, X15, X17, X9, X2, X3, X14
South America	0.5365	0.5373	X15, X5, X10, X3, X1, X16, X14, X13, X2, X17
Center America	0.7510	0.5304	X5, X15, X17, X7, X2, X1, X13, X14, X3
USA	0.7467	0.4480	X1, X5, X9, X15, X16, X17, X2, X13, X14

Table 13.c: Late rain season: Error rate - PW

Area	Error rate (Group variable selection) (%)	Error rate (Forward selection) (%)
MAC region	7.06	7.01
Antilles	7.81	6.74
South America	6.95	6.95
Center America	6.80	6.80
USA	5.17	5.50

Results provided by the two methodologies are similar as were described in previous section.

In terms of the important variable it is similar to the previous exercises the variables related to the time and position resulted important in every single model. Also, the difference between some BT (BT2-BT4) appear to be important. However, this group of difference are subdued by a new variable that appear important in many of the models this is the BT6. The NDVI is still being considered as an important variable.

It is observed that in terms of the results estimations obtained from satellite data provides a good set of approximations, however there are some region more problematic than others. South America, for example, is constantly providing a quite lower level of fit compared to the other regions. This effect is quite understandable looking the complexity of this region and its location compared to the other areas.

To divide the region in homogeneous areas provided a good set of estimations compared to one single model to represent the complete MAC region, this is also observed here and it also helps to reduce the computational time and to validate the model.

In summary, the model from different seasons are similar than the observed from MODIS Aqua models, the best R^2 coefficients were obtained from the early rain season; however, the minimum error values were observed in the dry season, and it can be explained by the lowest values of PW during this period.

7.3.3 LST-MODIS TERRA

This section presents the results corresponding to the estimation of LST based on MODIS data of Terra satellite. This section provides the best variables and the performance metrics to evaluate the developed modes. The Group variable selection and the Forward selection techniques were applied over a specific dataset, which could be the data corresponding to the entire MAC region or the data corresponding to each of the defined homogeneous regions.

Results are divided in three tables: table 14 correspond to the models for dry season, table 15 for early rain season and table 16 for late rain season. Each table are divided in three sub tables that follows the same structure defined for the previous models.

Table 14.a: Results Dry season from Group variable selection technique - LST

Area	R^2	MAE (Kelvin)	Important variables
MAC region	0.7055	3.1905	X15, X3, X16, X14, X10, X1, X17, X12, X13
Antilles	0.7526	3.2015	X16, X3, X14, X10, X1, X15, X13, X12, X17
South America	0.7834	2.0042	X15, X3, X14, X1, X11, X17, X10
Center America	0.7857	2.8037	X3, X15, X14, X7, X1, X16, X2, X11, X17
USA	0.7429	3.1575	X3, X15, X9, X16, X2, X14, X11, X17

Table 14.b: Results Dry season from Forward selection technique - LST

Area	R^2	MAE (Kelvin)	Important variables
MAC region	0.6928	3.2650	X15, X3, X16, X14, X1, X2, X17, X11, X9
Antilles	0.7568	3.1786	X16, X3, X14, X10, X1, X15, X2, X13, X12, X17
South America	0.7830	2.0018	X15, X3, X14, X1, X17, X13, X16, X11
Center America	0.7897	2.7809	X3, X15, X14, X6, X16, X1, X2, X11, X10, X17
USA	0.7373	3.1795	X3, X15, X16, X11, X2, X14, X17, X1

Table 14.c: Dry season: Error rate - LST

Area	Error rate (Group variable selection) (%)	Error rate (Forward selection) (%)
MAC region	5.58	5.71
Antilles	6.03	5.99
South America	3.64	3.63
Center America	4.91	4.87
USA	6.42	6.46

Table 15.a: Results Early rain season from Group variable selection technique - LST

Area	R^2	MAE (Kelvin)	Important variables
MAC region	0.7224	3.0489	X3, X14, X15, X8, X16, X2, X17, X1
Antilles	0.8050	2.6526	X3, X14, X7, X15, X16, X10, X2, X1, X17, X11
South America	0.6623	2.0991	X4, X15, X14, X3, X16, X1, X17, X2, X9
Center America	0.7034	2.9315	X3, X14, X15, X6, X5, X2, X10, X16, X1
USA	0.8437	2.5406	X4, X14, X1, X15, X2, X16, X9

Table 15.b: Results Early season from Forward selection technique - LST

Area	R^2	MAE (Kelvin)	Important variables
MAC region	0.7423	2.9432	X3, X14, X15, X5, X16, X2, X17, X7, X1, X13
Antilles	0.8025	2.6789	X3, X14, X7, X15, X16, X2, X1, X17, X9
South America	0.6631	2.0967	X4, X15, X14, X3, X16, X1, X17, X12, X2
Center America	0.7030	2.9344	X3, X14, X15, X7, X2, X12, X16, X17, X1, X13
USA	0.8125	2.8092	X14, X1, X15, X2, X17, X16, X9, X7, X5

Table 15.c: Early rain season: Error rate - LST

Area	Error rate (Group variable selection) (%)	Error rate (Forward selection) (%)
MAC region	4.87	4.70
Antilles	4.24	4.28
South America	4.35	4.35
Center America	4.69	4.70
USA	5.60	6.20

Table 16.a: Results Late rain season from Group variable selection technique - LST

Area	R^2	MAE (Kelvin)	Important variables
MAC region	0.7521	2.6967	X15, X3, X1, X14, X16, X17, X2, X12, X13
Antilles	0.8176	2.1823	X16, X15, X1, X3, X14, X13, X17, X5
South America	0.5866	1.8004	X4, X15, X14, X3, X16, X9, X2, X17, X8
Center America	0.7380	2.4044	X3, X14, X15, X1, X2, X13, X16, X17, X12
USA	0.8610	2.9316	X1, X15, X16, X14, X3, X11, X2, X9

Table 16.b: Late rain season from Forward selection technique - LST

Area	R^2	MAE (Kelvin)	Important variables
MAC region	0.7370	2.7780	X15, X13, X1, X14, X16, X17, X2, X10, X12
Antilles	0.8182	2.1814	X16, X15, X1, X3, X14, X13, X11, X17, X2
South America	0.5866	1.8001	X4, X15, X14, X3, X16, X9, X2, X17, X1, X11
Center America	0.7470	2.3658	X3, X14, X15, X1, X6, X2, X9, X16, X17, X12
USA	0.8353	3.2115	X1, X15, X16, X14, X12, X3, X2, X17, X9

Table 16.c: Late rain season: Error rate - LST

Area	Error rate (Group variable selection) (%)	Error rate (Forward selection) (%)
MAC region	4.95	5.09
Antilles	4.55	4.55
South America	3.66	3.66
Center America	4.46	4.39
USA	5.38	5.89

Results follow the same pattern observed in previous exercises both techniques provide good estimations of LST and it is hard to determine a winner technique from those two. In this step the technique selected to provide the estimation will be defined by the winner from the two methodologies on every season and for every region.

It is observed that the time and position components are significant in every model, these results are consistent with results observed in previous models. Variables obtained from GOES are also important variables, but in this time, they are not a single variable that capitalize the importance, instead depending on the region and the season a different group of variables appear as significant variables. NDVI also resulted an important variable in the models to estimate LST as it was theorized before.

It is observed in terms of the performance metrics that observations of LST are estimated better when they are divided in homogeneous areas than using a single model to estimate the entire MAC region. This is more noticeable in some season but the improvements are even more valuable specially on reducing the processing time.

Different from PW this models provides a better set of estimations in terms of the error specially looking the error rate that is lower than 7% in each model. It appears to be easier to estimate LST compared to the estimation of PW.

Two models were compared to estimate LST one model to estimate the entire MAC region and several models for the homogeneous climatic zones. Based on the results both R^2 and error are improved when the MAC region is divided.

7.3.4 LST-MODIS AQUA

This section present the results for estimating LST based on MODIS Aqua satellite data.

Results are presented in three tables: table 17 for the models corresponding to the dry season, table 18 for the early rain season and table 19 for the late rain season.

Table 17.a: Results Dry season from Group variable selection technique - LST

Area	R^2	MAE (Kelvin)	Important variables
MAC region	0.7499	3.6585	X3, X15, X16, X14, X1, X17, X12, X13
Antilles	0.7873	3.4658	X3, X7, X16, X1, X14, X15, X2, X9, X12
South America	0.7864	2.3954	X15, X3, X14, X1, X17, X12, X16, X13
Center America	0.8198	3.3210	X3, X14, X15, X16, X1, X2, X9, X7, X11
USA	0.8726	2.6421	X14, X6, X15, X16, X12, X3, X2, X1

Table 17.b: Results Dry season from Forward selection technique - LST

Area	R^2	MAE (Kelvin)	Important variables
MAC region	0.7414	3.7305	X3, X15, X16, X14, X10, X1, X17, X2, X11
Antilles	0.8080	3.3147	X3, X7, X16, X1, X14, X15, X2, X8, X10, X17
South America	0.7850	2.4103	X15, X3, X14, X1, X17, X8, X2, X16, X9
Center America	0.8049	3.4334	X3, X14, X15, X16, X1, X2, X8, X9, X17
USA	0.8729	2.6424	X14, X6, X15, X16, X12, X3, X2, X17, X1

Table 17.c: Dry season: Error rate - LST

Area	Error rate (Group variable selection) (%)	Error rate (Forward selection) (%)
MAC region	4.96	5.06
Antilles	5.41	5.17
South America	3.70	3.72
Center America	4.50	4.66
USA	4.17	4.17

Table 18.a: Results Early rain season from Group variable selection technique - LST

Area	R^2	MAE (Kelvin)	Important variables
MAC region	0.8187	3.2663	X3, X14, X15, X5, X13, X17, X2, X1
Antilles	0.8636	2.8245	X3, X14, X7, X9, X16, X15, X13, X2, X12, X17
South America	0.5725	2.1139	X15, X14, X3, X16, X1, X17, X13, X12, X9
Center America	0.8091	3.4921	X3, X14, X15, X9, X12, X1, X13, X17, X16
USA	0.8850	3.0090	X3, X14, X1, X15, X2, X17, X16, X11

Table 18.b: Results Early rain season from Forward selection technique - LST

Area	R^2	MAE (Kelvin)	Important variables
MAC region	0.8172	3.2983	X3, X14, X15, X13, X17, X16, X2, X12, X1
Antilles	0.8592	2.8719	X3, X14, X16, X15, X2, X1, X10, X11, X17
South America	0.5807	2.0961	X4, X15, X14, X3, X16, X1, X2, X17, X12
Center America	0.8089	3.4969	X3, X14, X15, X1, X17, X2, X16, X13, X9
USA	0.9010	2.6561	X4, X14, X1, X15, X2, X17, X9, X16, X12

Table 18.c: Early rain season: Error rate - LST

Area	Error rate (Group variable selection) (%)	Error rate (Forward selection) (%)
MAC region	4.36	4.40
Antilles	4.3	4.46
South America	3.63	3.60
Center America	4.75	4.76
USA	4.82	4.25

Table 19.a: Results Late rain season from Group variable selection technique - LST

Area	R^2	MAE (Kelvin)	Important variables
MAC region	0.7845	3.0891	X15, X3, X14, X1, X16, X17, X11, X9
Antilles	0.8414	2.3085	X16, X3, X1, X14, X15, X11, X13, X17
South America	0.5576	2.0897	X3, X15, X14, X4, X8, X16, X1, X2, X17, X13
Center America	0.7810	2.7721	X3, X15, X14, X1, X2, X13, X17, X16, X12
USA	0.9134	2.5189	X1, X3, X14, X15, X16, X2, X13, X5

Table 19.b: Results Late rain season from Forward selection technique - LST

Area	R^2	MAE (Kelvin)	Important variables
MAC region	0.7968	2.9957	X15, X3, X14, X1, X16, X17, X2, X11, X10
Antilles	0.8589	2.1609	X4, X16, X3, X1, X14, X15, X2, X12, X17
South America	0.5576	2.0898	X3, X15, X14, X4, X8, X16, X1, X2, X17, X13, X9
Center America	0.7848	2.7497	X3, X15, X14, X4, X1, X2, X9, X17, X16, X12
USA	0.9111	2.5650	X1, X3, X14, X15, X16, X17, X2, X6, X12

Table 19.c: Early rain season: Error rate - LST

Area	Error rate (Group variable selection) (%)	Error rate (Forward selection) (%)
MAC region	4.00	3.88
Antilles	3.36	3.15
South America	3.41	3.41
Center America	4.10	4.07
USA	3.93	4.00

Results are very similar to the observed in the previous section, based on that the decision criteria is the same that were provided in previous steps: to use the best model to produce the estimation, this model is selected independently for each portion of the dataset. Also, the important variables have similarities with the previous examples, BT and their differences are in fact important as well as the time spatial components and NDVI.

Results shows that the alternative to divide the MAC region in homogeneous areas provides better results compared to the option to develop one single model to estimate the entire MAC region. This effect is clearly appreciated when looking throughout the metrics and in a reduction in the processing time. Also, it is observed that difference between the observations and estimations of LST (MAE) is lower than 3.5 Kelvin degrees, and that the error rate is less than 5.5%.

The results obtained from MODIS Aqua appear to be slightly better than the obtained from the models trained with MODIS Terra observations. It corresponded to the observed in chapter 6 and can be related to the observation time from the satellites.

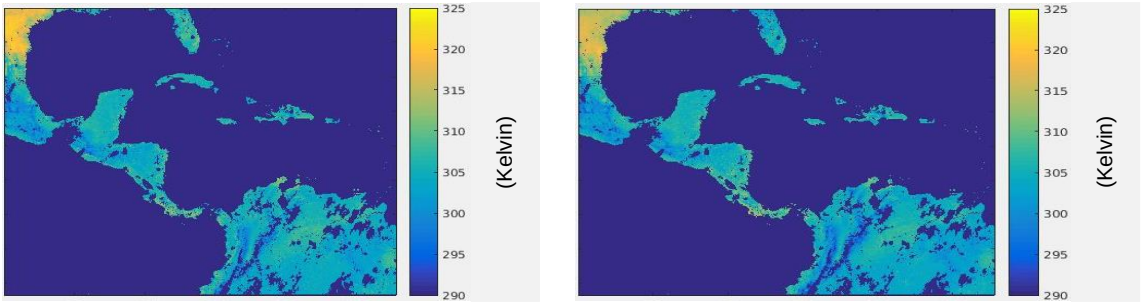
To corroborate the assumptions established in the errors are satisfied, the following residual test were implemented: normality test, independence test and test of constant variances. The Durbin-Watson test (Montgomery et al.,2012) was applied and it was found that the residuals are not independent. The problems with the independence can be attributed to the characteristics of the data that are measured in adjacent times, which induced this problem, especially in the studies with meteorological variables (Rawlings et al., 1998). The Bartlett test (Montgomery et al., 2012) was applied to measure whether or not the variance in the residual is constant, and it was found that the variance is not constant. Also, there was implemented a subroutine to analyze the distribution of the data to observe if it follows a normal distribution. The instability in the variance or heteroscedasticity is usually mitigated using transformations to the dataset (Verran and Ferketich,1984 and Lewis and Lewis, 2015) even when transformations as box-cox were implemented, the heteroscedasticity is still present. Assuming that the problems in the variance and in the autocorrelation in the residuals were negligible these models were used to derive the estimates of PW and LST. However, it is recommended to improve the models in the future.

7.3.5 MODEL EVALUATION

Regression equations are evaluated using the entire set of input variables: it includes information that were not considered in the training dataset because it cannot be linked to an observation of the studied physical parameter (LST or PW). This analysis has been developed to evaluate how adequate are the estimation compared to observe in a real scenario. It consists in generate hourly estimations of both physical parameters for the entire year, using the regression models developed before. It is difficult to present the images that represent every hourly estimation over the MAC region because it means to present a total of 8760 images for

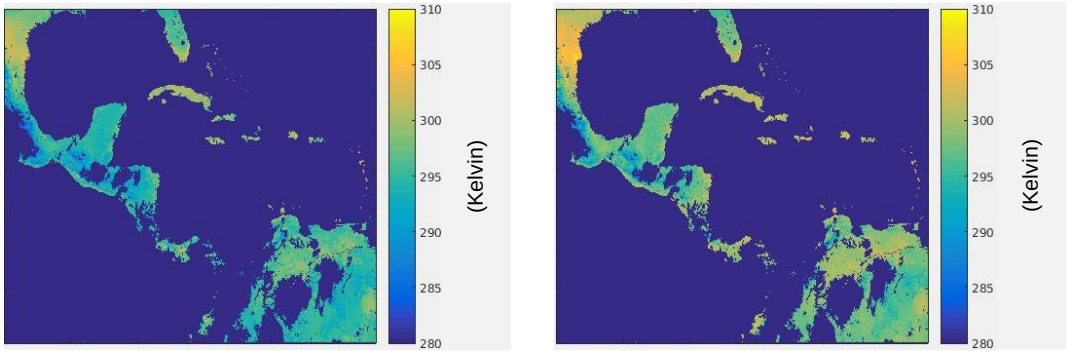
each physical parameter. It was selected 4 images per physical parameter, 2 for the model trained with MODIS Aqua observations and 2 for MODIS Terra. These images correspond to about the maximum sunlight (18 UTC) and the others corresponding to the approximate coldest hour of the day (8 UTC). Figure 17 shows the LST estimation at 18:00 UTC and figure 18 exhibits the estimation at 08:00 UTC. Figure 19 and 20 corresponds to the estimation at the same hours but for PW.

It is observed a good level of correspondence between the estimation of the physical parameters provided by the models and the natural behavior of those parameters that is expected in the environment. For example, looking the results from the Temperature, it increases during daytime hours and get lower during the night time. It also is observed how the topography affects the observed temperature. It can be noticed how the mountains are distinguishable by the low temperatures observed in those areas, similar effect is observed in the desert areas of Texas where the temperature goes to the peak values of the images during the daytime. Similar effects are appreciated in the forest areas of South America.



a. b.

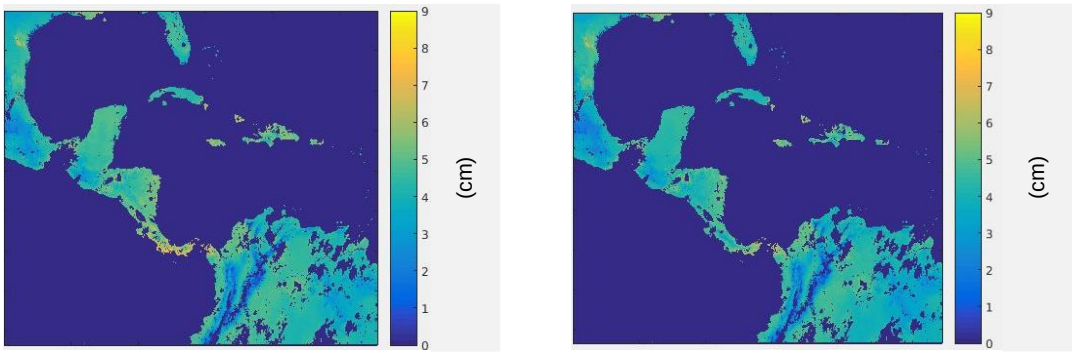
Figure 17: panel a.: Modeled LST Trained using MODIS Aqua. Panel b.: Modeled LST Trained using MODIS Terra Date: August 15 2011 at 18:00 UTC.



a. b.

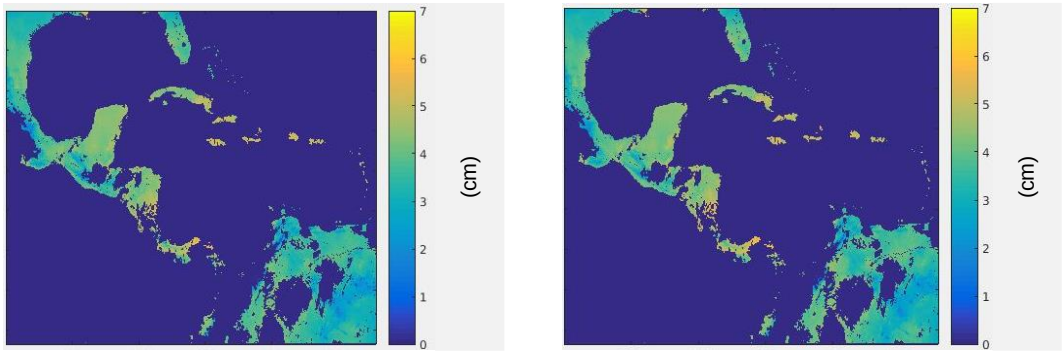
Figure 18: panel a.: Modeled LST Trained using MODIS Aqua. Panel b.: Modeled LST Trained using MODIS Terra Date: August 15 2011 at 08:00 UTC.

It is observed a similar effect in the results from the evaluation of PW figures 18 and 19. The estimation of this physical parameter correspond to the expected values to be observed in real conditions. For example, near to the clouds the values are larger probably because of the presence of rain. However, on the PW it is not expected to be observed a clear shift between day or night.



a. b.

Figure 19: panel a.: Modeled PW Trained using MODIS Aqua. Panel b.: Modeled PW Trained using MODIS Terra Date: August 15 2011 at 18:00 UTC.



a. b.

Figure 20: panel a.: Modeled PW Trained using MODIS Aqua. Panel b.: Modeled PW Trained using MODIS Terra Date: August 15 2011 at 08:00 UTC.

This estimation is not only valuable for evaluation purposes but also, they provide the products of LST and PW that will be introduced as input variables in the RH model. This dataset provides observations every hour covering the entire MAC region and only for land covered areas. However, to use this information, it is necessary to validate those products and the models used to develop.

7.3.6 VALIDATION

As it was mentioned in the methodology, 4 months were used for training and one month for each season were used for validation. The selected months for validation were: December 2011

to validate the dry season models, July 2012 to validate the early rain season and August 2012 to validate the late rain season. The scores to measure the accuracy of the models are: The coefficient of determination R^2, the mean absolute error (MAE), the Root Mean Squared Error (RMSE) and the error rate.

In addition is necessary to clarify that there exists a big number of combination that might be created from the model, and that the difference between them are negligible, based on this it has been decided to pick the best estimation method (group variable selection or forward selection), per each region extracted from the evaluation dataset.

The estimation for the entire MAC region was validated as a whole, this because it is necessary to observe the capability of the group of models to provide a map of estimation that match the observations. Results were organized by season and Table 20 provides results corresponding to the validation of the LST model trained with data from MODIS Aqua and table 21 to LST model trained with data from MODIS Terra.

Table 20: LST – MODIS Aqua validation: performance metrics

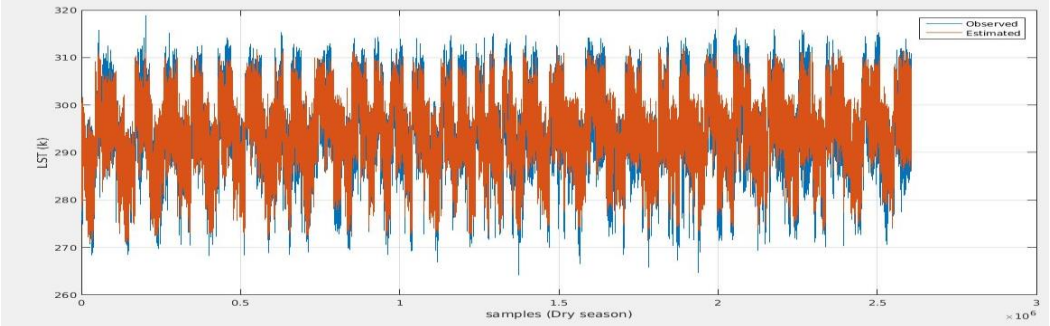
Season	MAE (Kelvin)	RMSE (Kelvin)	R^2	Error rate (%)
Dry season	3.0934	3.8888	0.7311	5.26
Early Rain Season	2.8976	3.7129	0.7497	4.00
Late Rain Season	2.6126	3.4056	0.8097	3.50

Table 21: LST – MODIS Terra validation: performance metrics

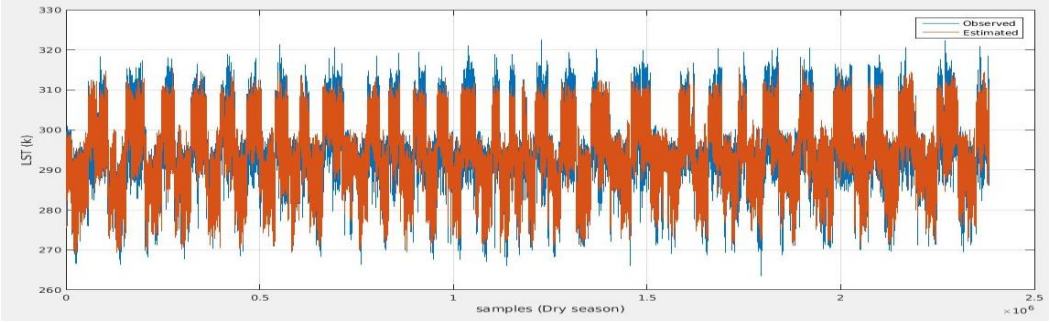
Season	MAE (Kelvin)	RMSE (Kelvin)	R^2	Error rate (%)
Dry season	2.6323	3.3968	0.7155	4.81
Early Rain Season	2.5262	3.2877	0.7290	4.59
Late Rain Season	2.4693	3.2273	0.7584	4.18

Results show that the model provides a good set of estimation under the validation period. The errors are low and the coefficient of determination R^2 are quite as good as they were in the training process. Models properly represent The LST during the validation process, and the highest error rate was about 5.3% of the total variation of values observed in the period but the average error rate was about 4.5% of the total variation of values.

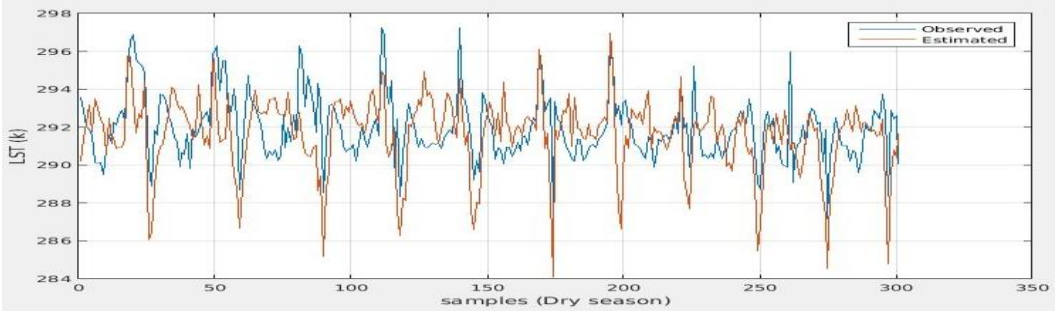
Figures 21, 22 and 23 Shows the time series corresponding to the estimation versus the observations of LST for the studied period. Each image corresponds to a particular validation period and has 4 panels: one corresponds to the estimations based on the observations from MODIS Aqua, and the second one to the estimations from MODIS Terra, the last two are also for MODIS Aqua and Terra, but present a time series with a portion of the data in order to better appreciate the comparison between observations and estimations.



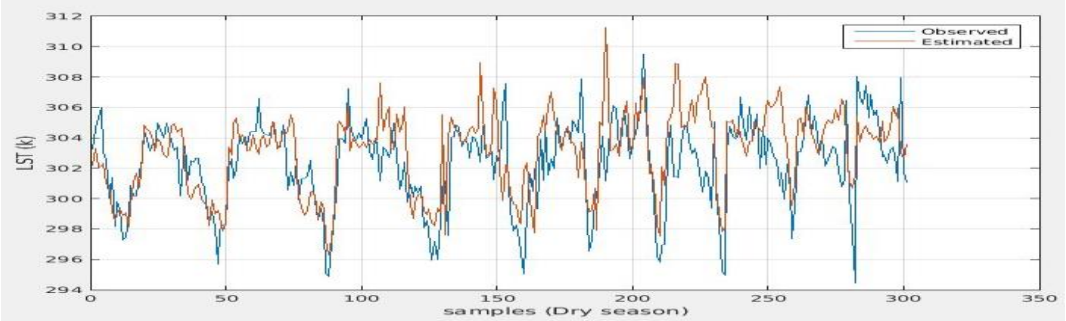
a.



b.

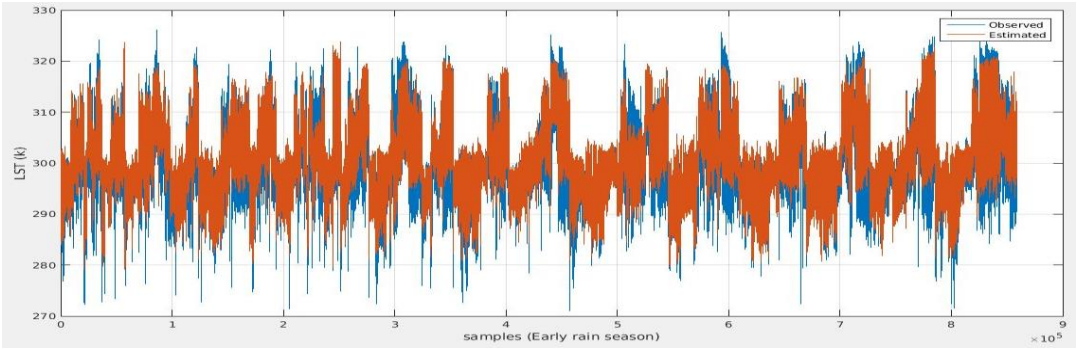


c.

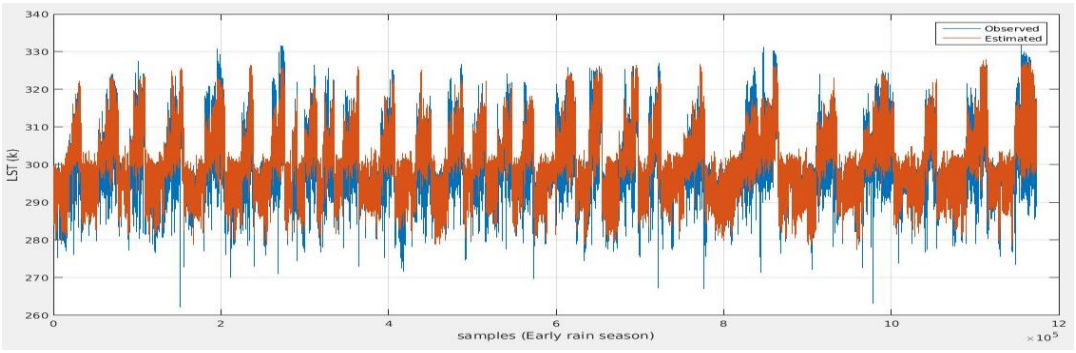


d.

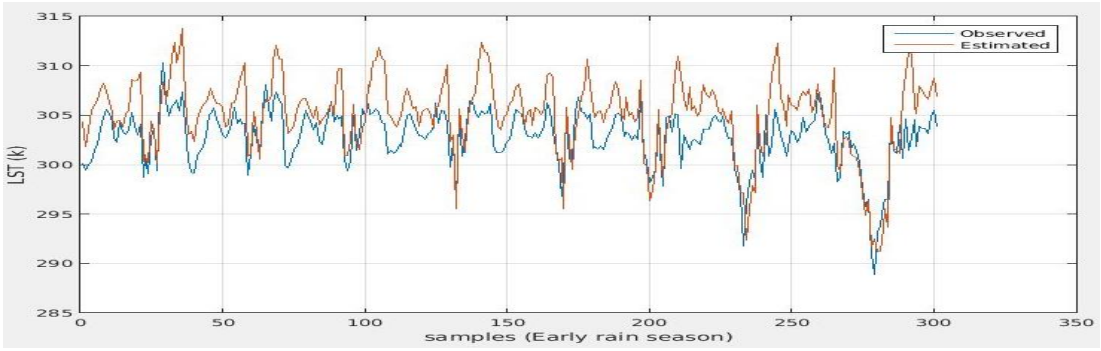
Figure 21: Time series December 2011. Panel a: observations vs MODIS Terra model. Panel b.: Observation vs MODIS Aqua model. Panel c. observations vs MODIS Terra model (sample). Panel d. observations vs MODIS Aqua model (sample).



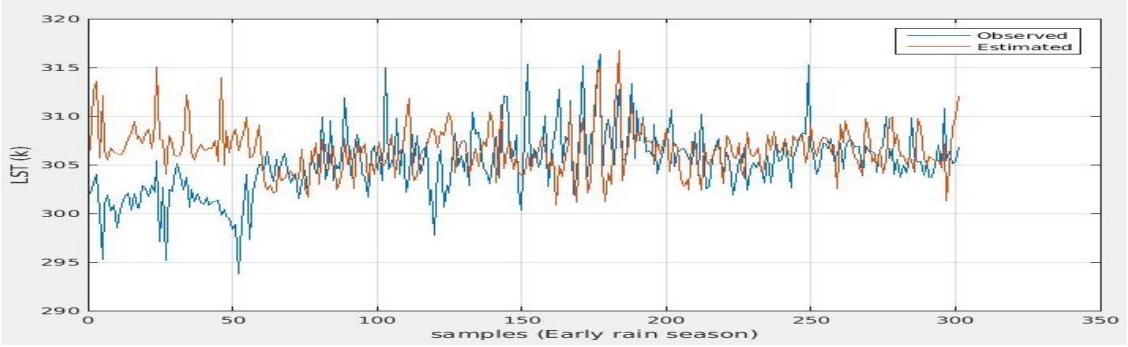
a.



b.

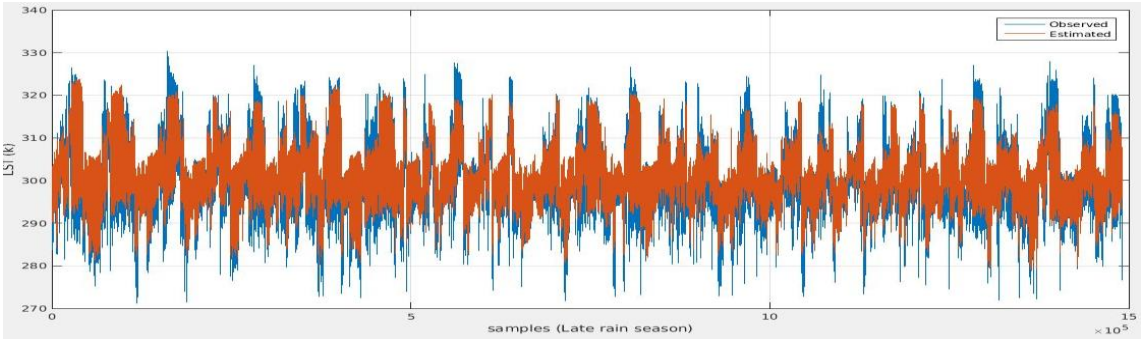


c.

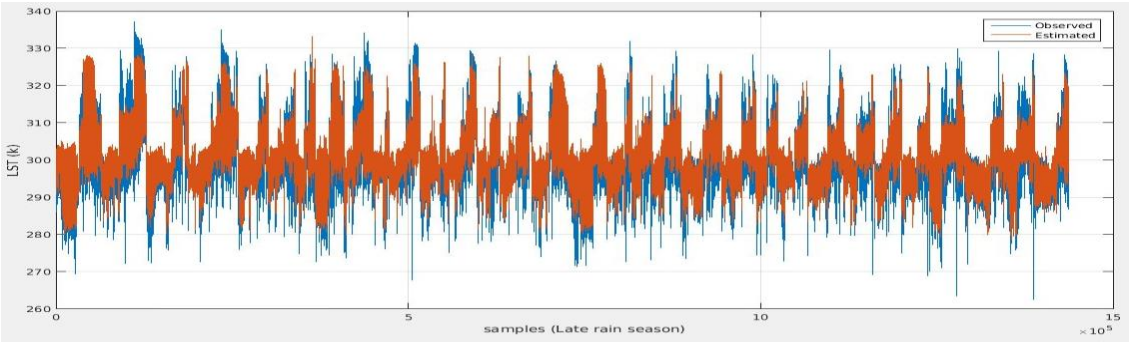


d.

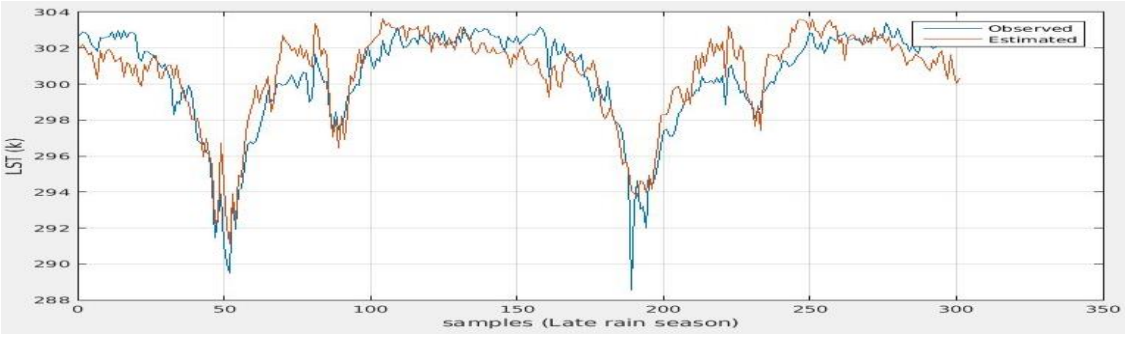
Figure 22: Time series July 2012. Panel a: observations vs MODIS Terra model. Panel b.: Observation vs MODIS Aqua model. Panel c. observations vs MODIS Terra model (sample). Panel d. observations vs MODIS Aqua model (sample).



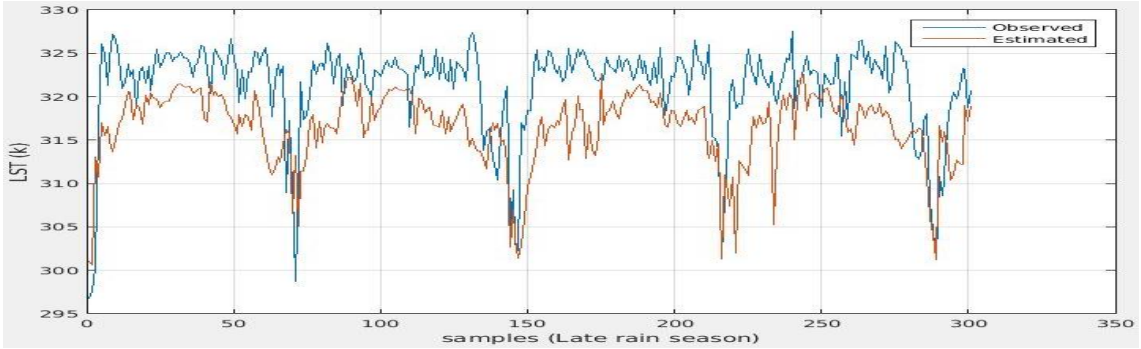
a.



b.



c.



d.

Figure 23: Time series August 2012. Panel a: observations vs MODIS Terra model. Panel b.: Observation vs MODIS Aqua model. Panel c. observations vs MODIS Terra model (sample). Panel d. observations vs MODIS Aqua model (sample)

It has been shown that the models are capable provide a good set of estimations of LST in the MAC region, as it was observed from results. In general, both set of models (MODIS Aqua and MODIS Terra) are good options to estimate LST as it was tested in the validation process.

Same analysis was performed for validation of PW models. Table 22 present the results from the models trained with data from MODIS Aqua and contrasted with observations from this instrument, and table 23 the results from the models trained from MODIS Terra and contrasted with observations from this instrument.

Table 22: PW – MODIS Aqua validation: performance metrics

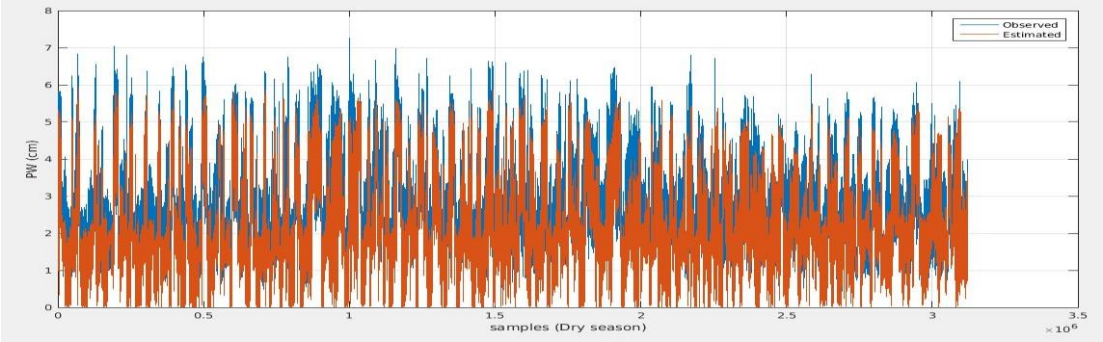
Season	MAE (cm)	RMSE (cm)	R^2	Error rate (%)
Dry season	0.5342	0.6886	0.7499	7.00
Early Rain Season	0.6786	0.8455	0.3078	7.74
Late Rain Season	0.6889	0.8630	0.3038	7.93

Table 23: PW – MODIS Terra validation: performance metrics

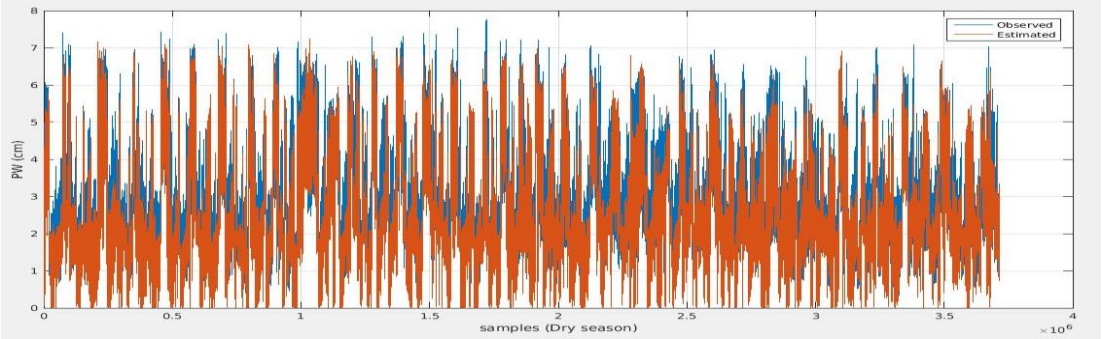
Season	MAE (cm)	RMSE (cm)	R^2	Error rate (%)
Dry season	0.4934	0.6384	0.7225	7.00
Early Rain Season	0.6607	0.8341	0.2222	9.47
Late Rain Season	0.5690	0.7288	0.4309	6.86

The results are not as good as the results observed for the LST. The R^2 values, especially for the periods of Early and Late Rain season, are considerably lower than the obtained during the training period. This may be caused by the selected months. July and August are in the frontier between two rain seasons and located during the summer time this might cause irregularities in the rain and in consequence poor estimation of the PW. However, results in terms of the error are very comparable to the observed during the training process and the error rate was below the 10% and this is consistent with the results observed in the training process.

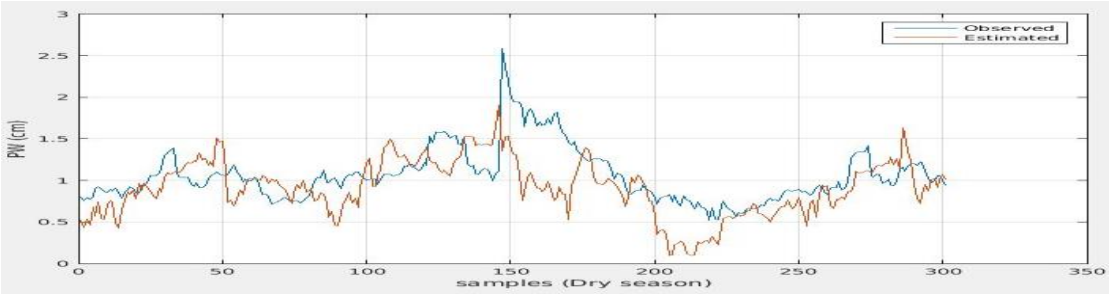
Time series helps to compare the observations of PW with the estimated values. Results are divided in 3 set of images, figure 24 shows the time series for the dry season (December 2011), figure 25 exhibits result for early rain season (July 2012) and figure 26 for the late rain season (August 2012). Each of the figures was divided into 4 panels: Panel a compares the observation gathered from MODIS Aqua and Panel b presents the same information but using MODIS Terra instrument and the last two are also for MODIS Aqua and Terra, but present a time series with a portion of the data in order to better appreciate the comparison between observations and estimations.



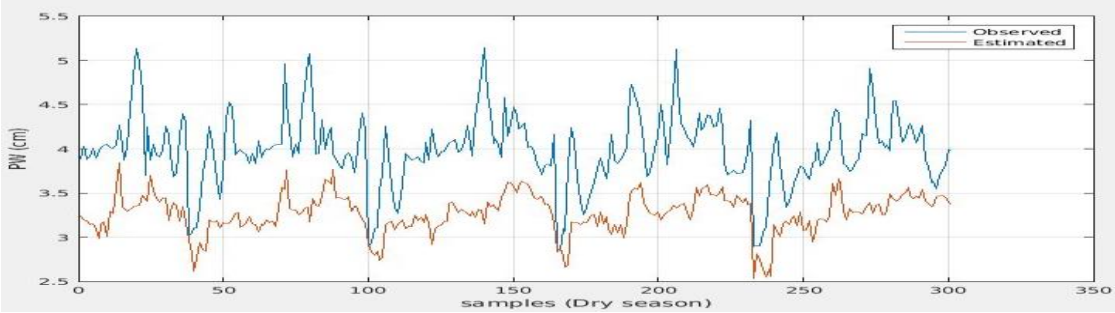
a.



b.

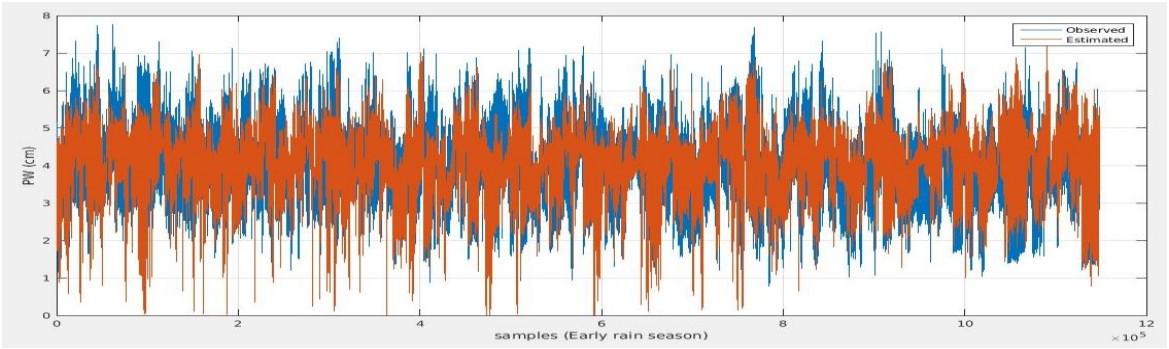


c.

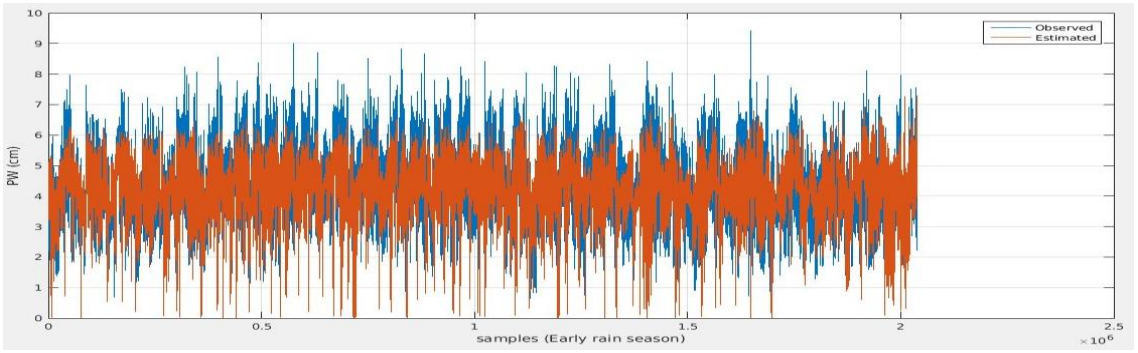


d.

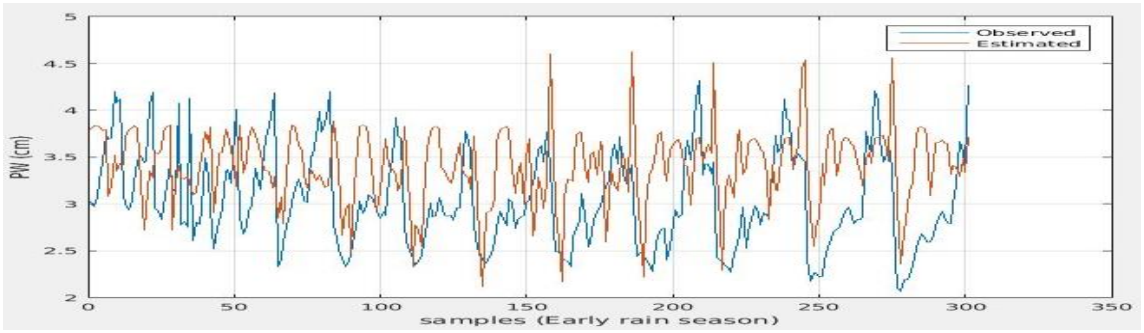
Figure 24: Time series December 2011. Panel a: observations vs MODIS Terra model. Panel b.: Observation vs MODIS Aqua model. Panel c. observations vs MODIS Terra model (sample). Panel d. observations vs MODIS Aqua model (sample).



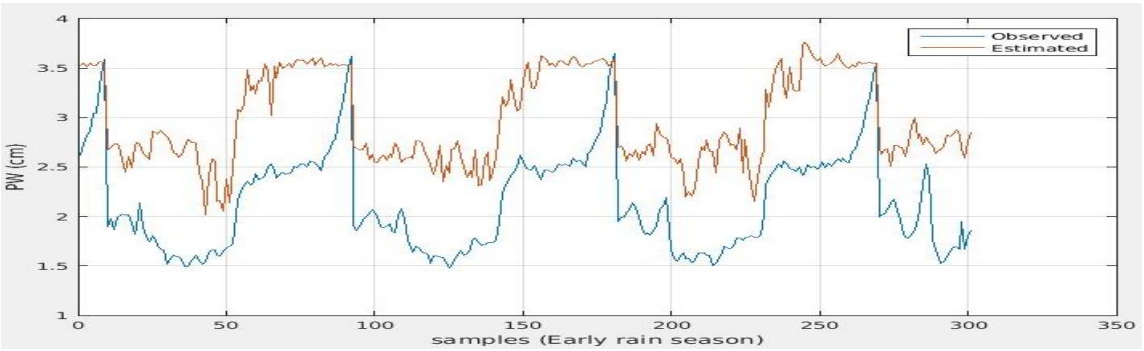
a.



b.

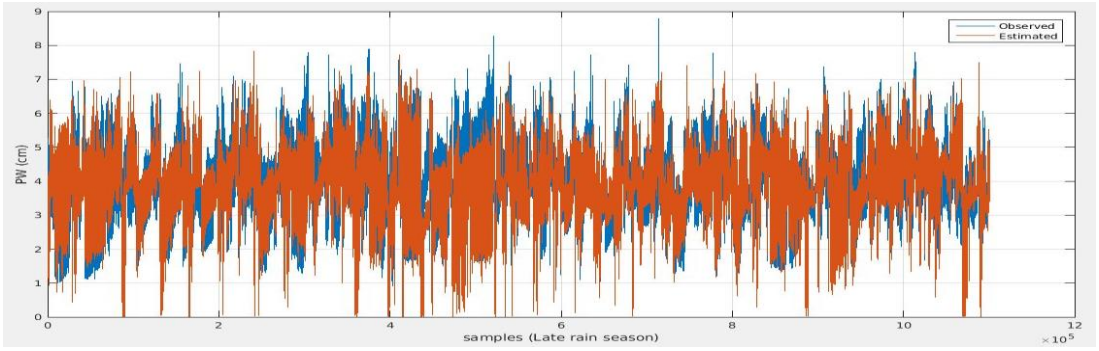


c.

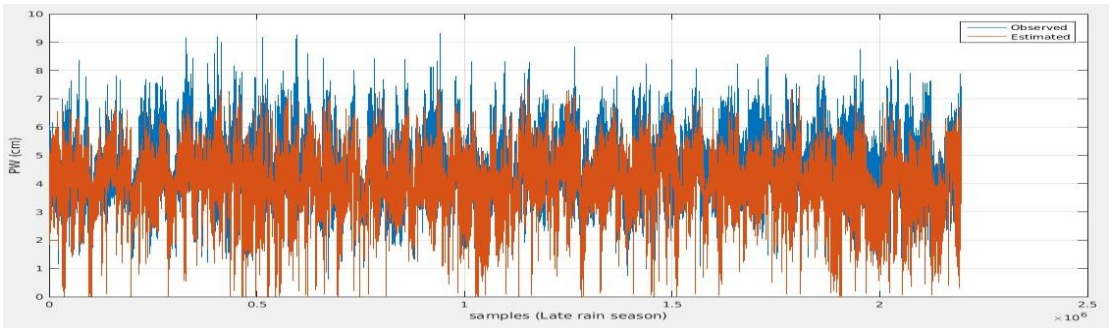


d.

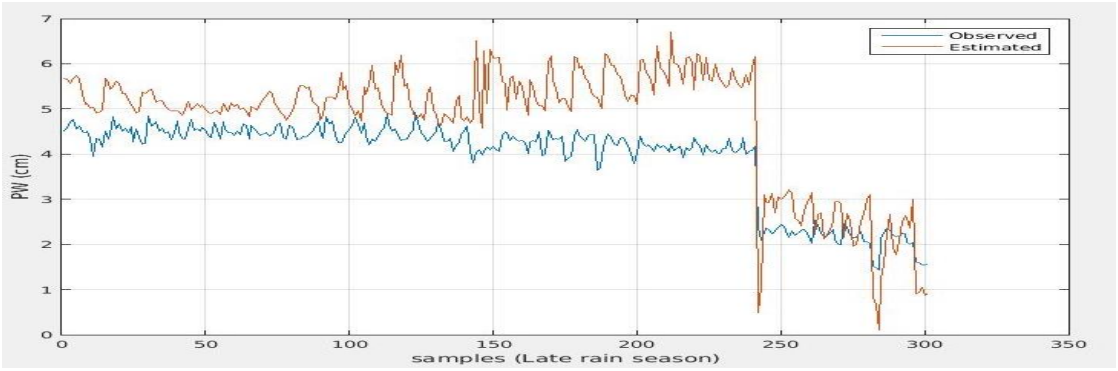
Figure 25: Time series July 2012. Panel a: observations vs MODIS Terra model. Panel b.: Observation vs MODIS Aqua model. Panel c. observations vs MODIS Terra model (sample). Panel d. observations vs MODIS Aqua model (sample).



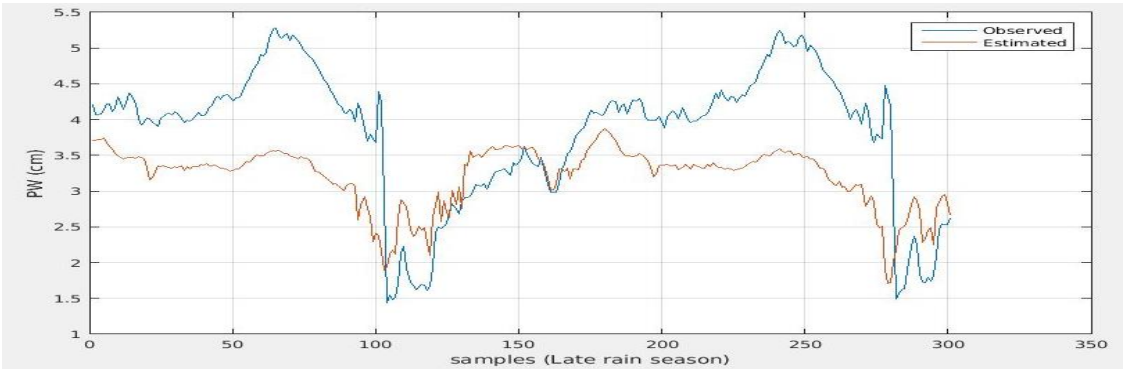
a.



b.



c.



d.

Figure 26: Time series August 2012. Panel a: observations vs MODIS Terra model. Panel b.: Observation vs MODIS Aqua model. Panel c. observations vs MODIS Terra model (sample). Panel d. observations vs MODIS Aqua model (sample).

It was observed a considerable quantity of values that appear to be extreme values in the estimation of PW. It was noted that this values may be caused by an anomaly in the input variables obtained from GOES. These values may be related with a rain event that causes a difference of about 1 cm. of PW and that may be responsible for the problem with low R^2 values.

In general PW estimations are acceptable since the error values and error rate and comparing them with the values obtained during the training process they are similar. Even the time series appears to follow the real behavior of this variable and their changes. The group of models are now feasible to obtain the input variables to build the hourly RH humidity models.

8. ESTIMATION OF RELATIVE HUMIDITY, BASED ON GOES AND MODIS DATA, USING REGRESSION AND ANN TECHNIQUES

This chapter represent the last step in the process to estimate RH from satellite data. Previously, a first exercise has been performed. From that exercise, it was discovered that RH can be expressed as a combination of different physical parameters as: LST, PW and NDVI. However, two of the physical parameters (LST and PW) were obtained from a satellite instrument that is not capable to offer hourly observations. Also, those observations do not cover the entire MAC region in hourly basis. Thus, the available information limits the RH in terms of it time and spatial resolution.

The second step was implemented to obtain estimations of the physical parameters. A new set of models based mainly on GOES imagery where developed to estimate the LST and PW and these estimations replace the observations obtained from MODIS. The estimates of PW and LST have a spatial resolution of 4 km and are obtained at every hour.

Now all the required input variables to estimate RH fulfill the proposed objectives. However, to estimate RH it became necessary to generate a new set of models, based on the new characteristics of the inputs. In this chapter a new approach to estimate RH is be developed.

8.1 DATA DESCRIPTION

The Data introduced in this group of models is divided into two different categories: The input variables and the response variable. The response variable is related to the observations of RH obtained from 584 stations across the MAC region.

The input variables correspond to a set of products, mostly from remote sensing, that will be introduced into the estimation model. This group of variables have been already studied in the first exercise presented in chapter 6; however, exists differences in terms of the data. The easiest and remarkable difference is the instrument where the physical parameters were obtained. In the first model, most of the products were obtained from MODIS instrument, but in this new exercise Most of them were derived from GOES imagery instrument. However, MODIS will still provide the NDVI input variable. This product has a spatial resolution of 4 km and are scaled to be given in hourly intervals.

GOES imagery replaces MODIS data providing LST and PW physical parameters at a spatial resolution of 4 km and a time resolution of hourly estimates. These estimates cover the entire land MAC region that are under clear sky conditions.

DEM is also being included in this model, motivated by the results from the previous set of RH estimation models where elevation resulted as an important variable. This product has a 4 km spatial resolution. The characteristics of the input variables are described in Table 24.

Table 24: Characteristics of the data

Product	Instrument	Spatial resolution	Time Resolution
Land surface temperature (Estimated)	GOES	4 km	Hourly
Precipitable Water (Estimated)	GOES	4 km	Hourly
NDVI (Observed)	MODIS	4 km	Hourly (scale)
Elevation (DEM)	DEM	4 km	N.A.*
RH (Observed)	Stations	N.A.*	Hourly

* Does not apply for this product.

The brightness temperature calculated from GOES imagery channels 2 to 6 (see Table 1) were also included as input variables into this set of models. It is important to include those because, as it was explained before, those channels have been designed to capture specific characteristics of the atmosphere, characteristics that may be significant to understand some changes in terms of the RH. This set of products have a spatial resolution of 4 km and are offer every hour for the entire MAC region. The products from GOES imagery does not differentiate between land covered areas and sea, it is necessary to include a mask to separate them. This mask has already been described in the chapter 7, section 7.1.

To model the RH, it is necessary to represent the trend and the seasonal components. The trend was represented as a linear combination of input variables and the seasonal behavior by a set of sinusoidal functions. (Newton, 1988, Brockwell and Davis, 2002 and Ramírez-Beltran et al., 2010).

To extract the periods a Fast Fourier Transform (FFT) has been applied over the dataset to identify the most important periods in a dataset. (Ramírez-Beltran et al., 2016). To find these periods it is important to work with a dataset (Station observation) from the response variable that has an enough amount of data, without missing values. From the group of station, it has been selected and analyzed a couple of station that fulfill this during the five years. Two factors are necessary to obtain the important periods: one is the value of the Fourier Transform which

provides the important values and the second one is the frequency of the Fourier Transform and its inverse provides the period. The results of the FFT are shown in figure 27.

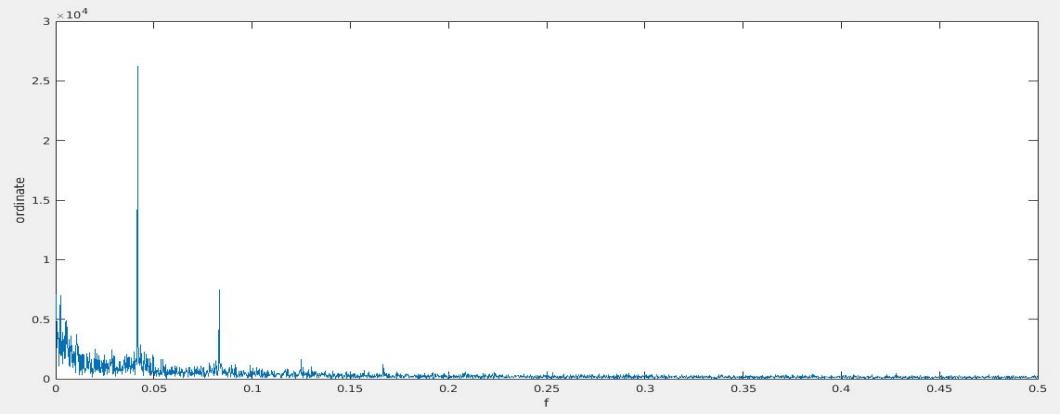


Figure 27: Spectral analysis of time series.

To find the most significant periods it resulted necessary to look the highest ordinate values and their corresponding frequencies (f), the inverse of the frequency correspond to the important periods. There are two important periods in this dataset, and those are:

- Ordinate value of 2.6248×10^4 corresponding to a frequency of 0.0417. The correspondent period value is 23.981 hours, or equivalent to 24 hours.
- Ordinate value of 7.4417×10^3 corresponding to a frequency of 0.0833. The correspondent period value is 12.005 hours, or equivalent to 12 hours.

There are two important periods, 12 and 24 hour cycles. To include those into the regression model it is necessary to develop a group of phase components, those are based on trigonometric functions. The selected trigonometric functions for this exercise are sine and cosine functions. (Ramírez-Beltran et al., 2016). The periods that have been found previously: 12 and 24 hours. However, should not be related to this exact number of hours, but with their atmospheric significance which is the day-night cycles. These cycles are not exactly every 24 hours, they are changing depending of the position of the earth respect to the sun. These changes during the year are related to the latitude longitude and to the solar zenith angle. These cycles can be tracked based on the sunrise and sunset time for each day during the studied period. The algorithm in this works is based an algorithm extracted from ed Williams' aviation page website (Williams, 2016) which is based on (Doggett et al. 1990). This algorithm was modified on this work in order to obtain the sunrise sunset hours for the entire MAC region during the entire studied period (training and validation), and with this information it was constructed the phase variables for the 12 and 24 hour periods.

8.2 METHODOLOGY

The methodology consists on using the developed predictors and apply regression technique to identify the variables that best express the variability of RH. ANN technique can also be explored to take advantages of modeling the nonlinear relationship that may be present between predictors and RH.

8.2.1 MATCH ALGORITHM

This algorithm match data from the stations dataset and the variables that will be introduced into the regression model. The variables were coded to manage a large amount of data and the involved variables are: RH data obtained from the stations, PW and LST estimated from GOES, the different infrared BT from the GOES channels, the MODIS NDVI data and the elevation.

This code starts opening the station dataset and extracting the position of each station. Data that are not required for training were eliminated from the file, which includes 5 years of data. But, before performing data elimination it is necessary to identify inconsistent data. Data from stations were processed by a quality control system to eliminate inconsistent data and this procedure was accomplished by using the Chebyshev inequality. One of the major advantages of this method is there is no need to know the probability distribution of the underlying data. This procedure is summarized as follows:

- Data from each station was opened, and the mean and standard deviation values were computed.
- Based on these parameters, and to the defined tolerance that is accepted, in terms of the standard deviation. Every value in the station series was evaluated and if the value falls outside the acceptable range then it is replaced by a missing value. This process is repeated for every station.

The training station data were selected and include from December 2010 through November 2011. Data from the different variables must be matched in time and space with station data. The procedure is explained bellow.

- Examine each hour from the RH dataset and search for the nearest observation in the different variables: LST, PW, GOES BT and NDVI. This closest images should be distant in no more than 30 minutes up or down from station data. If no image is found during this specific time, then the values of those variables corresponding to each

station at that time will be filled with a missing value. If images are found, then they will be matched to derive the regression model.

- The station positions are analyzed, searching for the nearest pixel from each station in every of the variable (images). Once a pixel is found then the value of that variable will be saved for each one of the stations. It is possible that a pixel in one or more variables may not be found at a specific time because it was covered by a cloud or rain, then this pixel will be replaced with a missing value for the specific variables. This process was repeated for every hour during the entire studied period.
- Also for the nearest pixel to each station, the elevation value was extracted and linked to that determined position. This process was performed only one time because the elevation value does not change during time.

The output of this algorithm was a set of tables. Those were related with the different studied variables and have different sizes. The table time have 3 columns (month, day and hour) and as many rows as hour exists during the studied period. The position table have 3 columns (Latitude, Longitude and Elevation) and as many rows as station exists. And all the other variables (LST, PW, GOES channels 2 to 6 brightness temperatures, NDVI, RH) were saved as individual tables and they will have as many columns as station exists, and as many rows as hour exists during the studied period. The data corresponding to the different periods will be aggregated in the following algorithm.

8.2.2 STRUCTURE AND CLEANING ALGORITHM

Two set of tables were imported in this algorithm; both were necessary to generate the final table with the variables needed for the regression. One set of tables has the variables obtained from GOES, MODIS and the elevation values expressed by station every instant of time. The second one provides the seasonal components for each of the pixels (stations) that are also expressed per station and every instant in time. Each of those variables are arranged in form of different tables with a particular distribution, those variables need to be rearranged. However, prior to that it is necessary to clean some of them using the same algorithm explained before based on Chebyshev' inequality. These variables were: PW, LST, and the Brightness temperature from GOES.

After doing this cleaning the rearrange process will be executed. The process is described below:

- The first variables to be rearranged were the LST, PW, NDVI, GOES brightness temperature and RH. They were presented as a table, where the columns represent the data from the nearest pixel to each station (584) and each of the rows represents the corresponding observation of the product for each pixel for every hour during the studied period (8760 hours). This table was converted and regrouped as a vector, moving one column bellow the other. The size of the new vector is 4599000×1
- Time matrix was also defined as a 3 column table: one for each one of the variables (Month, Day, and Hour). It has as many rows as the number of hours during the studied period (8760 hours). This process (table for station) was repeated 584 times one bellow the other, because every of the station-pixel have the same number of hours. With this rearrange now it is obtained the table corresponding to the 3 time variables. The size of this matrix is 4599000×3
- Four seasonal variables were imported as tables. There were 4 tables: 2 for 12 hour periods and 2 for 24-hour period, that correspond to 2 sinusoidal (sin and cosine) functions. The number of rows correspond to every hour during the studied period (8760 hours). The number of columns correspond to each station-pixel position (584). To rearrange this table every column will be allocated below the previous one, transforming each table into a vector of size 4599000×1.
- Another group of variables imported were the position. This table has 3 columns: latitude, longitude and elevation, and 584 rows, one for each station-pixel. Each of this rows are a position and this table is valid for every hour that has information for this station-position (8760 hours). Based on that, this table was rearranged in a new matrix repeating each individual row 8760 times, one bellows the other, repeating this same process for each of the rows until all of those are being processed. It was generated a new array that contains three variables: Latitude, Longitude and Elevation. The new size of the new matrix is 4599000×3

All these variables were grouped as a single table. This table has as columns the variables paste one next to each other.

The next step is to implement the quality control process. It is important to eliminate both missing and inconsistent values. The cleaning algorithms have already been discussed in detail in previous chapters. However, it is necessary to describe the thresholds defined for each of the variables that will be cleaned. Those are: 260 to 330 Kelvin degrees for the BT of channel 2,

235 to 270 Kelvin degrees for the BT of channel 3, 280 to 330 Kelvin degrees for the BT of channel 4, 240 to 290 Kelvin degrees for the BT of the channel 6, 0 to 320 for the LST, 0 to 1 for NDVI, and 0 to 1 for all the seasonal components.

This new table provides the entire set of input variables and the response variable for the entire studied cycle. But, before be saved, it is necessary to separate the data corresponding to each of the defined periods. The variables to discriminate between periods are the year and the month. Three different tables will be saved: one for the dry season (December 2010 through March 2011), one for Early Rain season (April 2011 through July 2011) and one for Late Rain Season (August 2011 through November 2011).

8.2.3 DATA PROCESSING

This algorithm is the responsible for developing regression models. It starts loading the tables generated before. There are 6 different modifications for this algorithm, depending on the data to be introduced. There are 2 different datasets that can be introduced depending if LST and PW were trained using MODIS Terra or Aqua and for each of those exists 3 different tables one for each season, it gives a total of 6 possible datasets. For each of those datasets an independent model will be generated.

Once the corresponding dataset has been loaded, it is necessary to divide them into the previously defined homogeneous zones. To create these zones different cutting rules will be applied over the latitude and longitude variables. Once the table for each zone are created, it is necessary to divide each of these into two different ones, one for the response variable (Y) and another for the input variables (X's). The variables that are introduced into the regression are presented in table 25:

Table 25: Description of the variables

Variable	Description	Variable	Description
Y	RH	X_9	NDVI
X_1	Latitude	X_{10}	BT2
X_2	Longitude	X_{11}	BT3
X_3	Elevation	X_{12}	BT4
X_4	Month	X_{13}	BT6
X_5	Day	X_{14}	Sinday1 (24 H)
X_6	Hour	X_{15}	Sinday2 (12 H)
X_7	PW	X_{16}	Cosday1 (24 H)
X_8	LST	X_{17}	Cosday2 (12 H)

The two regression techniques as well as their routines to eliminate multicollinearity will be applied. In addition, the ANN technique is included to explore the nonlinear connection of the input variable with RH. The input variable selected by the regression approach are the ones that are input into the ANN algorithm with the purpose of eliminating the multicollinearity problem.

The structure of the ANN includes one hidden and one output layer. The number of neurons in the hidden layer varies from one to three and each one includes a log-sigmoidal transfer function and the output layer has one neuron with a linear transfer function. The best number of neurons in the hidden layer was selected by the best performs of the algorithm. The log-sigmoidal transfer function scheme resulted better than the linear-linear scheme. The training process starts with a random initial point. However, before training the best initial point out of 7 was selected.

Three different performance metrics were selected to determine the best estimation techniques. These performance metrics are: the mean absolute error, the error rate and the coefficient of determination R^2 value.

- For the ANN, it was saved: the structure of the net, represented by the starting point, number of neurons, weight and bias, and the corresponding variables that enter into the net.
- For the regression models, it was saved: the important variables that enter into the model and also their corresponding coefficients including the constant.

In appendix 1, It can be found an example of the implemented algorithm, using MATLAB.

8.2.4 MODEL EVALUATION

To develop the models, only the pixels related to the stations have been studied. However, it has been downloaded data from the input variables that can be used to evaluate the identified models and create a map the RH for the entire MAC region and almost at every hour. To understand the behavior of the RH and to analyze their corresponding cycles and changes compared to the real observation, it becomes necessary to develop maps of RH.

An algorithm has been developed to obtain those maps and it will be explained bellow:

- The algorithm starts loading the images of the input variables corresponding to the entire studied year, as well as the parameters from identified models. The algorithm divides the data in 3 different periods based on the corresponding time from each image.

- The algorithm identifies all the hours that can be matched (that have all the necessary components), and it is necessary to study each one of the pixels inside of it.
- Each pixel is study first to find if it has a nearest pixel in other images; if it is not the case, then a missing value will be assigned to this pixel. Each of those pixels are matched and the information related to their position is studied and their corresponding region is identified. Based on their corresponding region and season the associated parameters of the models are found and applied. RH is estimated based on those input variables and the corresponding parameters of the model. This process is repeated for each pixel and for this hour.
- When all the pixels in a determined hour has been studied a new set of tables will be created and saved, and those are: The estimation of RH for every pixel in the image, the latitude and the longitude tables corresponding to those pixels. These images correspond to the estimated RH image for a determined hour. This process was be repeated for every available hour during the studied period.

This process was repeated for every hour during the training period (December 2010 to November 2011). This process was also repeated during the validation Period (December 2011, July 2012 and August 2012).

8.2.5 VALIDATION

It is necessary to validate the previously developed model to test the ability of those model to produce correct estimations of RH under a different time period. In this validation stage, three months of data were studied, one for each of the delimited seasons: December 2011 for dry season, July 2012 for early rain season and August 2012 for late rain season.

The validation algorithm compares observations with estimated RH values, and it is described below:

- This algorithm starts by cleaning all the observations of RH, by using the Chebyshev' inequality, which was explained before and is used to remove inconsistent observations. Then, it is necessary to extract from the observation dataset only the data corresponding to the validation months.
- The estimated RH products that correspond to the validation dataset will also be loaded and each hourly file to be analyzed. The image is opened and the pixels nearest to each station will be searched. The estimated RH values from each of those pixels will be

extracted and linked to the observed RH values in each of the station during that hour. This process will be repeated for every hour in the validation studied time period.

- The difference between the observed and the estimated RH values were calculated and the MAE, RMSE, the error rate and the R^2 coefficient were calculated and analyzed.

The error values will be calculated independently for every one of the seasons in order to compare each of the model performance.

In figure 28, it is presented a diagram that summarizes the methodology employed in this stage:

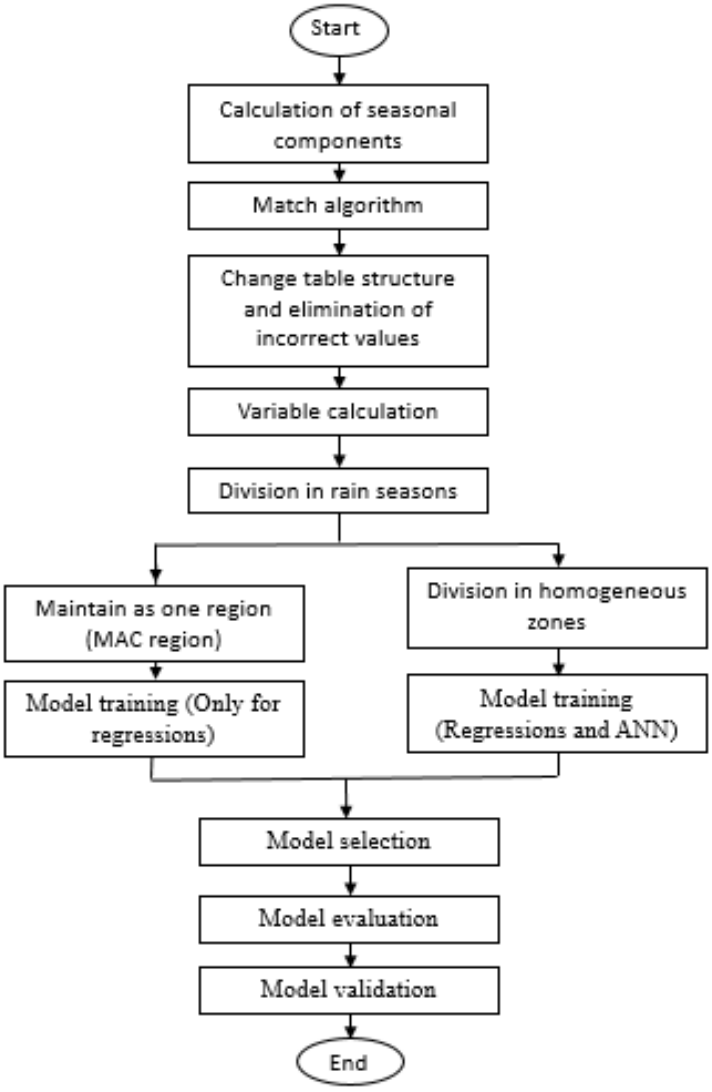


Figure 28: Methodology diagram. Final estimation of RH.

8.3 RESULTS

RH results are based on the Physical parameters estimated from GOES. These results present not only the model development but also the evaluation and validation stage.

To develop the models for estimating RH two important input variables were estimated from MODIS. On the previous chapters, it was concluded that exists 2 different set of models to estimate those variables, one trained using data from MODIS Aqua and other from MODIS Terra. It is necessary to work with two models and select the one that provides the best results.

8.3.1 RELATIVE HUMIDITY – MODIS AQUA

This section presents the group of models to estimate RH using PW and LST estimated from MODIS Aqua, and NDVI which is also obtained from MODIS Aqua.

This sections provides the results obtained from the three methodologies implemented: regression based on forward selection algorithm, regression based on group variable selection and ANN. The models are classified by rainfall season, and organized by homogeneous climatic zones or for the entire MAC region.

Table 26 presented the RH models corresponding to the dry season. This table is divided in 4 sub tables: a. shows the results corresponding to the group variable selection technique, b. the results corresponding to the forward selection technique c. shows the error rate for both techniques and d. exhibits results corresponding to the ANN. Tables 27 and 28 present the corresponding results for Early and Late rain season models, respectively, and using the same sub table structure.

Table 26.a: Results Dry season from Group variable selection technique - RH

Area	R^2	MAE (%)	Important variables
MAC region	0.5022	8.3532	X14, X16, X9, X17, X15, X10, X13, X11, X6
Antilles	0.5535	7.1470	X14, X16, X17, X7, X15, X6, X11, X10, X13
South America	0.5163	6.9581	X10, X16, X14, X2, X15, X1, X7, X8, X5, X9
Center America	0.5091	8.8009	X14, X16, X9, X17, X15, X10, X13, X11, X6
USA	0.5449	9.4297	X14, X16, X3, X5, X13, X11, X15, X17

Table 26.b: Results Dry season from Forward selection technique - RH

Area	R^2	MAE (%)	Important variables
MAC region	0.5098	8.3339	X14, X16, X9, X17, X15, X13, X1, X10, X6, X5, X3, X11, X2, X4
Antilles	0.5619	7.0323	X14, X16, X17, X9, X15, X6, X13, X10, X4, X11, X2, X3, X5
South America	0.5271	6.8393	X10, X16, X14, X2, X4, X15, X1, X17, X5, X6, X9, X11, X13, X3
Center America	0.5296	8.6026	X14, X16, X9, X17, X1, X2, X15, X10, X3, X4, X5, X13, X6
USA	0.5643	9.2680	X14, X16, X7, X3, X5, X13, X17, X2, X15, X9, X4

Table 26.c: Dry season: Error rate - RH

Area	Error rate (Group variable selection) (%)	Error rate (Forward selection) (%)
MAC region	9.83	9.81
Antilles	9.83	9.67
South America	9.01	8.86
Center America	10.95	10.70
USA	11.10	10.91

Table 26.d: Results Dry season from ANN - RH

Area	R^2	MAE (%)	Error Rate (%)
Antilles	0.6337	6.3456	8.72
South America	0.5953	6.3132	8.18
Center America	0.6218	7.6265	9.49
USA	0.6298	8.3573	9.84

Table 27.a: Results Early rain season from Group variable selection technique - RH

Area	R^2	MAE (%)	Important variables
MAC region	0.5688	7.5013	X10, X16, X14, X7, X17, X8, X11, X15, X13
Antilles	0.6238	6.2842	X14, X16, X17, X7, X8, X10, X13, X15, X11
South America	0.4142	7.3861	X8, X16, X14, X17, X11, X15, X10, X13
Center America	0.5090	8.4022	X10, X16, X9, X14, X6, X5, X4, X3, X2
USA	0.6241	7.9711	X16, X3, X14, X7, X13, X17, X15, X11

Table 27.b: Results Early rain season from Forward selection technique - RH

Area	R^2	MAE (%)	Important variables
MAC region	0.5819	7.3916	X10, X16, X14, X9, X17, X15, X1, X13, X2, X5, X4, X3, X11, X6
Antilles	0.6417	6.1338	X14, X16, X17, X7, X2, X10, X9, X15, X5, X13, X11, X3, X6
South America	0.4303	7.2224	X16, X14, X17, X11, X2, X9, X15, X6, X4, X12, X3, X5, X13
Center America	0.5283	8.1957	X10, X16, X9, X14, X17, X15, X13, X4, X5, X3, X11, X6, X2
USA	0.6276	7.8509	X16, X3, X14, X13, X17, X15, X9, X2, X1, X6, X4

Table 27.c: Early rain season: Error rate - RH

Area	Error rate (Group variable selection) (%)	Error rate (Forward selection) (%)
MAC region	8.47	8.34
Antilles	8.95	8.74
South America	9.26	9.05
Center America	10.25	10.00
USA	9.00	8.86

Table 27.d: Results Early rain season from ANN - RH

Area	R^2	MAE (%)	Error Rate (%)
Antilles	0.7161	5.4035	7.70
South America	0.4891	6.7619	8.48
Center America	0.6002	7.4758	9.12
USA	0.6903	7.0769	7.99

Table 28.a: Results Late rain season from Group variable selection technique - RH

Area	R^2	MAE (%)	Important variables
MAC region	0.5455	7.8147	X14, X16, X7, X13, X17, X15, X10, X11, X6
Antilles	0.5408	6.0830	X14, X16, X17, X15, X10, X13, X6, X11
South America	0.4588	7.2258	X17, X16, X14, X2, X6, X4, X3, X9
Center America	0.5102	7.1506	X14, X16, X7, X2, X8, X9, X3, X5, X1, X4
USA	0.6389	8.8061	X14, X16, X2, X1, X3, X4, X6, X9, X5

Table 28.b: Results Late rain season from Forward selection technique - RH

Area	R^2	MAE (%)	Important variables
MAC region	0.5842	7.5197	X14, X16, X9, X2, X17, X13, X15, X3, X4, X5, X1, X6, X10, X11
Antilles	0.5707	5.8636	X14, X16, X17, X9, X15, X10, X2, X3, X13, X6, X11, X5, X4
South America	0.4725	7.0779	X17, X16, X14, X2, X1, X15, X3, X11, X6, X10, X13, X5
Center America	0.5466	6.7952	X14, X16, X17, X2, X9, X15, X13, X10, X1, X5, X11, X3, X6
USA	0.6756	8.2981	X14, X16, X2, X1, X15, X11, X17, X3, X9, X5, X13, X4

Table 28.c: Late rain season: Error rate - RH

Area	Error rate (Group variable selection) (%)	Error rate (Forward selection) (%)
MAC region	8.90	8.56
Antilles	8.81	8.49
South America	9.25	9.06
Center America	8.89	8.45
USA	10.03	9.45

Table 28.d: Results Late rain season from ANN - RH

Area	R^2	MAE (%)	Error Rate (%)
Antilles	0.6192	5.4582	7.90
South America	0.5149	6.8948	8.83
Center America	0.6121	6.2936	7.83
USA	0.7333	7.3688	8.39

The best results were obtained from the ANN. This technique offered the biggest R^2 values and correspondingly the lowest errors and the ratio between the error rate and the total variability is also the lowest. However, to achieve those results different assumptions have been applied over the dataset. The input variables that were included in the ANN are the most important variables from the best regression techniques, for each model. This decision has been adopted to reduce the multicollinearity issues. Also, the routine of ANN used a nonlinear transfer function, specifically the log sigmoidal function. The results become better compared to the ones obtained with regression, on the downside the processing time increases exponentially. It was decided to divide the MAC region into homogeneous zones to improve estimation and to reduce the computational time.

South America was the hardest regions to estimate RH, this especially noticeable in their higher error rate and poor R^2 value. This is somehow expected based on the low quantity of stations that are presented in this region that are not sufficient to explain the entire variability of the data. This problem is also a heritage from the estimations of LST and PW that were used as input variables in this model. The dry season correspond to the winter time, a period with complex behavior of the RH.

In terms of the important variables, at least two of the seasonal components were important variables, it confirms the fact that the RH follows a periodical behavior. Also, the variables related to the GOES brightness temperature as well as some of the physical parameters appear as important variables in most of the models, but depending on the combination of region with season then the GOES channel and physical parameter combination are different.

8.3.2 RELATIVE HUMIDITY – MODIS TERRA

This section presents a set of models to estimate RH using PW and LST estimated from MODIS Terra and NDVI obtained from MODIS Terra.

Table 29 presented results of the models corresponding to the dry season. This table is divided in 4 sub tables: a. shows the results corresponding to the group variable selection technique; Table 29 b. shows results of the forward selection technique, Table29 c. presents the error rate from both techniques and Table 29 d. exhibits results corresponding to the ANN. Tables 30 and 31 present the associated results for Early and Late rain season models.

Table 29.a: Results Dry season from Group variable selection technique - RH

Area	R^2	MAE (%)	Important variables
MAC region	0.5026	8.3576	X14, X16, X9, X17, X15, X10, X13, X11, X6
Antilles	0.5506	7.1038	X14, X16, X17, X7, X8, X2, X4, X3, X5
South America	0.5131	6.9760	X10, X16, X14, X2, X4, X1, X8, X5, X3, X9
Center America	0.5321	8.6240	X14, X16, X9, X17, X12, X1, X2, X15, X7
USA	0.5113	9.6825	X14, X16, X3, X13, X17, X11, X15, X6

Table 29.b: Results Dry season from forward selection technique - RH

Area	R^2	MAE (%)	Important variables
MAC region	0.5088	8.3346	X14, X16, X9, X17, X15, X1, X13, X10, X5, X6, X11, X3,
Antilles	0.5557	7.0672	X14, X16, X9, X17, X7, X15, X13, X10, X4, X2, X3, X11, X5
South America	0.5294	6.8343	X16, X14, X2, X4, X15, X1, X17, X5, X9, X5, X3, X11, X12, X6
Center America	0.5320	8.6356	X14, X16, X9, X17, X1, X2, X15, X10, X3, X4, X13, X5, X11, X6
USA	0.5557	9.2950	X14, X16, X3, X5, X7, X13, X17, X8, X15, X4, X1

Table 29.c: Dry season: Error rate - RH

Area	Error rate (Group variable selection) (%)	Error rate (Forward selection) (%)
MAC region	9.84	9.81
Antilles	9.77	9.72
South America	9.03	8.85
Center America	10.73	10.74
USA	11.40	10.94

Table 29.d: Results Dry season from ANN - RH

Area	R^2	MAE (%)	Error Rate (%)
Antilles	0.6344	6.2995	8.66
South America	0.5676	6.4912	8.41
Center America	0.5892	8.0238	9.98
USA	0.6331	8.2587	9.72

Table 30.a: Results Early rain season from Group variable selection technique - RH

Area	R^2	MAE (%)	Important variables
MAC region	0.5530	7.6164	X10, X16, X14, X7, X17, X15, X13, X11, X6
Antilles	0.6276	6.2190	X14, X16, X12, X17, X2, X4, X8, X13, X15
South America	0.4282	7.2541	X6, X16, X14, X17, X11, X2, X9, X15, X7, X12
Center America	0.5122	8.3875	X10, X16, X9, X14, X6, X4, X5, X3, X2
USA	0.6212	7.8996	X10, X16, X3, X14, X4, X5, X6, X2, X1

Table 30.b: Results Early rain season from Forward selection technique - RH

Area	R^2	MAE (%)	Important variables
MAC region	0.5809	7.3700	X10, X16, X14, X9, X17, X15, X1, X2, X13, X5, X3, X4, X11, X6
Antilles	0.6362	6.1214	X14, X16, X17, X2, X4, X9, X15, X5, X1, X10, X11, X13, X3, X6
South America	0.4306	7.2160	X6, X16, X14, X17, X11, X2, X4, X15, X5, X3, X1, X12, X13
Center America	0.5319	8.1764	X10, X16, X9, X14, X17, X15, X13, X4, X5, X3, X2
USA	0.6289	7.8193	X16, X3, X14, X15, X13, X17, X9, X1, X2, X6, X4, X5

Table 30.c: Early rain season: Error rate - RH

Area	Error rate (Group variable selection) (%)	Error rate (Forward selection) (%)
MAC region	8.60	8.32
Antilles	8.86	8.72
South America	9.09	9.05
Center America	10.24	9.98
USA	8.92	8.82

Table 30.d: Results Early rain season from ANN - RH

Area	R^2	MAE (%)	Error Rate (%)
Antilles	0.7221	5.2862	7.53
South America	0.5060	6.7379	8.45
Center America	0.6027	7.4449	9.09
USA	0.6945	7.0014	7.90

Table 31.a: Results Late rain season from Group variable selection technique - RH

Area	R^2	MAE (%)	Important variables
MAC region	0.5714	7.7582	X14, X16, X9, X12, X2, X7, X1, X4, X5, X6
Antilles	0.5662	5.8829	X14, X16, X7, X17, X2, X8, X3, X15, X9
South America	0.4487	7.3207	X14, X16, X17, X11, X1, X10, X15, X6, X13
Center America	0.5235	7.0731	X14, X16, X7, X8, X2, X3, X12, X9, X5, X1
USA	0.6419	8.7088	X14, X16, X3, X15, X11, X17, X13, X7

Table 31.b: Results Late rain season from Forward selection technique - RH

Area	R^2	MAE (%)	Important variables
MAC region	0.5828	7.5391	X14, X16, X9, X2, X17, X13, X15, X4, X3, X5, X1, X6, X10, X11
Antilles	0.5591	5.9225	X14, X16, X17, X2, X3, X6, X15, X13, X9, X10, X5, X4
South America	0.4796	7.0149	X14, X16, X2, X17, X1, X11, X9, X6, X4, X12, X15, X3, X5
Center America	0.5475	6.8131	X14, X16, X17, X2, X3, X9, X15, X1, X11, X5, X10, X4, X13
USA	0.6715	8.3489	X14, X16, X2, X1, X15, X11, X17, X3, X5, X13, X9, X4

Table 31.c: Late rain season: Error rate - RH

Area	Error rate (Group variable selection) (%)	Error rate (Forward selection) (%)
MAC region	8.84	8.59
Antilles	8.52	8.58
South America	9.37	8.98
Center America	8.80	8.47
USA	9.92	9.51

Table 31.d: Results Late rain season from ANN - RH

Area	R^2	MAE (%)	Error Rate (%)
Antilles	0.6200	5.4237	8.58
South America	0.4825	7.0024	8.98
Center America	0.6279	6.1855	8.47
USA	0.7113	7.6602	9.51

It is noticeable that similar from the models obtained in the previous section, the best results were obtained modeling RH based on ANN. Results shows that the computational time grows exponentially when ANN was trained. This time was reduced after dividing the MAC region into homogeneous zones.

In terms of the results itself once again is noticeable that models provide good estimations of RH, but South America is still being the hardest region to estimate compared to the results obtained on the other regions. In terms of the rainy seasons, the hardest region to estimate RH was the dry season which provides the lowest results in terms of the error, error rate and R^2 coefficient.

The sinusoidal functions resulted important predictors to estimate RH. The position and time related variables resulted also important variables to estimate RH. The satellite variables BT

and some of the physical parameters contribute with a set of important variables, but this set is different depending on the combination of season and region that was modeled.

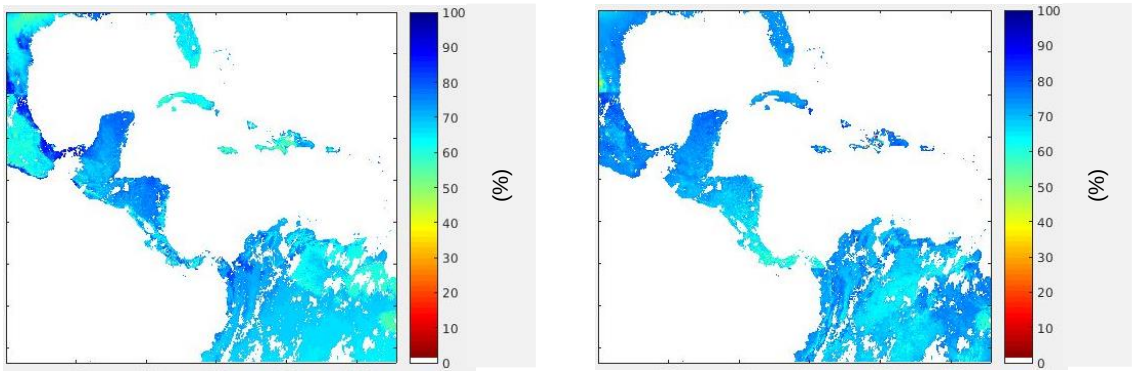
Comparing the results obtained in this section with the models from MODIS Aqua, both models resulted competitive. In general terms, it is difficult to determine which of those models provide the best set of estimators. Most of the input variables remains the same for both models except for three variables: LST, PW and NDVI.

It has been observed that even using a nonlinear estimation technique (ANN) the residuals still exhibits problems related to the constant variance and the independence. To attend these problems, require further research.

8.3.3 MODEL EVALUATION

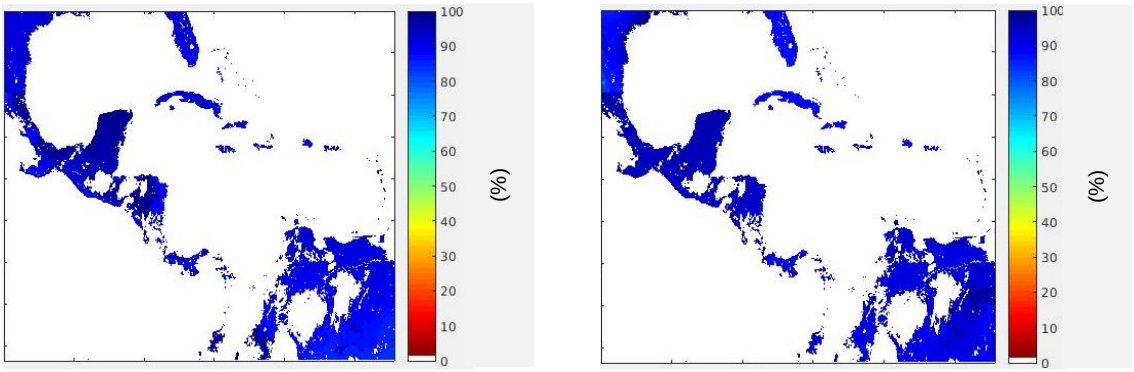
The model has been developed using 584 weather stations. However, the main objective of this work is to derive hourly and gridded estimates of RH; therefore, model evaluation produces the RH over the MAC region.

Model evaluation provides images of estimated RH from the entire MAC region during every hour during the training period. Each image has been studied to determine how adequate are the estimation for a specific hour of time and for a specific region. The change of the values during the day and night cycles and during the different months and seasons have also been analyzed. However, it is not possible to present the entire set of developed images based on the huge number of images that has been generated and studied. Figure 29 shows estimates of RH for August 15, 2011 during the daytime (at 18 UTC) the left panel exhibits estimates based on Aqua and the right panel shows estimates based on Terra. Figure 30 is similar but it present estimates for August 15, 2011 during nighttime (at 08 UTC).



a. b.

Figure 29: panel a.: Modeled RH Trained using MODIS Aqua. Panel b.: Modeled RH Trained using MODIS Terra Date: August 15 2011 at 18:00 UTC.



a. b.

Figure 30: panel a.: Modeled RH Trained using MODIS Aqua. Panel b.: Modeled RH Trained using MODIS Terra Date: August 15 2011 at 08:00 UTC.

Results shows that the estimations follow correctly the variation of RH throughout the day, in general it is observed that estimated RH increase during the night hours reaching its peak during the early morning and their minimum values during the daytime about the 18 UTC, but with small time shifts depending on the local times. This effect is clearly appreciated in the figures, for the 08 UTC image, the values or RH are very high about 80 or 90%. In contrast, at the 18 UTC images the values of RH decreases to about 60%. Also, it is observed that the minimum values do not fall below 40% as it was expected. Thus, it is not expected to see values of RH close to 0% specially in a month that is the starting of the late rain season where the large percentage of RH is expected.

A particular effect is noticed on the South America estimations. The models provide acceptable results on the training process, however looking the evaluation images it is noticeable that it has some limitations and issues. The estimations follow correctly the changes for the day night cycles, but they look very even specially for an area with a complex geography. It was expected that the influence of the amazon forest (the Andean mountains) would be more notorious the estimation of RH but it was not the case. It can be explained looking for the location and quantity of the stations used to train the model, most of these were located in the coastal areas of this region and in consequence the effect of the elevation and NDVI were marginal. This effect was not completely noticeable in the training process and is not expected to be noticed in the validation stage, because in those stages the estimations of the stations were only compared with these stations and their position. To solve this limitation, it is necessary to include weather station located in the mountains to obtain a more representative sample.

Even that, the evaluations have shown that the models develop are capable to offer estimations of RH that matches with the real behavior of the RH product. Even the limitations with South America, the model follows the patterns and trends that are expected for the estimates of RH. It observed that estimates of RH follow the correct day night cycles and they follow the changes in value expected for the mountains, deserts, Caribbean islands. In general, it presents a correct panorama of the behavior of the RH in areas with low quantity of stations, which was one of the objectives proposed for this model. However, this evaluation was performed under the same period of time used to train the model, it is necessary to validate this model in order to conclude that it can be used as a tool to estimate RH.

Comparing the two different set of developed models, it can be confirmed that both provides similar values of RH. However, results based on MODIS Aqua is quite superior to represent the joint areas between different zones and this effect can be appreciated in the limits of center and North America for example, and have a bigger level of contrast between areas with high RH and areas with low values of RH.

8.3.4 VALIDATION

Validation consists on comparing a set of RH model estimates with station observations. It is based on the ANN models, which have resulted to be the best technique for this application. It is important to clarify in the validation, even when the estimations were obtained from different models depending on the region, the comparison are performed based on the MAC region indistinctly of those regions. This is to be coherent with the objective to model the entire MAC region in a correct way.

Table 32 shows validation results in terms of four performance metrics MAE, RMSE error rate and R^2. This table is divided in two sub tables: a. presents results when using input variables from MODIS Aqua, b. table shows results based on MODIS Terra. Each of the sub tables present the results corresponding to each individual season.

Table 32.a: RH Validation results: MODIS Aqua models

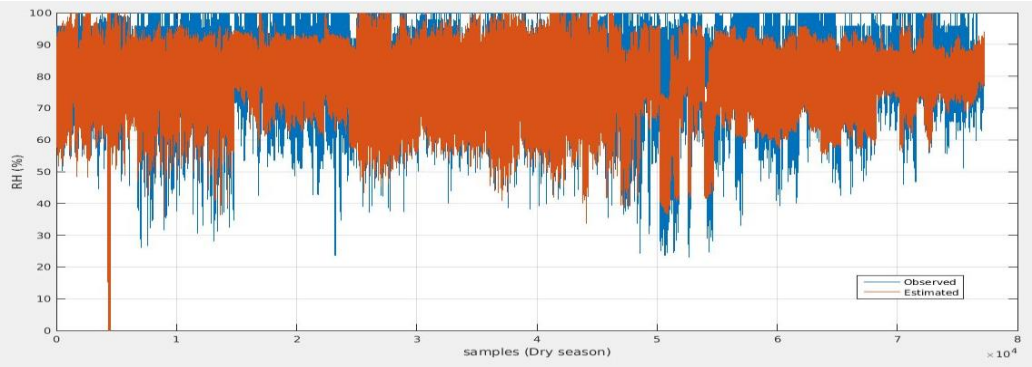
Season	MAE (%)	RMSE (%)	R^2	Error rate (%)
Dry season	7.3800	9.6350	0.4911	9.59%
Early Rain Season	5.7822	7.7372	0.6479	7.58%
Late Rain Season	5.7579	7.7068	0.6982	7.46%

Table 32.b: RH Validation results: MODIS Terra models

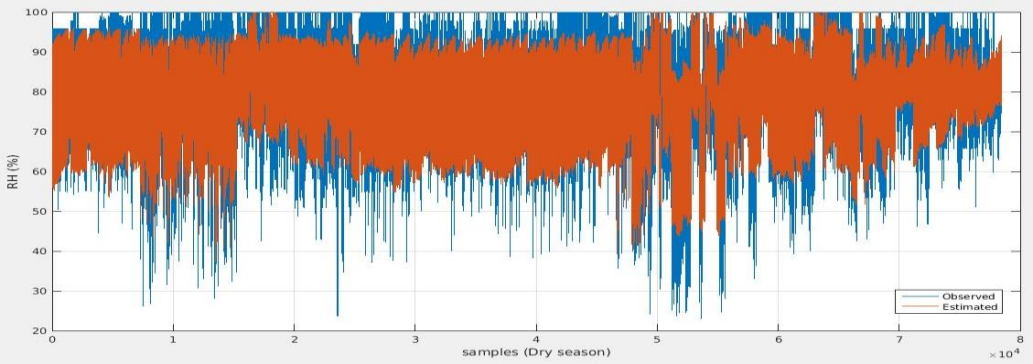
Season	MAE (%)	RMSE (%)	R^2	Error rate (%)
Dry season	7.5456	10.0054	0.4522	9.81%
Early Rain Season	5.9381	8.1868	0.6010	7.78%
Late Rain Season	5.7869	7.7463	0.6979	7.49%

It is observed that the difference between the observations and estimations remains about the same in comparison with what has been observed on the training results. The four metrics are being evaluated using a different period of time and the results matched the observed during the training process being very close one to the other. It is good to observe these kinds of results and to observe how small was in general the error and the error rate. All the error performance metrics were in general below 10%. This is a positive sign since in different times and space the models provided satisfactory estimates.

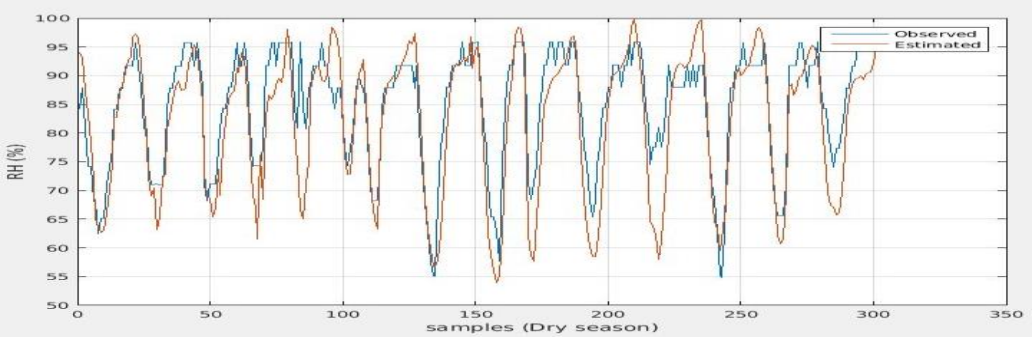
To represent in a more visual way the difference between the estimations and observations of RH, a group of figures corresponding to the time series of both for the validation set are presented below. There are presented three figures: 31, 32 and 33; one for each of the validation months, and each figure was divided into 4 panels: panel a. compares observations versus estimations based on MODIS Terra models and panel b. observation versus estimations based on MODIS Aqua models. Panels c. and d. present the time series for a small sample of data to observe in more detail the difference between observations and estimations.



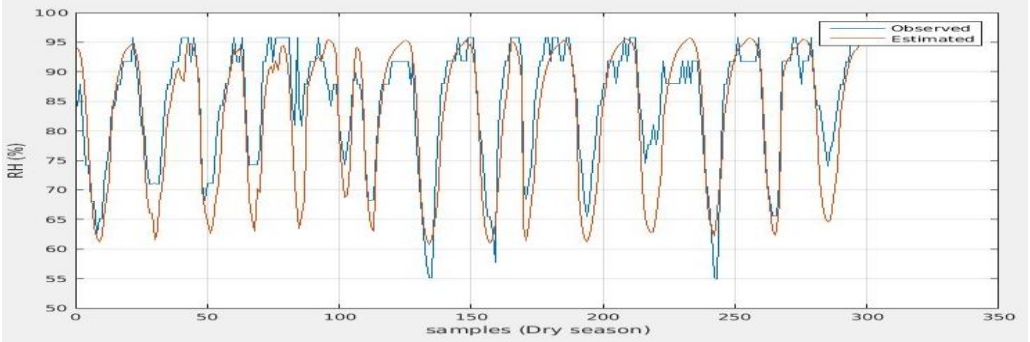
a.



b.

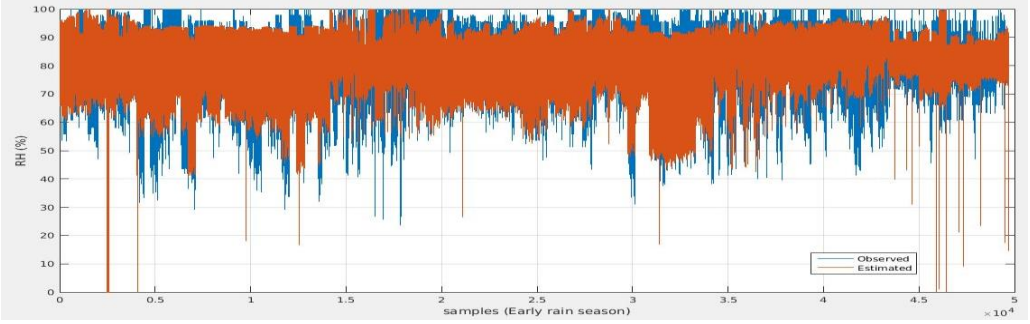


c.

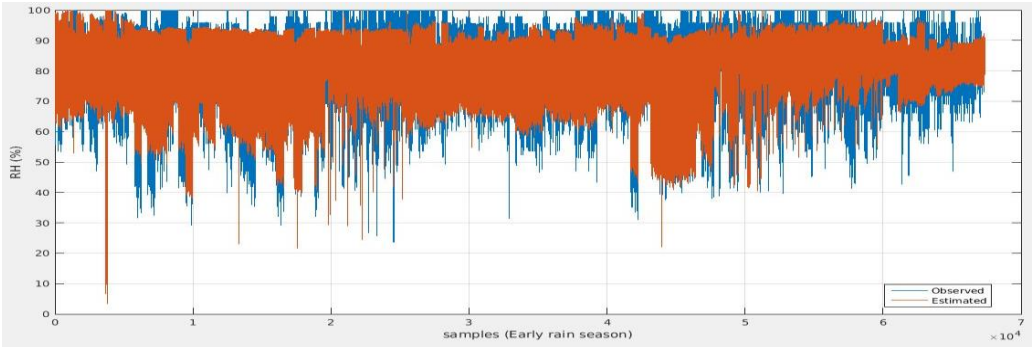


d.

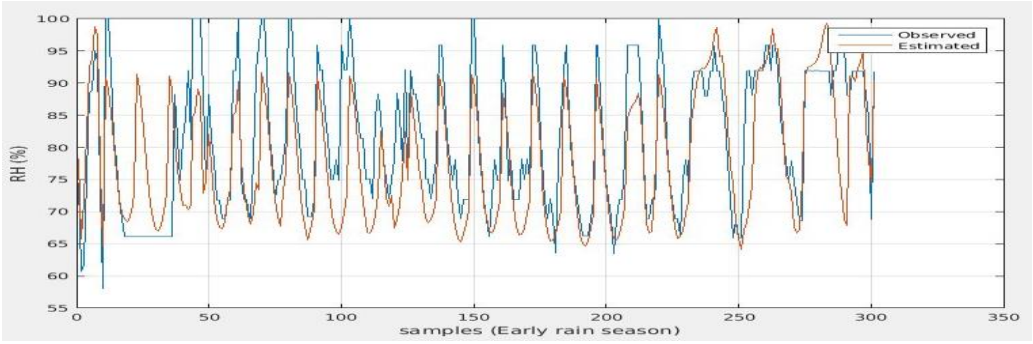
Figure 31: Time series December 2011. Panel a: observations vs MODIS Terra model. Panel b.: Observation vs MODIS Aqua model. Panel c. observations vs MODIS Terra model (sample). Panel d. observations vs MODIS Aqua model (sample).



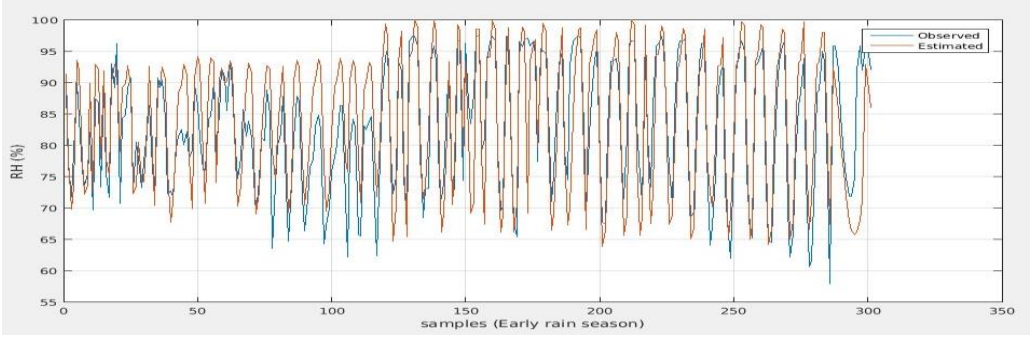
a.



b.

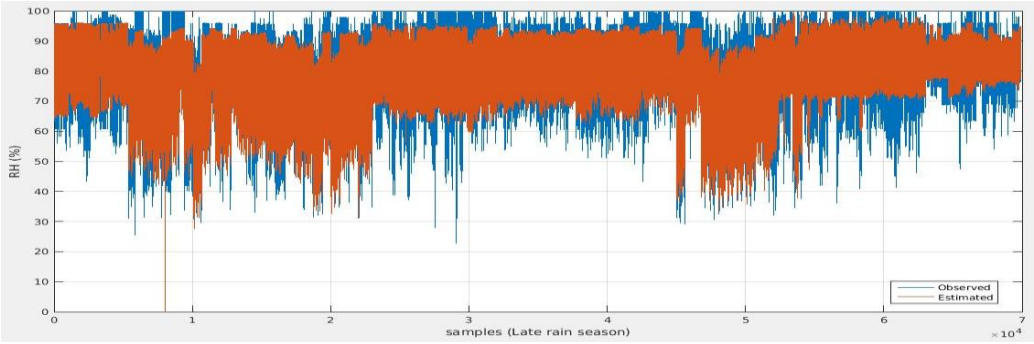


c.

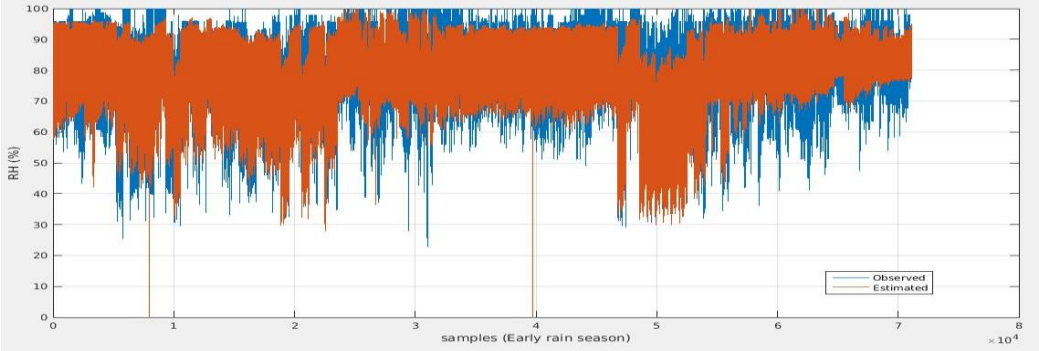


d.

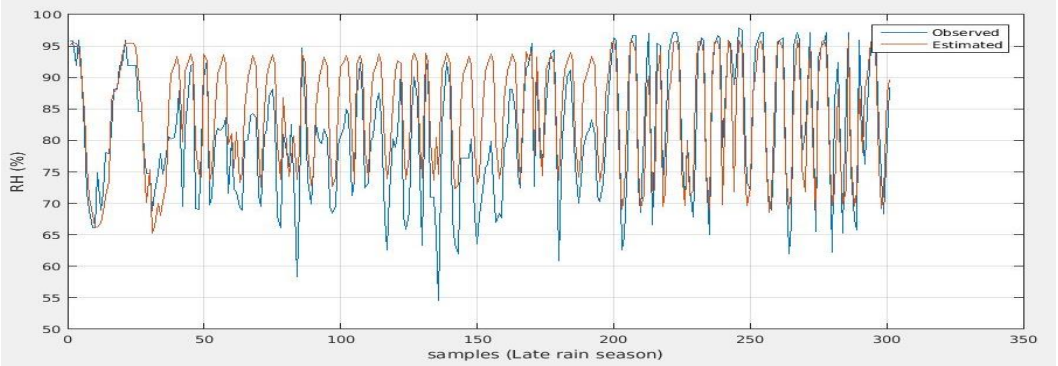
Figure 32: Time series July 2012. Panel a: observations vs MODIS Terra model. Panel b.: Observation vs MODIS Aqua model. Panel c. observations vs MODIS Terra model (sample). Panel d. observations vs MODIS Aqua model (sample).



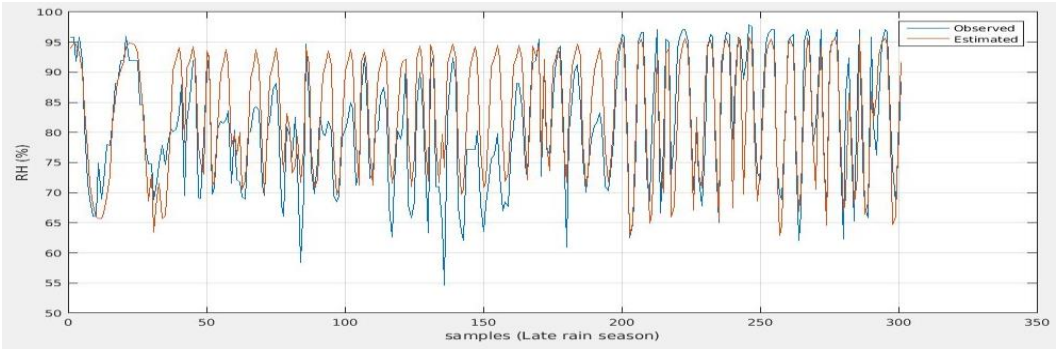
a.



b.



c.



d.

Figure 33: Time series August 2012. Panel a: observations vs MODIS Terra model. Panel b.: Observation vs MODIS Aqua model. Panel c. observations vs MODIS Terra model (sample). Panel d. observations vs MODIS Aqua model (sample).

The RH errors between MODIS Terra and MODIS Aqua appear to be small. More important, the figures 31, 32 and 33 show that MODIS Aqua provides a more stable set of estimations than MODIS Terra.

It can be noted that both models provide a good set of RH, however in some the estimations do not cover the entire range of values given by the station dataset.

9. CONCLUSIONS

9.1 CONCLUSIONS CHAPTER 6

In this chapter, it was proposed to estimate RH based on variables from remote sensing. A group of models were developed to test the importance of three proposed physical parameters (LST, PW and NDVI) in the estimation of RH. These physical parameters are obtained from MODIS instruments. The conclusions are presented below:

- It has been confirmed, based on the results, that the proposed physical parameters result significant to estimate RH. In each model, at least one of the physical parameters variations appear as significant variables. It is also noted that to obtain a good representation of RH it is necessary to include some other variables as the location or variables related with the time of the observation. Furthermore, it may be necessary the inclusion of some other variables to represent the variability and behavior of this product in a more precise way. In future experiments, it is recommended to study new variables that may be included to improve the model result.
- From the two MODIS datasets used to train the models, MODIS Aqua provides better estimations compared with MODIS Terra. It is theorized that this may be related to the time when the observations are taken. MODIS Aqua provides images during the early morning and about 18 or 19 both in UTC time. And those can be related to the time where the higher and lower observation of RH are captured in the stations. On the other hand, MODIS Terra provide images approximately at 2 to 3 and about 15 to 17 hours, both in UTC, when there is large variability of RH, and in consequence introducing larger errors.
- The results from the best set of models provides R^2 between 0.68 and 0.81 and the MAE has values between 7.33 and 9.63 percent. Those results are acceptable but it is used just as a reference point. In future experiments, it might be obtained different result in the estimation of the same product.
- MODIS physical parameters in general provides a good set of variables to estimate RH. However, the amount of information obtained is not enough to generate a model that fit the required characteristics of hourly observations over the entire MAC region.

Furthermore, it is observed the importance of the hour as a variable to estimate RH as well as the position and elevation component. However, LST and PW observations from MODIS are limited to only twice a day. To solve this limitation, it is proposed to estimate LST and PW using data from another satellite that increase the amount of available data to hourly observations and covering MAC region in every observation, and the selected satellite is GOES-13.

9.2 CONCLUSIONS CHAPTER 7

This second step in the process of estimating RH, has resulted in an improvement of the limited information provided for MODIS. The major conclusions of this chapters are summarized below:

- Regression techniques with MODIS and GOES data is a feasible and efficient alternative to develop the gridded and hourly estimates of LST, PW and NDVI, which are required to estimate the RH
- It has been confirmed the importance to include the elevation, the time and the position to explain the variability of LST and PW.
- It has been concluded that better estimates are usually obtained when the MAC region is divided in Homogeneous zones.
- The division of the year in different seasons provide good results in terms of the estimations. It has helped to solve two different problems that are the processing time, and to reduce the problems with the variation distributing the period in seasons with defined characteristics.
- There are residual problems related to the independence and the constant variance. However, these problems may be associated to the data and a more advance transformation process or a more detail dataset would help on the solution of it.
- The evaluation of the models shows how good the models follow the natural behavior of the observations. It is also confirmed in the validation stage where the models provide low errors compared to the range of values from the variables. However, there is a problem with the validation for the PW that may be explained by the nature of the selected months.

9.3 CONCLUSIONS CHAPTER 8

A new set of models to estimate RH from satellite data have been developed. These new models improved the developed product because it is trained with more information and are capable to produce estimations with a most accurate hourly interval.

- The proposed model provides hourly RH estimations at 4km spatial resolutions over the entire MAC region under clear sky condition.
- The implementation of the ANN resulted in improves in the estimation of RH. It has constantly produced the best results when compared to the regression techniques. Also, there is observed better results when using a nonlinear transfer function, which provides evidence that the RH and their input variables follows a nonlinear behavior that requires more advanced techniques to fully understand it. The main drawback of the ANN is the large computational time required for model training.
- It has been shown that there is a small improvement in the models trained using LST and PW estimated from MODIS Aqua and NDVI also observed from MODIS Aqua, this advantage is hard to observe in the metrics, but it became more evident during the validation and evaluation process.
- There persist some problems related to the residuals related to the independence and the constant variance. Thus, it is required to perform further research to mitigate this problem.
- South America is one of the hardest region to estimate RH. It is caused by the lowest quantity of information obtained from this region. The selected sample is not representative of the characteristics of this complex region.
- It is important to include the seasonal components to captured the variations of the RH during the 24 hours' variation. At least one of the sinusoidal functions appears to be significant variables in each of the models. During the evaluation period, it can be observed that this change of RH during the daytime correspond to the real changes. This effect is also observed in the validation process noticing that it corresponds to the real shift observed in the stations.

9.4 GENERAL CONCLUSIONS

- It has been shown the importance of Industrial Engineering techniques as: regression, optimization, quality control and neural network in the solution of problems related to meteorological applications.
- The proposed model is capable to estimate hourly RH, LST and PW at 4km spatial resolutions over the entire MAC region under clear sky condition.
- The inclusion of the seasonal components, division in homogeneous zones and in rain seasons resulted important to improve the performance of the RH estimations.
- It was noticed that it is hardly to generate estimation mostly in the southern region. The proposed model can be improved by using additional transformations to stabilize the variance and independence on residuals.

10. CONTRIBUTIONS

This work helped in the estimation of an important atmospheric product which is the RH. This product can be estimated now in a spatial resolution of 4 km and every hour. Also, this product has been trained to cover the entire MAC region offering estimation for all the land covered areas that are under clear sky conditions.

This models contribute as the possibility to obtain an operational product from RH, product that can be useful in the calculations of products as Heat index. Also, thank to this model now it can be developed an operational RH product that can be introduced as an input variable in correlation analysis with different atmospheric and non-atmospheric variables as energy consumption.

This model also offer a model trained with a big amount of data and in a good time resolution. It was trained using an entire year of data divided in 3 different rain seasons. Also, the validation time include one entire month per each season.

11. FUTURE WORK

Several models have been developed, evaluated and validated to estimate RH for the MAC region. However, these models are not perfect, as it was discussed in the previous chapters. Here a group of alternatives are introduced to improve current work:

- Include some other physical parameters and satellite products as input variables. Some of those products are the air temperature that could also be estimated from MODIS data. Also, it will be recommended to include some other RH humidity products from numerical models, as the one offered by NCEP that provides a low-resolution product. This product may be included as an input variable to increase the level of variability captured by the model.
- It is necessary to solve the problems of instable variance and autocorrelation in the residuals on the models for LST, PW and RH. Different approximation has been tested to achieve a solution; however, advanced techniques specifically designed for this end should be implemented. It is necessary to include some advanced transformation combined with a new data segmentation based in more advanced clustering techniques. It is expected that those implementations may be helpful to reduce these problems.
- The RH model requires of the LST, PW and NDVI data in hourly basis. To derive these variables regression models were developed; however, these models include some estimation errors. Thus, to avoid these errors it would be desirable to have measurements of the actual physical parameters. In the future, it is recommended to replace the estimation with direct measurements from GOES-R which was launched in November 19, 2016. In the near future, this satellite will offer LST, PW, NDVI and other products that can be used to improve the RH (GOES-R, 2016 and GOES-R. b, 2016). It is expected that this change will reduce the error in the estimation of RH, specifically the error generated from LST and PW.

12. REFERENCES

Ahrens, C. Donald (2013). *Meteorology Today: an introduction to weather, climate, and the environment (Tenth Edition)*. Belmont, CA. Brooks/Cole, Cengage Learning.

Air University (U.S.). Air Command and Staff College Space Research Elective Seminars (ACSC). (2009). *AU-18 Space Primer*. Maxwell Air Force Base, Alabama: Air University Press.

Akhoondzadeh, M. and Saradjian, M.R. (2008). Comparison of Land Surface Temperature Mapping Using MODIS and ASTER Images in Semi-Arid Area. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 873-876.

Akbari, V., Amini, J., Saradjian, M. R., and Motagh, M. (2008). Estimation of atmospheric temperature and humidity profiles from MODIS and radiosond data using artificial neural network. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Beijing*, *37*, 35-40

AMETSOC, 2016. "American Meteorological Society glossary of meteorology." [Online]. Available: http://glossary.ametsoc.org/wiki/Precipitable_water [Accessed: 18-March-2016].

Ba, M. B. and Gruber, A. (2001). GOES Multispectral Rainfall Algorithm (GMSRA). *Journal of Applied Meteorology, 40(8)*, 1500-1514.

BOM, 2016. "Australian Government Bureau of meteorology" [Online]. Available: http://www.bom.gov.au/climate/austmaps/about-ndvi-maps.shtml [Accessed: 19-March-2016].

Brockwell, Peter J. and Davis, Richard A. (2002). *Introduction to Time Series and Forecasting (Second Edition)*. New York: Springer-Verlag New York, Inc.

Castro, J. M. (2007). *Nuevas estrategias para pronosticar la trayectoria e intensidad de los huracanes en el Oceano Atlantico* (Unpublished MS dissertation). University of Puerto Rico Mayaguez Campus. Puerto Rico, USA.

Doggett, L.E., Tangren, J.T., and Panossian, S.P. (1990). *Almanac for Computers 1990*. , Washington DC: U.S. Government Printing Office.

GES DISC, 2016. “Goddard Earth Sciences Data and Information Services Center”. [Online]. Available:http://disc.sci.gsfc.nasa.gov/data-holdings/PIP/brightness_temperature.shtml [Accessed: 22-December-2016]

GOES-R, 2016. “GOES-R Series Products”. [Online]. Available: http://www.goes-r.gov/products/overview.html [Accessed: 30-December-2016]

GOES-R. b, 2016. “Advanced baseline imager (ABI)” [Online]. Available: http://www.goes-r.gov/spacesegment/abi.html [Accessed: 30-December 2016]

Gunst, R. F. and Webster, J. T. (1975). Regression analysis and problems of multicollinearity. *Communications in Statistics-Theory and Methods, 4(3),* 277-292.

Hagan, M.T., Demuth, H.B., Beale M.H. and De Jesus O. (2014). *Neural Network Design (Second Ed.).* Boston, MA.: PWS Publishing Company.

Han, K.S., Viau, A.A., Kim, Y.S. and Roujean, J.L. (2005). Statistical estimate of the hourly near-surface air humidity in eastern Canada in merging NOAA/AVHRR and GOES/IMAGER observations. *International Journal of Remote Sensing Vol. 26, No. 21*, 4763–4784.

Hillger, D.W. and Schmit, T. J. (2010). *The GOES-14 Science Test: Imager and Sounder Radiance and product validations.* Washington, DC: U.S. DEPARTMENT OF COMMERCE National Oceanic and Atmospheric Administration.

Ji Zhou, Xu Zhang, Wenfeng Zhan and Huailan Zhang (2014). Land Surface Temperature Retrieval from MODIS Data by Integrating Regression Models and the Genetic Algorithm in an Arid Region. *Remote Sensing 2014, 6,* 5344-5367.

Jun Cheng, Xin Wang, Tingting Si, Fan Zhou, Junhu Zhou and Kefa Cen. (2016). Ignition temperature and activation energy of power coal blends predicted with back-propagation neural network models. *Fuel 173*, 230-238.

Kuligowski, Robert J. (2002). A Self-Calibrating Real-Time GOES Rainfall Algorithm for Short-Term Rainfall Estimates. *Journal of Hydrometeorology, 3(2),* 112-130.

Kuligowski, Robert J., and Barros, Ana P. (2001). Combined IR-Microwave Satellite Retrieval of Temperature and Dewpoint Profiles Using Artificial Neural Networks. *Journal of Applied Meteorology*, *40*(11), 2051-2067.

Kunkel, K. E., Pielke Jr., R. A., and Changnon, S. A. (1999). Temporal fluctuations in weather and climate extremes that causes economic and human health impacts: A review. *Bulletin of the American Meteorological Society, 80(6)*, 1077-1098.

LAADS WEB, 2016. "LAADS WEB Level 1 and Atmosphere Archive and Distribution System" [Online]. Available: https://ladsweb.nascom.nasa.gov/data/ [Accessed: 08-April-2016].

Levy B.S., Wegman D.H., Baron S.L. and Sokas R.K. (2011). *Occupational and Environmental Health: Recognizing and Preventing Disease and Injury (Sixt Ed.).* New York, NY.: Oxford University Press, Inc.

Lewis-Beck, C. and Lewis-Beck, M. (2015). *Applied Regression: An Introduction (Second Edition) (Vol. 22)*. Sage Publications.

Linyi Li, Yun Chen, Tingbao Xu, Rui Liu, Kaifang Shi and Chang Huang (2015). Super-resolution mapping of wetland inundation from remote sensing imagery based on integration of back-propagation neural network and genetic algorithm. *Remote Sensing of Environment 164*, 142-154.

LPDAAC, 2016. "Land Processes Distributed Active Archive Center LP DAAC" [Online]. Available: https://lpdaac.usgs.gov/dataset_discovery/modis/modis_products_table/mod11_l2 [Accessed: 05-April-2016].

Marin, J.C., Pozo, D. and Curé, M. (2015). Estimating and forecasting the precipitable water vapor from GOES satellite data at high altitude sites. *Astronomy and Astrophysics 573, A41*.

MODIS, 2015. "MODIS Moderate Resolution Imaging Spectroradiometer" [Online]. Available: http://modis.gsfc.nasa.gov/about/ [Accessed: 22-November-2015].

MODIS, 2016. "MODIS Moderate Resolution Imaging Spectroradiometer - Specifications" [Online]. Available: http://modis.gsfc.nasa.gov/about/specifications.php [Accessed: 22-March-2016].

MODIS, 2017. "MODIS Atmosphere MOD05-L2. Product Description" [Online]. Available: http://modis-atmos.gsfc.nasa.gov/MOD05_L2/ [Accessed 28-February-2017]

Montgomery, D.C., Peck, E.A. and Vining, G.G. (2012). *Introduction to Linear Regression Analysis. (Fifth Ed.).* Hoboken, NJ.: John Wiley & Sons, Inc.

NASA, 2016. "My NASA data" [Online]. Available: http://mynasadata.larc.nasa.gov/glossary/relative-humidity-2/ [Accessed: 14-june-2016].

NCDC, 2016. "NOAA National Center for Environmental Information" [Online]. Available: http://www.ncdc.noaa.gov/data-access [Accessed: 06-July-2016].

NCEP, 2016. "NCEP/NCAR Reanalysis 1: Pressure" [Online]. Available: https://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.pressure.html [Accessed: 09-June-2016].

Newton, H.J. (1988). *TIMESLAB: A Time Series Analysis Laboratory*. Pacific Grove, CA.: Wadsworth & Brooks/Cole.

NGDC, 2016. "Digital Elevation Model DEM Discovery Portal" http://www.ngdc.noaa.gov/mgg/dem/demportal.html [Accessed: 05-June-2016]

NOAA, 2016. "NOAA Comprehensive Large Array-Data Stewardship System" [Online]. Available: http://www.class.ngdc.noaa.gov/saa/products/welcome [Accessed: 08-April-2016].

OSPO, 2016. "Conversion of GVAR Infrared Data to Scene Radiance or Temperature" [Online]. Available: http://www.ospo.noaa.gov/Operations/GOES/calibration/gvar-conversion.html [Accessed: 03-November-2016].

Peng Guangxiong, Li Jing, Chen Yunhao, Norizan Abdul P. and Tay Liphong. (2006). High-resolution Surface Relative Humidity Computation Using MODIS Image in Peninsular Malaysia. *Chinese Geographical Science, 16(3)*, 260–264.

Ramírez-Beltran Nazario D., Salazar Cesar M., Gonzalez Jorge E. and Castro Joan M. (2016). A Time Series Model and Satellite Data to Estimate Air Temperature. Proceedings of the IIE annual Conference: Industrial & Systems Engineering Research Conference (ISERC), 2016 H. Yang, Z. Kong, and Sarder, eds. May 21-24 2016, Anaheim, CA.

Ramírez-Beltran, N. D., Calderon Arteaga, C., Harmsen E., Vasquez, R., and Gonzalez, J. (2010). An Algorithm to estimate soil moisture over vegetated areas based on in situ and remote sensing information. *International Journal of Remote Sensing, 1366-5901, Volume 31, Issue 10, 2016*, 2655-2679.

Ramírez-Beltran, N. D., Kuligowski, R. J., Castro, J. M., Cardona-Soto, M., Vasquez, R. (2009). A Projection Algorithm for Satellite Rainfall Detection. *WSEAS Transactions on Systems, Issue 6, Vol. 8*, pp. 763-772.

Ramírez-Beltran, N.D, Lau, W. K., Winter, A., Castro, J. M. and Escalante, N. R. (2007). Empirical probability models to predict precipitation levels over Puerto Rico stations. *Monthly weather review, 135(3)*, 877-890

Rawlings J. O., Pantula S. G. and Dickey D. A. (1998). *Applied Regression Analysis. A Research Tool (Second Ed.)*. New York.: Springer-Verlag New York, Inc.

Rumerman, J. A. (2009). *NASA Historical Data Book, Vol. VII: NASA Launch Systems, Space Transportation, Human Spaceflight, and Space Science, 1989-1998*. Washington D. C.: U.S. Government Printing Office.

Ulivieri, C., Castronouvo, M.M., Francioni, R. and Cardillo, A., (1994). A split-window algorithm for estimating land surface temperature from satellites. *Advances in Space Research, 14(3),* 59-65.

Verran, J. A. and Ferketich, S. L. (1984). Residual analysis for statistical assumptions of regression equations. *Western Journal of Nursing Research, 6(1), 27-40.*

Wenhui Wang, Shunlin Liang and Tilden Meyers (2008). Validating MODIS land surface temperature products using long-term nighttime ground measurements. *Remote Sensing of Environment,* 623-635.

Williams, 2016. "Sunrise Sunset Algorithm" [Online]. Available: http://williams.best.vwh.net/sunrise_sunset_algorithm.htm [Accessed: 22-December-2016]

13. APPENDICES

13.1 APPENDIX 1

In this appendix, it is presented an example of the code developed using MATLAB software. This code corresponds to the model training process to estimate RH, based on MODIS Aqua, in the final stage (explained in chapter 8). Also, this code requires functions for specific tasks as the regression or ANN functions. The code is presented bellow

```
clear all;
clc;
close all;
warning off all
%%Load
tables%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%
directory='/usr/local/Data/jean/Estimate Relative Humidity/Tables AQUA/1 Create Inputs';
file='tabledata_rh_aqua_late_2011.mat';
regressiontable=table_generator(directory,file);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%
tic
%% Creating clusters
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
aas1=find(regressiontable(:,2)<=25 & regressiontable(:,2)>=16 & regressiontable(:,3)<=-60 & regressiontable(:,3)>=-85);
aas2=find(regressiontable(:,2)<=16 & regressiontable(:,2)>=12 & regressiontable(:,3)<=-60 & regressiontable(:,3)>=-65);
aas3=find(regressiontable(:,2)<=30 & regressiontable(:,2)>=25 & regressiontable(:,3)<=-60 & regressiontable(:,3)>=-87);
aas=[aas1;aas2;aas3]; %Lesser Antilles
mas=setdiff([1:length(regressiontable)]',[aas]); %Greater Antilles
reg=regressiontable;
reg([aas],2)=-9999;
usa=find(reg(:,2)>23); %USA
cas=find(reg(:,2)<=23 & reg(:,2)>=8 & regressiontable(:,3)<=-77); %Central America
sas=setdiff(mas,[usa;cas]); %South America
%%Selecting a particular region%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
regressiontable=regressiontable(sas,:);
seltime=find(regressiontable(:,4)>=0);
```

```
regressiontable=regressiontable(seltime,:);

seltime2=find(regressiontable(:,8)>=0);

regressiontable=regressiontable(seltime2,:);

%%Dividing Regressors and response variables%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

x1=regressiontable(:,[2:4 6:8 11:13 15:18 19:22]);

y=regressiontable(:,1);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

clear regressiontable reg

%%Model Training%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

n=size(x1);

toc

tic

x=x1;

disp('Division Method')        %% Division method

[varws3,xwins3,beta3,R23,pval3,t03]=division_sel(1,10,x,y,0.95);

disp('VIF')

[Xa3,vars3a,beta3a,R23a,VIF3a,pval3a,t03a]=variance_factor(y,xwins3,5);

ye5=[ones(n(1),1) x(:,varws3)]*beta3';

ye6=[ones(n(1),1) Xa3]*beta3a';

%%

disp('Forward Selection')       %% Forward selection

[R24,Xs4,beta4,Ye4,pval4,Fu4,vars4,t04]=forward_sel(y,x,0.95);

disp('VIF')

[Xb4,vars4a,beta4a,R24a,VIF4a,pval4a,t04a]=variance_factor(y,Xs4,5);

ye7=Ye4;

Xb4=[ones(n(1),1) Xb4];

ye8=Xb4*beta4a';

%%

R2p2=[R23 R23a;R24 R24a];

%

R2f(1)=R23a;

R2f(2)=R24a;

MAEf(1)=sum(abs(y-ye6))/length(y);

MAEf(2)=sum(abs(y-ye8))/length(y);

RMSEf(1)=sqrt(sum((y-ye6).^2)/length(y));

RMSEf(2)=sqrt(sum((y-ye8).^2)/length(y));
```

```
errate(1)=mean(abs(y-ye6))/(max(y)-min(y));
errate(2)=mean(abs(y-ye8))/(max(y)-min(y));

figure
plot([y ye5 ye6 ye7 ye8],'linewidth',2);
grid
title('Relative Humidity')
xlabel('samples (Center America)')
ylabel('RH')
legend({'Observed';'Division Method';'Division Method VIF';'Forward Sel';'Forward Sel VIF'});
toc

ed=(sum(abs(y-ye6)))/length(y);
ef=(sum(abs(y-ye8)))/length(y);

savevar1=varws3(vars3a);
savevar2=vars4(vars4a);

if(R24a>=R23a)
    xnn=x(:,vars4(vars4a));
    savevar3=vars4(vars4a);
else
    xnn=x(:,varws3(vars3a));
    savevar3=varws3(vars3a);
end

net=tryNN_fun(xnn,y);  %%ANN
ye9=sim(net,xnn');
ye9=ye9';


figure
plot([y ye9],'linewidth',2);
grid
title('Relative Humidity')
xlabel('samples (South America)')
ylabel('RH')
legend({'Observed';'Neural Network'});
```

```
e9=y-ye9;
SSE=e9'*e9;
SST=y'*y-length(y)*mean(y)^2;
R29=1-SSE/SST;
en=(sum(abs(y-ye9)))/length(y);
beta5a=net;

R2f(3)=R29;
MAEf(3)=sum(abs(y-ye9))/length(y);
RMSEf(3)=sqrt(sum((y-ye9).^2)/length(y));
errate(3)=mean(abs(y-ye9))/(max(y)-min(y));
%%Figure obs. vs
est.%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
figure
subplot(2,2,1)
plot([y ye6 ye8 ye9],'linewidth',2);
grid
title('Relative Humidity')
xlabel('samples (South America)')
ylabel('RH')
legend({'Observed';'Division Method VIF';'Forward Sel VIF';'Neural Network'});
subplot(2,2,2)
plot(y-ye6,'linewidth',2);
grid
title('Division Method')
xlabel('samples (South America)')
ylabel('RH')
%legend()
subplot(2,2,3)
plot(y-ye8,'linewidth',2);
grid
title('Foward Selection')
xlabel('samples (South America)')
ylabel('RH')
%legend()
subplot(2,2,4)
plot(y-ye9,'linewidth',2);
grid
title('Neural Network')
```

```
xlabel('samples (South America)')

ylabel('RH')

%legend()

%%Saving
results%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%

results=[MAEf' RMSEf' R2f' 100*errate'];

save rh_aqua_clear_coef_late_sas.mat R2p2 beta3a beta4a beta5a savevar1 savevar2 savevar3
```

13.2 APPENDIX 2

Calculation of the error rate:

$$Error\ rate = \frac{MAE}{max_o - min_o}$$

$$MAE = \frac{\sum |v_e - v_o|}{n_{el}}$$

Where:

- max_o = Maximum observed value.
- min_o = Minimum observed value.
- v_e = Estimated value
- v_o = Observed value
- n_{el} = Number of elements