COLEGIO DE POSTGRADUADOS
INSTITUCIÓN DE ENSEÑANZA E INVESTIGACIÓN EN CIENCIAS AGRÍCOLAS

CAMPUS CAMPECHE

POSGRADO EN BIOPROSPECCIÓN Y SUSTENTABILIDAD AGRÍCOLA EN EL TRÓPICO

MÉTODOS EMPÍRICOS Y DE SOFT-COMPUTING PARA ESTIMAR EVAPOTRANSPIRACIÓN DE REFERENCIA EN EL ESTADO DE CAMPECHE

LUIS ALBERTO RAMOS CIRILO

TESIS

PRESENTADA COMO REQUISITO PARCIAL
PARA OBTENER El GRADO DE

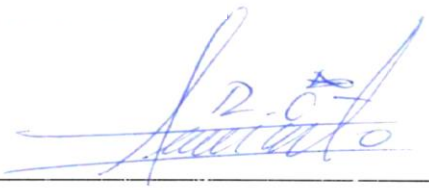
MAESTRO EN CIENCIAS

SIHOCHAC, CHAMPOTÓN, CAMPECHE

2019

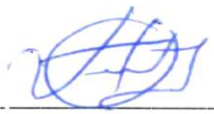
CARTA DE CONSENTIMIENTO DE USO DE LOS DERECHOS DE AUTOR Y DE LAS REGALIAS COMERCIALES DE PRODUCTOS DE INVESTIGACION DE PRODUCTOS DE INVESTIGACION

En adición al beneficio ético, moral y académico que he obtenido durante mis estudios en el Colegio de Postgraduados, el que suscribe **Luis Alberto Ramos Cirilo**, Alumno de esta institución, estoy de acuerdo en ser partícipe de las regalías económicas y/o académicas, de procedencia nacional e internacional, que se deriven del trabajo de investigación que realicé en esta institución, bajo la dirección del Profesor **Víctor Hugo Quej Chi**, por lo que otorgo los derechos de autor de mi tesis **"Métodos empíricos y de soft-computing para estimar evapotranspiración de referencia en el estado de Campeche"** y de los productos de dicha investigación al Colegio de Postgraduados. Las patentes y secretos que se pueden derivar serán registrados a nombre del Colegio de Postgraduados y las regalías económicas que se deriven serán distribuidas entre la Institución, El Consejero o Director de Tesis y El que suscribe, de acuerdo a las negociaciones entre las tres partes, por ello me comprometo a no realizar ninguna acción que dañe el proceso de explotación comercial de dichos productos a favor de esta institución.



Firma

Ing. Luis Alberto Ramos Cirilo



Vo.Bo. del Consejero o Director de Tesis

Dr. Víctor Hugo Quej Chi

La presente tesis, titulada: **Métodos empíricos y de soft-computing para estimar evapotranspiración de referencia en el estado de Campeche**, realizada por el alumno **Luis Alberto Ramos Cirilo** bajo la dirección del consejo particular indicado, ha sido aprobada por el mismo y aceptada como requisito parcial para obtener el grado de:

MAESTRO EN CIENCIAS

EN BIOPROSPECCIÓN Y SUSTENTABILIDAD AGRÍCOLA DEL TRÓPICO

CONSEJO PARTICULAR

DIRECTOR DE TESIS: _____

Dr. Víctor Hugo Quej Chi

ASESOR: _____

Dr. Everardo Aveces Navarro

ASESOR: _____

Dr. Eugenio Carrillo Ávila

ASESOR: _____

M.C. Benigno Rivera Hernández

Sihochac Campeche, 2019

MÉTODOS EMPÍRICOS Y DE SOFT-COMPUTING PARA ESTIMAR EVAPOTRANSPIRACIÓN DE REFERENCIA EN EL ESTADO DE CAMPECHE

Luis Alberto Ramos Cirilo, M. en C.

Colegio de Postgraduados, 2019

RESUMEN

En el presente estudio, utilizando solamente datos de temperatura, se evaluó el desempeño de tres modelos de soft-computing (máquinas de soporte vectorial (SVM), Programación de Expresión Genética (GEP) y XGBoost) y dos ecuaciones empíricas (Hargreaves- Samani y Camargo) para predecir la evapotranspiración de referencia (ETo) en el estado de Campeche, México. El desempeño de los modelos empíricos y de soft-computing se evaluaron de acuerdo a los índices estadísticos: Error Absoluto Medio (MAE), Raíz Cuadrada Media del Error (RMSE) y el coeficiente de determinación (R^2). Se evaluaron dos técnicas de interpolación de datos (PCHIP y SPLINE) siendo PCHIP la mejor técnica de interpolación para estimar valores faltantes en series históricas de datos meteorológicos; y tres técnicas de detección de datos atípicos (Grubbs, Cuartiles y Mean), siendo el método de Mean elegido como el mejor, ya que permite mayor tolerancia a valores atípicos causados por eventos de lluvias y/o nubosidades. Los resultados muestran que, entre los modelos empíricos evaluados, la ecuación de Camargo (MAE = 0.563, RMSE = 0.721 y R^2 = 0.723) obtuvo mejor eficiencia en la estimación de la ETo en comparación con la ecuación de Hargreaves-Samani (MAE = 0.588, RMSE = 0.750 y R^2 = 0.703) confirmando que el modelo funciona en climas cálido subhúmedo como el del estado de Campeche. Respecto a los modelos de soft-computing, el modelo SVM obtuvo mejor desempeño global entre las técnicas evaluadas (MAE = 0.480, RMSE = 0.637 y R^2 = 0.786) siendo el modelo recomendado para estimar la ETo en el estado de Campeche. El modelo GEP superó ligeramente a los modelos empíricos, sin embargo, tiene la ventaja de proporcionar un modelo algebraico programable en una hoja de cálculo para realizar predicciones de ETo, siendo otra opción viable en la determinación de la ETo en el estado de Campeche.

Palabras claves: Evapotranspiración de referencia; ecuaciones empíricas; técnicas soft-computing; estaciones meteorológicas automatizadas.

EMPIRICAL AND SOFT-COMPUTING METHODS TO ESTIMATE EVAPOTRANSPIRATION OF REFERENCE IN THE STATE OF CAMPECHE

Luis Alberto Ramos Cirilo, M. en C.

Colegio de Postgraduados, 2019

ABSTRACT

In the present study, the performance of three soft-computing models and two empirical equations were evaluated using only temperature data for predict the ETo in Campeche, México. The evaluated soft-computing models were support vector machines (SVM), Gene expression programing (GEP) and XGBoost, the empirical approaches Hargreaves-Samani and Camargo models were evaluated. The soft computing and empirical models performance were evaluated accord to the statistics rates; Mean absolute error (MAE), root mean square error (RMSE) and the determination coefficient (R^2). Two data interpolation techniques (PCHIP and SPLINE) were evaluated, with PCHIP being the best interpolation technique for estimating missing values in historical meteorological data series; and three outliers techniques (Grubbs, Quartiles and Mean), being the Mean method chosen as the best, since it allows greater tolerance to outliers caused by rainy events and / or cloudiness. As results of the empirical approaches, the Camargo model (MAE = 0.563, RMSE = 0.721 y R^2 = 0.723) obtained a better efficiency in the ETo prediction in comparison with the Hargreaves-Samani model (MAE = 0. 0.588, RMSE = 0.750 y R^2 = 0.703) confirming that the model operate in sub humid warm climate like that of the state of Campeche. About the soft computing models, the SVM model obtained a better performance (MAE = 0.480, RMSE = 0.637 y R^2 = 0.786) being the recommended model to estimate the ETo in the state of Campeche. The GEP model slightly surpassed the empirical models and provides a programmable algebraic model in a spreadsheet, to make predictions of ETo, being another viable option in the determination of ETo in the state of Campeche.

Keywords: Reference evapotranspiration; empirical equations; soft computing techniques; automated weather stations.

DEDICATORIA

A mi compañera de vida Guadalupe, quien gracias a su amor, paciencia, cariño y esfuerzo he podido llegar a cumplir hoy un sueño más.

A mis padres Eneida y Ángel, por inculcar en mí el ejemplo de esfuerzo y valentía, de no temer las adversidades y de saber salir adelante aun en las adversidades.

A mis hermanos Carlos, Ángel y Alex, que con consejos y palabras de aliento hicieron de mí una mejor persona y de una u otra forma me acompañan en todos mis sueños y metas.

A Tony…

AGRADECIMIENTOS

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo financiero que me brindó a través de una beca de manutención para poder concluir mi posgrado.

Al Dr. Víctor Hugo Quej Chi, por brindarme su confianza, extenderme una mano y compartir sus conocimientos, por prestarme siempre de su atención pero sobre todo por creer en mí y por ayudarme a culminar una meta más.

Al Dr. Everardo Aceves, por ser una excelente persona y por ser una inspiración para todas las personas que lo rodean.

Al Dr. Eugenio Carillo y al M.C. Benigno Rivera por el esfuerzo, la dedicación, el tiempo, el apoyo y la confianza que me han brindado.

A la M.C. Guadalupe Arena, por estar siempre conmigo, por levantarme de todas mis caídas, por brindarme siempre ese impulso de seguir adelante, por no dejarme declinar, por creer siempre en mí, por darme todo ese amor y cariño que toda persona necesita para seguir adelante.

A mis compañeros de generación, de las siguientes generaciones y generaciones pasadas, Doctores, Maestros, cuerpo administrativo, etc. integrantes del Colegio de Postgraduados Campus Campeche, quienes de alguna manera me han acompañado hasta el día de hoy.

CONTENIDO

RESUMEN iv

ABSTRACT v

LISTA DE FIGURAS xii

LISTA DE CUADROS xiv

CAPÍTULO I. INTRODUCCIÓN 1

CAPÍTULO II. REVISIÓN DE LITERATURA 4

2.1 Evapotranspiración (ET) 4

2.1.1. Conceptos básicos relacionados 4

2.1.2 Evapotranspiración del cultivo (ETc) 5

2.1.3 Evapotranspiración real 6

2.2 Evapotranspiración de referencia (ETo) 6

2.2.1 Fórmula de la FAO56 Penman-Monteith 7

2.2.2 Ecuaciones empíricas 8

2.2.2.1 Ecuación de Hargreaves y Samani (1985) 9

2.2.2.2 Ecuación de Camargo (1999) 10

2.2.2.3 Calibración de las fórmulas empíricas 10

2.2.3 Métodos de Inteligencia Artificial o Soft-computing 11

2.2.3.1 Conceptos básicos 11

2.2.3.2 Máquinas de Soporte Vectorial. 12

2.2.3.2.1 Funciones del Kernel 15

2.2.3.3 Programación de Expresión Genética. 16

2.2.3.4 XGBoost. 19

2.3 Factores climáticos que afectan la ETo. 21

2.3.1 Temperatura 21

2.3.2 Radiación solar 22
2.3.3 Humedad relativa 22
2.3.4 Velocidad del viento 23
2.4 Estaciones Meteorológicas 23
2.4.1 Estaciones Meteorológicas Automatizadas (EMAs) 23
2.4.2 Componentes de una EMAs 24
2.4.3 Tiempo Universal Coordinado (TUC) 26
2.5 Calidad de los datos meteorológicos 26
2.5.1 Técnicas de interpolación 26
2.5.1.1 Interpolación Polinómica de Hermite (PCHIP) 28
2.5.1.2 Interpolación "SPLINE" 29
2.5.2 Datos atípicos 30
2.5.2.1 Diagrama de caja y "bigotes"; cuartiles) 31
2.5.2.2 Método de "Mean" (método de la media) 31
CAPÍTULO III. PLANTEAMIENTO DEL PROBLEMA 33
CAPÍTULO IV. JUSTIFICACIÓN 34
CAPÍTULO V. HIPÓTESIS 35
CAPÍTULO VI. OBJETIVOS 35
6.1. Objetivo General 35
6.2. Objetivos específicos 35
CAPÍTULO VII. MATERIALES Y MÉTODOS 36
7.1 Descripción del sitio de estudio 36
7.1.1 Ubicación de las EMAs 37
7.2 Datos y análisis de calidad 37
7.2.1 Estado de las bases de datos meteorológicos 38

7.2.1.1 Adecuación de los datos meteorológicos 38
7.2.1.2 Ajuste del formato de fecha y de hora 38
7.2.1.3 Arreglo de series consecutivas 39
7.2.2 Manejo de datos faltantes 39
7.2.3 Detección de valores atípicos 39
7.3 Estimación de la ETo 40
7.3.1 Fórmula de la FAO-56 PM 40
7.3.2 Ecuaciones Empíricas 41
7.3.2.1 Calibración 41
7.3.3 Métodos de inteligencia artificial o "*Soft-computing*" 41
7.3.3.1 Máquinas de Soporte Vectorial (MSV) 41
7.3.3.2 Programación de Expresión Genética (GEP) 42
7.3.3.3 Máquina de aumento del Gradiente Extremo (XGBoost) 43
7.4 SOFTWARE 44
7.4.1 MATLAB 44
7.4.2 The R Project 44
7.4.3 GeneXproTools 5.0 45
7.5 Índices estadísticos para la evaluación de modelos de predicción de la ETo 46
CAPÍTULO VIII. RESULTADOS Y DISCUSIÓN 48
8.1 Análisis de la calidad de datos 48
8.1.1 Estado de la calidad de datos 48
8.1.1.1 Ajuste del formato de fecha y hora 50
8.1.1.2 Arreglo en las series consecutivas 51
8.1.2 Evaluación de métodos de relleno de datos meteorológicos 53

8.1.3 Detección de valores atípicos 58

8.2 Evaluación de los modelos empíricos y de soft-computing en la estimación de la ETo. 60

8.2.1 Evaluación de ecuaciones empíricas en la estimación de la ETo 62

8.2.2 Evaluación de los métodos de inteligencia artificial (soft-computing) en la estimación de la ETo. 65

CAPÍTULO IX. CONCLUSIONES 76

CAPÍTULO X. LITERATURA CITADA 78

CAPÍTULO XI. ANEXOS 83

LISTA DE FIGURAS

Figura 1. Función de pérdida del modelo SVM 14
Figura 2. Función de minimización 15
Figura 3. Árbol de expresión del modelo GEP 17
Figura 4. Árbol de expresión para la formulación de una función matemática 18
Figura 5. Diagrama de flujo del funcionamiento del modelo GEP 18
Figura 6. Esquema de funcionamiento del modelo XGBoost 21
Figura 7. Estructura tipo andamio 25
Figura 8. Estructura tipo Torre triangular 25
Figura 9. Representación gráfica de una interpolación de datos 27
Figura 10. Representación gráfica de interpolación por PCHIP 29
Figura 11. Representación gráfica de interpolación por Spline 30
Figura 12. Diagrama de caja y “bigotes” 31
Figura 13. Representación de la técnica Mean 32
Figura 14. Valores de entrada de la fórmula de la FAO-56 PM en la hoja de cálculo de Excel 40
Figura 15. Parámetros de la hoja de cálculo de la FAO-56 PM 41
Figura 16. Hora y fecha en columnas diferentes (febrero 2004). Estación Calakmul. .. 49
Figura 17. Hora y fecha en la misma columna (octubre 2009). Estación Calakmul. 49
Figura 18. Periodo corto sin registro de datos meteorológicos. 49
Figura 19. Periodo largo sin registro de datos meteorológicos. 50
Figura 20. Ausencia de periodos de registro en las bases de datos. 50
Figura 21. Ajuste en los formatos de hora y fecha mediante códigos macros implementados en Microsoft Excel. 51
Figura 22. Series consecutivas completadas (sin datos meteorológicos). 51
Figura 23. Series incompletas 52
Figura 24. Series completas. 53
Figura 25. Llenado de filas vacías mediante técnicas de interpolación. 54
Figura 26. Interpolación de datos en radiación solar. 55
Figura 27. Interpolación de datos en velocidad de viento 56
Figura 28. Interpolación de datos en humedad relativa 57

Figura 29. Interpolación de datos en temperatura 57
Figura 30. Detección de datos atípicos. 58
Figura 31. Representación gráfica de la detección de datos atípicos. A) Mean, B) Quartiles, C) Grubbs, en una misma serie de datos. 59
Figura 32. Adecuación de los valores de entrada en la fórmula de la FAO-56 PM. 60
Figura 33. Regresión lineal, valores estimados (Ecuaciones empíricas) vs valores observados (FAO 56 PM). Estación meteorológica A) Calakmul, B) Campeche, C) Cd del Carmen, D) Escárcega, E) Monclova, F) Los Petenes. 65
Figura 34. Gráficos de regresión lineal de valores estimados de ETo en etapa de entrenamiento y validación del modelo SVM vs valores calculados de ETo con la fórmula FAO56-PM, por estación meteorológica A) Calakmul, B) Campeche, C) Cd del Carmen, D) Escárcega, E) Monclova, F) Los Petenes. 69
Figura 35. Regresión lineal, entre los valores estimados de ETo en las etapas de entrenamiento y validación del modelo GEP vs valores calculados de ETo con la fórmula FAO56 por estación meteorológica: A) Calakmul, B) Campeche, C) Cd del Carmen, D) Escárcega, E) Monclova, F) Los Petenes. 71
Figura 36. Regresión lineal entre los valores estimados de ETo en las etapas de entrenamiento y validación con el modelo XGBoost vs valores calculados de ETo con la fórmula de la FAO56 por estación meteorológica A) Calakmul, B) Campeche, C) Cd del Carmen, D) Escárcega, E) Monclova, F) Los Petenes. 75

LISTA DE CUADROS

Cuadro 1. Ubicación geográfica en coordenadas de las estaciones meteorológicas automatizadas EMAs utilizadas en el presente trabajo. 37

Cuadro 2. Información histórica de las estaciones meteorológicas automatizadas (EMAs) 37

Cuadro 3. Parámetros del modelo GEP 43

Cuadro 4. Índices estadísticos (MAE, RMSE y R2) de los modelos empíricos y de soft-computing usados para la estimación de la ETo de cada estación meteorológica. 61

Cuadro 5. Índices estadísticos (R^2, MAE y RMSE) de los modelos SVM, GEP y XGBoost durante la etapa de entrenamiento y validación, de cada estación meteorológica. 66

Cuadro 6. Parámetros ajustados del modelo SVM por estación meteorológica utilizando el algoritmo genético. 67

Cuadro 7. Expresiones matemáticas obtenidas por el modelo GEP para cada estación. 72

CAPÍTULO I. INTRODUCCIÓN

La evapotranspiración (ET) es el proceso combinado entre la transpiración realizada por los cultivos y la evaporación del agua del suelo en el que se desarrollan, o de la superficie de las plantas que integran el cultivo, además de ser uno de los procesos principales en el ciclo del agua. Bajo este concepto, la evapotranspiración de referencia (ETo) se define como la tasa de evapotranspiración de una superficie de referencia que corresponde a un cultivo hipotético con características específicas, que ocurre sin restricciones de agua y representa la pérdida de agua de una superficie cultivada estándar (FAO, 2006). La importancia de la ETo radica en la aplicación de varios estudios en ciencias hidrológicas y agrícolas, como la gestión de los recursos hídricos y programación de riego, la cuantificación de las necesidades de agua de los cultivos y la planificación del uso de la tierra, por lo que su medición y/o estimación llega a ser un insumo importante en diferentes áreas de la gestión del agua (Allen *et al*., 1998; Sentelhas *et al*., 2010). La ET puede medirse de manera directa mediante el uso de lisímetros, evapotranspirómetros y torres de Eddy covarianza, donde la principal desventaja de estos métodos radican en los altos costos de instalación, mantenimiento de los equipos y uso de personal altamente capacitado para su operación, por lo que su uso muchas veces se limita a centros de investigación. La ETo, al ser afectada por variables climáticas, puede ser un parámetro climático que puede ser medido o estimado a partir de datos meteorológicos. Entre los métodos que se pueden implementar para la medición o estimación de la ETo se encuentran los indirectos que pueden estimar el requerimiento de agua a través de todo el ciclo vegetativo mediante la utilización de fórmulas la mayoría empíricas, entre los que se encuentran los modelos analíticos; como la ecuación Penman-Monteith (FAO, 2006) que desde el año de 1990 es utilizado como el método estándar para estimar la evapotranspiración del cultivo (ETc), que incluye parámetros tanto de intercambio de energía como de flujo de calor, (llamado también evapotranspiración) que pueden ser medidos o estimados utilizando datos meteorológicos y son implementados para la determinación de la ET de cualquier cultivo ya que la resistencia superficial y aerodinámica es distinta para cada tipo de cultivo. La

FAO (Food and Agriculture Organization) propuso una combinación del modelo Penman-Monteith con las ecuaciones de resistencia aerodinámicas. Este nuevo modelo es recomendado como método estándar para el cálculo de la evapotranspiración de referencia, y es denominado método FAO Penman-Monteith. La ETo representa la evapotranspiración de una superficie de referencia con el uso de un cultivo hipotético (pasto), con una altura, resistencia aerodinámica y albedo determinado, que crece de manera activa y tiene un adecuado riego (Valiantzas, 2013; FAO, 2006). Sin embargo, para su implementación y obtención de resultados precisos, se requiere de datos de variables climáticas de temperatura, radiación solar, velocidad del viento y humedad relativa, datos que son obtenidos de estaciones meteorológicas automatizadas (EMAs) (Hernández-Cruz, 2013), pero que en muchas ocasiones, algunas regiones no cuentan con este estaciones y solo cuentan con estaciones meteorológicas estándar que miden pocas variables climáticas. Aunado a esto, diferentes investigadores han propuesto muchos modelos empíricos que estimen con precisión la ETo. Estos modelos están basados en parámetros meteorológicos como temperatura, radiación solar, humedad relativa, etc., y para su implementación de forma confiable y precisa es necesario calibrarlos utilizando la ecuación de la FAO 56 Penman-Monteith (FAO56 PM) mencionado anteriormente como método estándar para la medición de la ETo.

Por otra parte, recientes investigaciones realizadas para la determinación de la ETo han establecido que puede ser modelada mediante técnicas denominadas de inteligencia artificial o “soft-computing”, basadas en el aprendizaje estadístico y auto aprendizaje, útiles para resolver problemas de clasificación, regresión y reconocimiento de patrones. Entre los modelos basados en “soft-computing” el modelo “Gene Expression Programming” (GEP) propone un enfoque alternativo que genera algoritmos y/o expresiones para resolver problemas de forma automática, y que recientemente se han aplicado con buenos resultados en estudios hidrológicos (Mattar, 2018). Entre los trabajos que han utilizado este modelo para la determinación de la evapotranspiración de referencia, se encuentra el trabajo realizado por Mehdizadeh *et al.,* (2017) quienes realizaron estimaciones de evapotranspiración de referencia utilizando los modelos de programación de expresión génica (GEP), máquinas de soporte vectorial y 16 ecuaciones empíricas. Los resultados obtenidos en ese estudio mostraron que el modelo

GEP presenta buenas funciones, pero no los mejores resultados, ya que los otros modelos de soft-computing utilizados en este estudio mostraron mejor resultado. En otro estudio, (Mattar, 2018) evaluó la estimación de la ETo mediante el modelo GEP utilizando una cantidad mínima de datos climáticos y comparándolos con modelos empíricos; los resultados mostraron que el modelo GEP es más precisa y por lo tanto se puede emplear en la estimación de la ETo.

Por otra parte, el modelo de “Support Vector Machines” (SVM), es una técnica de aprendizaje supervisado robusta, utilizada para resolver problemas de clasificación y de regresión aplicados a grandes conjuntos de datos complejos con ruido. Dado un conjunto de datos de entrenamiento (variables), la flexibilidad de SVM es asociada con las funciones del Kernel, que seleccionan los datos y los cambian a un hiperplano único de separación que equidiste de los ejemplos más cercanas de cada clase, para, de esta forma, conseguir un margen máximo de separación entre cada clase. Durante el entrenamiento solo se consideran los ejemplos que se encuentran en el margen de separación. Estos últimos se llaman vectores de soporte (Mehdizadeh, 2018).

En años recientes, se ha propuesto un nuevo algoritmo denominado XGBoost (Extreme Gradient Boosting), resultado de una versión mejorada del aumento de gradiente (Gradient Boosting), con una mayor eficiencia de cálculo y capacidad para resolver problemas de ajustes excesivos. La implementación de este modelo solo se ha dado en trabajos como el de Fan *et al*., 2018, donde se hace una comparación del modelo XGBoost con el modelo SVM, evaluando la precisión en la estimación de la radiación solar global diaria utilizando datos meteorológicos limitados. Los resultados mostraron que el modelo XGBoost obtuvo el mejor rendimiento durante la fase de entrenamiento, pero durante la fase de prueba su rendimiento fue ligeramente más bajo. No obstante, los autores consideran que la precisión en la predicción, la estabilidad del modelo y la eficiencia computacional son términos que hacen recomendable la utilización del modelo XGBoost en la estimación de la radiación solar.

CAPÍTULO II. REVISIÓN DE LITERATURA

2.1 Evapotranspiración (ET)

2.1.1. Conceptos básicos relacionados

La evapotranspiración (ET) es un proceso combinado por medio del cual se pierde agua de la superficie del suelo a la atmósfera mediante el proceso de evaporación y por la pérdida de agua en las plantas por el proceso de transpiración que se presenta en cada etapa de desarrollo. Además de ser uno de los componentes primordiales del ciclo hidrológico, es un factor esencial para la gestión de recursos hídricos y juega un papel importante en diferentes campos de estudio, principalmente en la parte agrícola (Fan *et al*., 2018; Straatmann *et al*., 2018). Tanto la evaporación como la transpiración, son procesos que se dan de manera simultánea, por lo que diferenciar cada proceso es algo difícil de realizar.

La evaporación (E) es un proceso físico esencial en el ciclo hidrológico, donde el agua líquida regresa a la atmósfera en forma de vapor de agua; proceso llamado vaporización. El agua se puede evaporar de cualquier superficie y es la única forma de transferencia de humedad de la tierra y los océanos a la atmósfera. La radiación solar y la temperatura tienen un papel importante en el proceso de evaporación, porque brindan la energía necesaria para que se lleve a cabo dicho proceso (Cascone *et al*., 2019) . La evaporación dada en una superficie de suelo cultivada es determinada por la fracción de radiación solar que llega a dicha superficie, donde esta fracción de radiación solar depende de la etapa de desarrollo del cultivo, esto se debe principalmente a que el dosel del cultivo proyectará más sombra y abarcará una mayor fracción de la superficie evaporante. En las primeras etapas de desarrollo de la planta, el cultivo no tiene el suficiente tamaño para abarcar mucha superficie evaporante, por lo que el agua se pierde por el proceso de evaporación de la superficie del suelo. Conforme el cultivo crece, el espacio a ocupar en el suelo será mayor, y en ocasiones cubre en su totalidad el suelo; de esta manera, el proceso de transpiración se convierte en el principal causante en la pérdida de agua (Macías, 2009).

La transpiración (T) es el proceso físico-biológico por el cual gran cantidad de agua absorbida del suelo por las raíces es transportada hasta la parte superior de las plantas donde se convierte en vapor de agua y se libera a través de los estomas, que son pequeñas aperturas en las hojas que regulan el intercambio del vapor hacia la atmósfera (Cascone *et al*., 2019).

Estos procesos dependen en gran medida de la energía obtenida de la radiación solar que llega a la superficie, pero también dependen del gradiente de presión de vapor y de la velocidad del viento, por lo que para poder determinar la tasa de transpiración de un cultivo es necesario conocer la radiación solar, la humedad atmosférica, la temperatura del aire y la velocidad del viento (FAO, 2006; Macias, 2009)

La ET de un cultivo es un indicador de su crecimiento, debido a que su rendimiento está íntimamente relacionado con la cantidad de agua evapotranspirada. Por lo tanto, cuantificar de manera precisa la evapotranspiración (ET) es de gran ayuda al momento de realizar programaciones de riego que coadyuven en la gestión de la disponibilidad del agua. (Martel *et al*., 2018; Mehdizadeh, 2018; Mehdizadeh *et al*., 2017; FAO, 2006).

2.1.2 Evapotranspiración del cultivo (ETc)

La evapotranspiración del cultivo puede ser considerada como la necesidad de agua de cualquier cultivo cuando este se encuentra libre de enfermedades, con buena fertilización y su desarrollo está dado en bajo óptimas condiciones de suelo y agua, para alcanzar la máxima producción de acuerdo a las condiciones climáticas de la región (Almorox, 2016; Webb, 2010; FAO, 2006; Macias, 2009).

La evapotranspiración de los cultivos se puede estimar haciendo mediciones de forma directa; utilizando los lisímetros, torres de flujos de calor sensible y latente, sistemas de cámaras, método del flujo de sabia, entre otros; o de forma indirecta, mediante la aplicación del concepto de evapotranspiración de referencia (ETo), el cual, al combinarlo con los coeficientes de cultivo Kc permite realizar una estimación de la

evapotranspiración potencial ETc para un cultivo especifico (Alberto *et al*., 2014; Mahmoud y Gan, 2019; Negm *et al*., 2017).

2.1.3 Evapotranspiración real

El valor de la evapotranspiración potencial ETc se refiere a la que se presenta en un cultivo creciendo bajo condiciones óptimas de humedad en el suelo y expresando su máximo potencial productivo, generalmente en campos con buen manejo agronómico. Pero en algunos casos, las condiciones del campo en las que se encuentra un cultivo difieren de las de uno que cuenta con un manejo deficitario, por ejemplo, bajo condiciones de baja fertilidad, salinidad, suelos inundados o en su defecto con falta de agua, o la incidencia de plagas o enfermedades. La presencia de horizontes duros o impenetrables en la zona radicular puede generar en la planta un crecimiento deficiente. Todo lo anterior resulta en la reducción de la tasa de evapotranspiración por debajo de la ETc. Para estimar la ET real (ETr), es común multiplicar la ETc por coeficientes de estrés hídrico menores de la unidad (Kr) para cada cultivo, ajustando los valores del coeficiente de cultivo (Kc) a las condiciones ambientales presentes (FAO, 2006).

2.2 Evapotranspiración de referencia (ETo)

La ETo es un parámetro agrometeorológico de gran importancia para muchas áreas de estudio como geotécnica, climatológica e hidrológica, donde su mayor importancia recae en el cálculo de la ETc (Čadro *et al*., 2017; Jovic *et al.*, 2018; Webb, 2010; Zhang *et al*., 2018). Es definida como la tasa de evapotranspiración de una superficie de referencia hipotética que presenta características específicas, y se emplea principalmente para estimar la ETc de un cultivo en específico. La FAO en el manual 56 de riego y drenaje define a la ETo como un “parámetro relacionado con el clima que expresa el poder evaporante de la atmósfera”, sin embargo, el concepto de la ETo se introdujo para estimar la demanda ambiental de evapotranspiración de la atmosfera, esto sin tener en cuenta el tipo de cultivo, su desarrollo y las prácticas de manejo dadas a dicho cultivo (FAO,2006; Mattar, 2018). Algunos otros autores la definen como “la tasa de

evapotranspiración de un cultivo", "tasa de evapotranspiración de una superficie de referencia, que ocurre sin restricciones de agua" donde la superficie de referencia corresponde de un cultivo hipotético (pasto) que presenta características específicas de altura (0.12 m), resistencia de superficie de cultivo de (70 s m^{-1}) y albedo (0.23) fijos, similar a la evapotranspiración de una superficie extensa cubierta de pasto verde con una altura uniforme, creciendo de manera activa, sombreando completamente el suelo y con disponibilidad de agua adecuada, cuyo principal objetivo es determinar la demanda de agua perdida por la evapotranspiración, dependiendo del tipo, etapa de desarrollo y prácticas de manejo del cultivo, asociado con las condiciones climáticas del lugar (Gong *et al.*, 2016; Jovic *et al.*, 2018; Zhang *et al.*, 2018).

2.2.1 Fórmula de la FAO56 Penman-Monteith

La fórmula de la FAO 56 PM es el modelo estándar para estimar la ETo, y fue propuesto por la Organización de las Naciones Unidas para la Agricultura y la Alimentación (FAO por sus siglas en inglés) en el Documento de Irrigación y Drenaje No. 56. Para implementar esta fórmula se requieren variables meteorológicas que hacen la función de valores de entrada (FAO, 2006; Fan *et al*., 2018), considera muchos parámetros meteorológicos relacionados con el proceso de evapotranspiración, tales como la radiación neta, la temperatura del aire, el déficit de presión de vapor, la velocidad del viento y ha presentado muy buenos resultados en comparación con los resultados obtenidos con la aplicación de otros métodos como los lisímetros (Leszek *et al*., 2011). Sin embargo, existe el inconveniente de que no en todas las estaciones meteorológicas se registran todas las variables necesarias para la implementación de este modelo, y en muchas ocasiones se opta por la utilización de modelos que ocupen menor cantidad de variables (Fan *et al*., 2018; Shiri, 2017)

La ecuación de la FAO56-PM, incorpora aspectos termodinámicos y aerodinámicos, ha demostrado ser un método relativamente preciso bajo diferentes condiciones o regiones. Estos aspectos han sido incorporados en la siguiente ecuación:

$$ETo = \frac{0.408\Delta(R_n - G) + \gamma \frac{900}{T_{media}+273} u_2 (e_s - e_a)}{\Delta + \gamma(1 + 0.34 u_2)} \quad (1)$$

ETo	evapotranspiración de referencia (mm día^{-1})
R_n	radiación neta en la superficie del cultivo (MJ m^{-2} día^{-1})
G	flujo del calor de suelo (MJ m^{-2} día^{-1})
T_{media}	temperatura media del aire a 2 m de altura (°C)
u2	velocidad del viento a 2 m de altura (m s^{-1})
e_s	presión de vapor de saturación (kPa)
e_a	presión real de vapor (kPa)
e_s - e_a	déficit de presión de vapor (kPa)
Δ	pendiente de la curva de presión de vapor (kPa °C^{-1})
γ	constante psicrométrica (kPa °C^{-1})

2.2.2 Ecuaciones empíricas

Utilizar la fórmula de la FAO 56-PM en sitios donde no se cuenta con datos de algunas de las variables meteorológicas requeridas para su aplicación es poco viable; muchos investigadores han desarrollado e implementado otros métodos alternos para estimar la ETo, que utilizan menos datos climáticos y pueden clasificarse en diferentes tipos dependiendo de la disponibilidad de variables meteorológicas. Dentro de estos métodos alternos se encuentran las ecuaciones empíricas, que son métodos basados en el uso de menos variables meteorológicas, por ejemplo, empleando únicamente datos de radiación solar, o de temperatura, etc. Dentro de las ecuaciones empíricas podemos mencionar el modelo Thornthwaite, basado en datos de temperatura para calcular la evapotranspiración potencial; el modelo Hargreaves y Samani que utiliza datos de temperatura máxima (T_{max}) y mínima (T_{min}) en relación con la radiación solar extraterrestre (Ho) para calcular la radiación solar. Estos mismos autores propusieron el modelo Hargreaves basado solo en datos de temperatura, el cual fue considerado como uno de los métodos más fácil de implementar. Howard Penman (1948) propuso un

modelo para la estimación de la evaporación a partir de superficies abiertas mediante la combinación del balance de energía con los métodos de transferencia de masa. Priestley y Taylor propusieron (1972) el modelo basado en radiación, que era una simplificación del modelo de Penman (Feng *et al.*, 2016)

2.2.2.1 Ecuación de Hargreaves y Samani (1985)

El modelo de Hargreaves y Samani ha sido considerado como un modelo alternativo para estimar la ETo cuando solo los registros de temperatura están disponibles en el lugar de estudio. Es uno de los métodos que ha sido utilizado consecutivamente por su simple implementación y la precisión en sus resultados (Gong et al., 2016; Shiri, 2017). Sin embargo, Shiri et al., (2015) menciona que muchos estudios realizados han demostrado que este modelo sobrestima la ETo en regiones húmedas y subestima la ETo en regiones secas.

La ecuación del modelo de Hargreaves y Samani está estructurada de la siguiente manera:

$$ETo = 0.408\, K_{HS}\, (T_{media} + 17.8)(T_{max} - T_{min})^{0.5}\, Ho \qquad (2)$$

Dónde:

ETo = Evapotranspiración de referencia (mm día^{-1}); K_{HS} = Es un coeficiente empírico, que inicialmente fue establecido en 0.0023 pero se ha recalibrado acorde al lugar empleado; T_{media} = Temperatura media (ºC); T_{max} = Temperatura máxima (ºC); T_{min} = Temperatura mínima (ºC); *Ho* = Radiación solar extraterrestre (MJ m^{-2} día^{-1}); 17.8 y 0.5 son constantes (Almorox, 2016; Ayyoub *et al*., 2017).

2.2.2.2 Ecuación de Camargo (1999)

La ecuación de Camargo es una modificación de la ecuación de Thornthwaite, y es un modelo basado en la variable climática de temperatura; Camargo sustituyó el valor de la temperatura media de la ecuación de Thornthwaite por la temperatura efectiva, en función de la amplitud térmica diaria que pudiera ser utilizada en el modelo de Thornthwaite y similares, buscando corregir la estimación de la evapotranspiración potencial (ETp) en condiciones especiales de aridez y de mucha humedad, en las cuales la ecuación de Thornthwaite carece de exactitud, en la medida en que subestima la evapotranspiración potencial, en el primer caso, y la sobreestima, en el segundo (Camargo *et al.*, 1999).

$$ETo = K_{CA1} * \left(10 * \frac{K_{CA2} * (3T_{max} - T_{min})}{I}\right)^{a} * N/360 \qquad (3)$$

Donde:

ETo = Evapotranspiración de referencia (mm día^{-1}); K_{CA1} y K_{CA2} = Coeficientes empíricos, cuyos valores originales son 16 y 0.36 respectivamente y deben ser calibrados acorde al lugar de empleo; I = índice de calor anual; N = Tiempo de insolación en horas; ($3T_{max}$ - T_{min}) = Temperatura efectiva, remplazando a la temperatura media en la ecuación de Thornthwaite; 10 = Constante (Čadro *et al*., 2017; Moeletsi *et al*., 2013).

2.2.2.3 Calibración de las fórmulas empíricas

Las ecuaciones empíricas han sido desarrolladas en regiones donde las condiciones climáticas son diferentes del resto del mundo, por lo que implementarlas sin previo ajuste de sus componentes conduce a la subestimación o sobrestimación de los valores de evapotranspiración, dependiendo de la región donde se implemente.

La ecuación de Hargreaves y Samani (HS) fue desarrollada bajo las condiciones climáticas semiáridas presentes en California, EE.UU. y el uso de esta ecuación en

regiones húmedas, tiende a sobreestimar los valores de la ETo. Caso contrario sucede cuando se usa en regiones aún más secas, los valores de la ETo tienden a ser subestimados (Shiri *et al*., 2015; Zanetti *et al*., 2019). Por ello la calibración de las ecuaciones empíricas se realiza específicamente para el sitio en estudio y no se pueden implementar en otros lugares con condiciones de climas diferentes, motivo por el cual, la ecuación de HS se debe de calibrar o ajustar de manera local para obtener un rendimiento óptimo (Gong *et al*., 2016).

2.2.3 Métodos de Inteligencia Artificial o Soft-computing

2.2.3.1 Conceptos básicos

Las técnicas de soft-computing en los últimos diez años han ganado popularidad en muchas disciplinas, incluyendo estudios hidrológicos como es la estimación de la ETo, donde han mostrado buenos resultados (He *et al*., 2014).

El soft-computing es una colección de metodologías que apuntan a explotar la tolerancia a la imprecisión y la incertidumbre para lograr trazabilidad, robustez y bajo costo de solución. A veces se denomina inteligencia computacional, que abarca una gama de técnicas computacionales en informática, inteligencia artificial y aprendizaje automático. Algunas veces, el término “soft-computing” se usa indistintamente con el de sensores blandos o sensores virtuales. El sensor suave es un nombre común para una pieza de software que se utiliza para derivar información deseable de las mediciones disponibles. Los sensores blandos son especialmente útiles en la fusión de datos, donde las mediciones de diferentes características y dinámicas se fusionan y combinan. Los algoritmos de software conocidos que se utilizan en sensores blandos incluyen filtros de Kalman. Recientemente, las redes neuronales o la computación difusa se han utilizado para implementar sensores blandos. Hasta cierto punto, el software desarrollado basado en técnicas de soft-computing, con el propósito de monitoreo o medición, puede considerarse como un sensor blando (Yan *et al*., 2018)

El cálculo de ETo puede considerarse como un proceso de regresión no lineal complicado, en función de varios factores climáticos, por lo que cada vez más investigadores han presentado modelos para estimar la ETo basados en técnicas de soft-computing (Feng *et al*., 2016)

En los últimos años, además de las ecuaciones empíricas, los enfoques de soft-computing se han utilizado ampliamente para estimar parámetros ambientales, hidrológicos y climatológicos (por ejemplo, ETo). Los enfoques de soft-computing son válidos como métodos para modelar procesos complejos no lineales (Mehdizadeh *et al*., 2017).

2.2.3.2 Máquinas de Soporte Vectorial.

La teoría de las máquinas de soporte vectorial (SVM, por sus siglas en inglés, Support Vector Machines) fue desarrollada por (Vapnik, 2000), y es uno de los enfoques basados en el aprendizaje automático. Es una técnica de aprendizaje supervisado robusta para resolver problemas de clasificación y regresión aplicados a grandes conjuntos de datos complejos con ruido. Dado un conjunto de datos de entrenamiento (variables), la técnica de SVM selecciona un hiperplano único de separación que equidiste de los ejemplos más cercanos de cada clase, para, de esta forma, conseguir un margen máximo de separación entre cada clase. La idea básica es mapear los datos *x* en un espacio de características de alta dimensión a través de un mapeo no lineal y hacer una regresión lineal en este espacio. Durante el entrenamiento solo se consideran los ejemplos que se encuentran en el margen de separación. Estos últimos se llaman vectores de soporte. Es un modelo de inteligencia artificial supervisada para el análisis de datos y el reconocimiento de patrones, que se utiliza ampliamente para la regresión y la predicción. El modelo SVM estima la regresión en función de una serie de funciones Kernel, que convierten los datos de entrada originales de dimensiones inferiores a un espacio de características de mayor dimensión de una manera implícita. En comparación con el modelo ANN (Red Neuronal Artificial) normalmente con múltiples mínimos locales, la SVM ofrece una solución única como resultado de la naturaleza convexa del problema de optimización. Los modelos SVM se aplican con éxito en problemas de regresión,

generalmente denominados SVR (regresión de vectores de soporte), utilizando SVM para un conjunto de datos {(X_i, Y_i)} N/i = 1, donde X_i es el vector de entrada, Y_i es el valor de salida y *N* es el número total de conjuntos de datos mediante el mapeo de *X* en un espacio característico a través de una función no lineal φ(x) para después encontrar una función de regresión (Ecuación 1) (Fan *et al*., 2018; Mehdizadeh *et al.*, 2017; Quej *et al.*, 2017; Topi y Vanita, 2017; Wen *et al.*, 2015).

$$f(x) = \omega\varphi(x) + b \quad (4)$$

Donde $\varphi(x)$ es la función de mapeo no lineal; ω es un vector de peso y *b* es un valor de sesgo. ω y *b* son los parámetros de la función de regresión, los cuales pueden ser calculados minimizando la siguiente función de riesgo regularizado:

$$R(C) = C \sum_i^N L_\varepsilon (f(\chi_i), y_i) \frac{1}{2} \|\omega\|^2 \quad (5)$$

Donde el término $\frac{1}{2}\|\omega\|^2$ mejora la generalización del modelo SVM, normalizando el grado de complejidad del modelo; C es un parámetro de compensación positiva que determina el grado de error en el problema de optimización elegido por el usuario y (ε) es la función de pérdida de Vapnik (tamaño del tubo del modelo de SVM) y está definida como:

$$L_\varepsilon (f(\chi_i), y_i) = \begin{cases} 0 & for |f(x_i) - y_i| \leq \varepsilon \\ |f(x_i) - y_i| - \varepsilon & otra\ manera \end{cases} \quad (6)$$

Es decir, si la diferencia entre los valores predichos y los medidos es menor que ε, entonces la pérdida es igual a 0. Si los valores predichos están dentro del tubo, el error de pérdida es igual a 0. Para el resto de los puntos predichos encontrados fuera del tubo, la pérdida es igual a la diferencia entre el valor predicho y el radio ε del tubo. Para evitar los valores atípicos, se introducen las variables de holgura ξ y ξ^*, que miden de arriba y abajo en el tubo de ε, ambas variables de holgura tienen valor positivo.

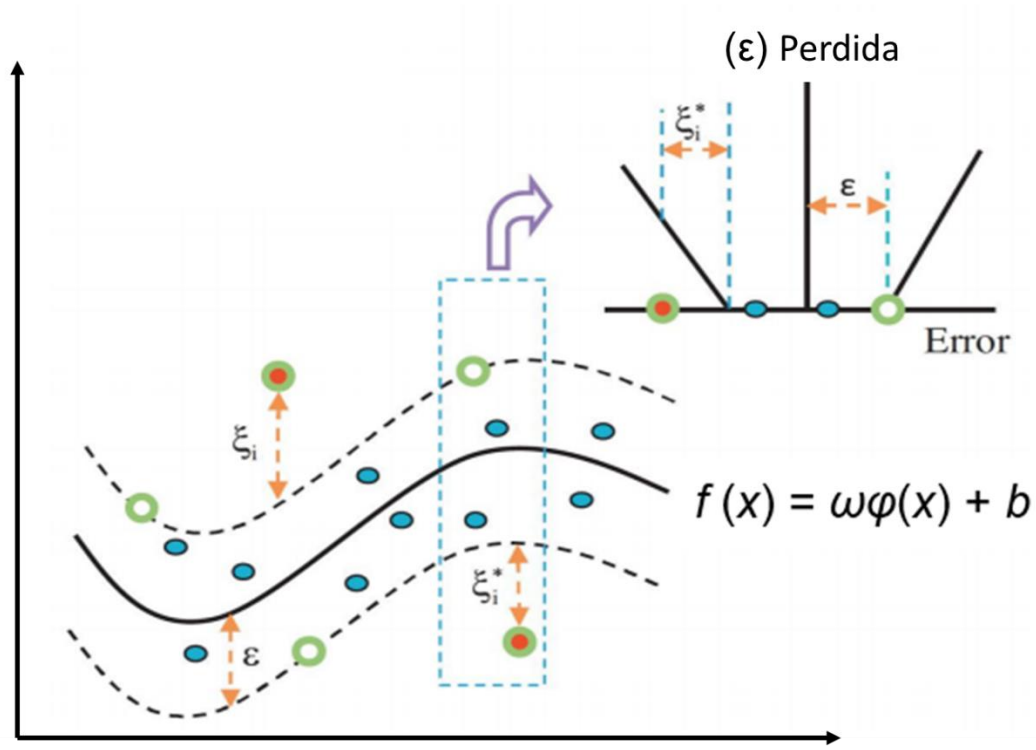


Figura 1. Función de pérdida del modelo SVM

Debido a que ambas variables adquieren valores positivos, se tiene que minimizar el riesgo con la siguiente ecuación:

$$R\ (\xi, \xi^*, \omega, b) = \frac{1}{2}\ \|\omega\|^2 + C\ \sum_{i=1}^{n}(\xi_i\ +\ \xi_i^*) \qquad (7)$$

Sujeto a $\begin{cases} y_i\ -\ \omega\phi(x_i)\ -\ b_i\ \leq\ \varepsilon\ +\ \xi_i \\ \omega\phi(x_i) + b_i\ -\ y_i\ \leq \varepsilon\ +\ \xi_i^* \\ \xi,\quad \xi_i^* \geq 0 \end{cases}$

Donde la $C\sum_{i=1}^{n}(\xi_i\ +\ \xi_i^*)$ controlan los grados de riesgo empírico.

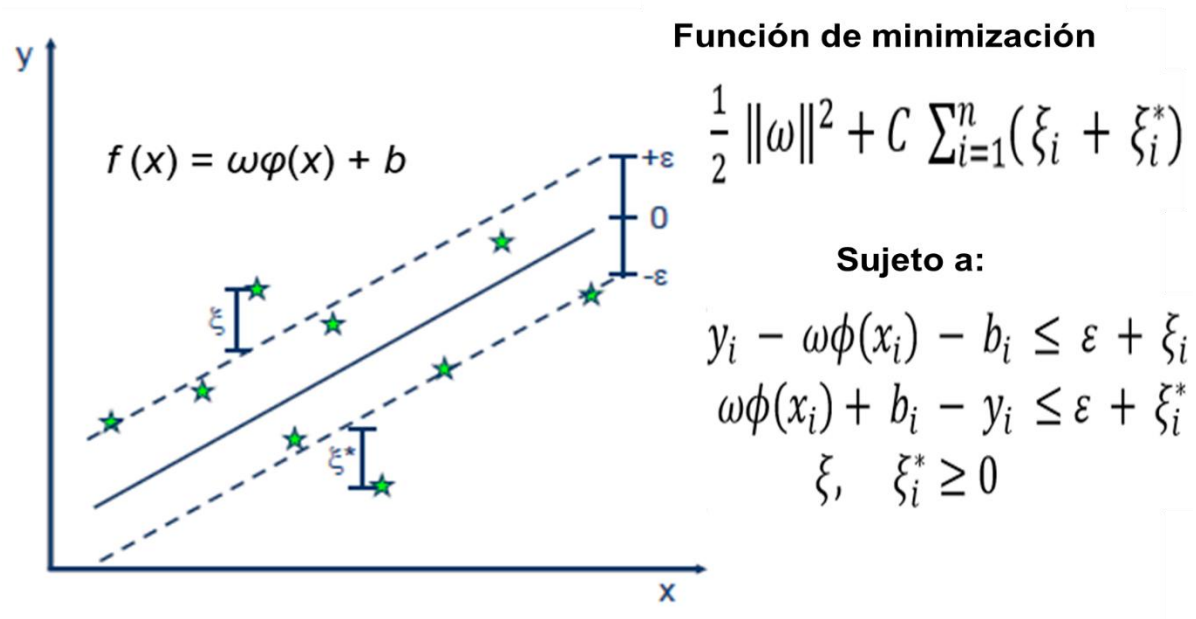


Figura 2. Función de minimización

2.2.3.2.1 Funciones del Kernel

Las funciones Kernel cambian los datos a un espacio de características dimensionales más alto. Entre los Kernel más utilizados se halla el SVM polinomial (SVM-Poly) y la función de base radial SVM (SVM-RBF), cuyos parámetros del Kernel deberán ajustarse previamente mediante un algoritmo. Por ejemplo, los parámetros óptimos del Kernel y del modelo SVM generalmente se obtienen utilizando el método de búsqueda de cuadrícula (Mehdizadeh et al., 2017). Las funciones Kernel también pueden ser del tipo lineal o sigmoidal, y su elección dependerá del problema que se requiera resolver. Cada función Kernel presenta su respectiva ecuación:

Función Kernel de Base Radial:

$$K\left(x_i, x_j\right) = \exp\left(-\gamma \left\|x_i - x_j\right\|^2\right), \qquad \gamma > 0 \qquad (8)$$

Función Kernel Polinomial:

$$K\left(x_i, x_j\right) = \left(\gamma x_i . x_j + r\right)^{\mathrm{d}}, \qquad \gamma > 0 \qquad (9)$$

Función Kernel Lineal:

$$K\left(x_i, x_j\right) = x_i . x_j \quad (10)$$

Función Kernel Sigmoidal:

$$K\left(x_i, x_j\right) = \tanh\left(\gamma x_i . x_j + r\right) \quad (11)$$

2.2.3.3 Programación de Expresión Genética.

La programación de la expresión genética (GEP) fue presentada por (Ferreira, 2001). Es una rama de los algoritmos evolutivos que tiene la capacidad de modelar los procesos dinámicos y no lineales. Es un algoritmo que pertenece a la familia de los algoritmos genéticos (GA) y programación genética (GP) tradicionales. Puede emular la evolución biológica basada la programación por computadora para resolver un problema definido por el usuario.

Los GEP se consideran un hibrido entre los GA y GP. Utilizan programación genética para la solución del problema en forma de árbol, donde existen dos tipos de nodos:

- Terminales, u hojas del árbol. No tienen descendientes, se asocian a las variables o constantes.
- Funciones. Tienen descendientes, se asocian a operadores del algoritmo que se desea desarrollar.

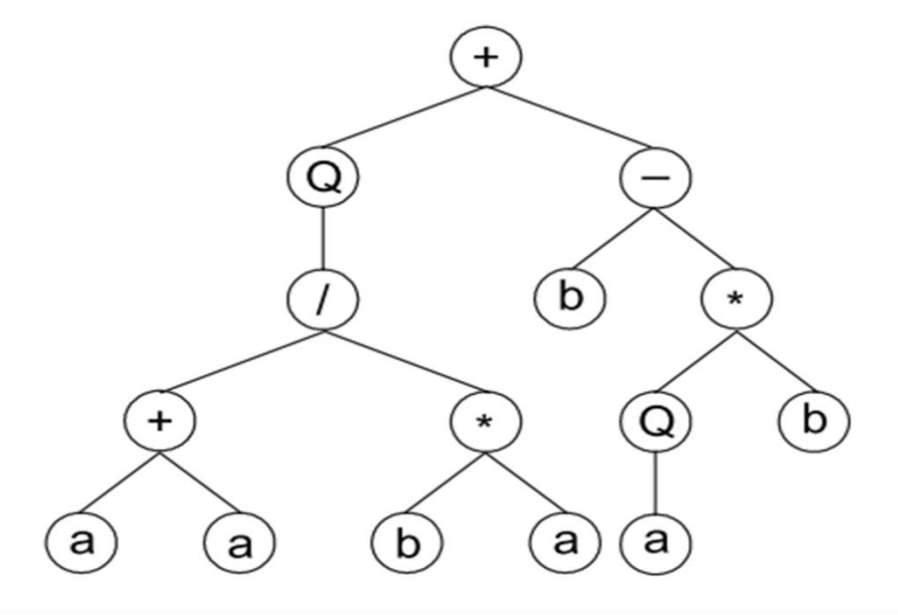


Figura 3. Árbol de expresión del modelo GEP

Las poblaciones de individuos se utilizan en los algoritmos evolutivos. La principal diferencia entre tres algoritmos es la naturaleza de los individuos. En GEP, los individuos se codifican primero como cadenas lineales de longitud fija como el GA. Luego, se expresan como entidades no lineales de diferentes tamaños y formas, como el GP (Ferreira, 2001). Además, un conjunto de terminales (coeficientes y predictores), funciones y operadores matemáticos se utilizan en el GEP para estimar la variable dependiente (Mehdizadeh *et al*., 2017) creando de manera aleatoria las funciones creadas y seleccionadas que se ajuste mejor a los resultados experimentales, permitiendo la generación de algoritmos y expresiones matemáticas de manera automática para la solución de problemas, donde estos algoritmos son implementados para la generación de una función matemática que se adecue a un determinado conjunto de datos (Mattar, 2018; Shiri, 2017).

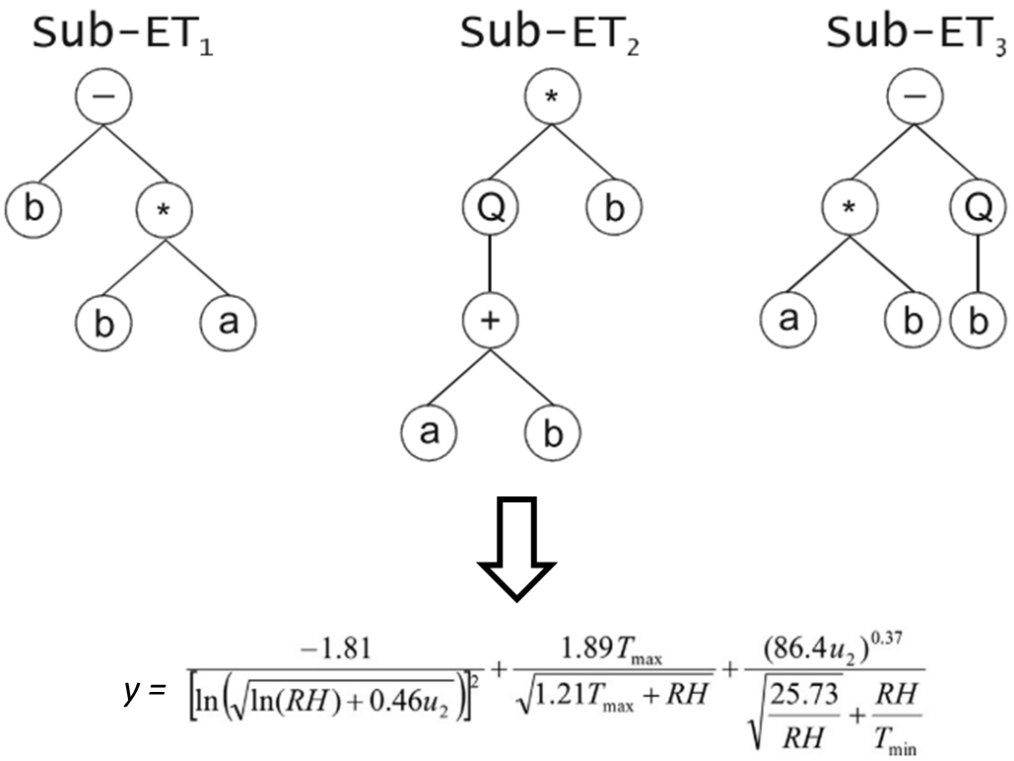


Figura 4. Árbol de expresión para la formulación de una función matemática

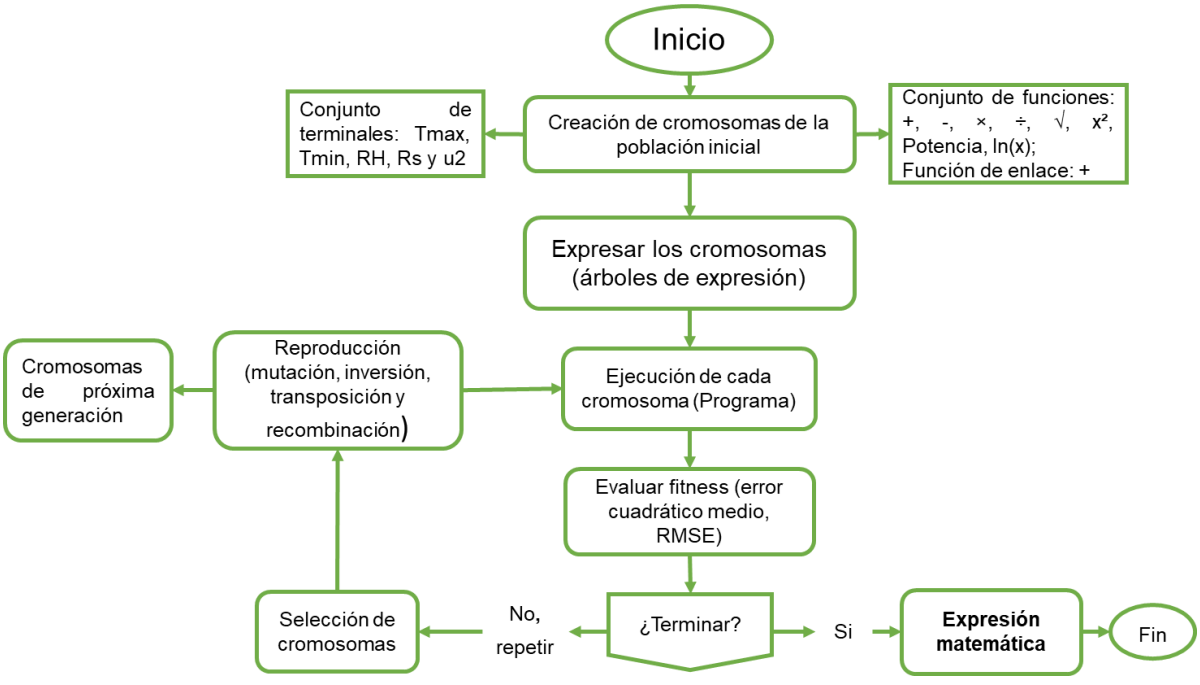


Figura 5. Diagrama de flujo del funcionamiento del modelo GEP

2.2.3.4 XGBoost.

Es uno de los algoritmos más importantes y potentes del “Machine Learning” (aprendizaje automático) creado por (Chen y Guestrin, 2016), utilizado para el análisis de problemas de regresión y clasificación estadística, el cual produce un modelo de predicción complejo a partir del ensamblaje de árboles de decisión (modelos simples), en un contexto de aprendizaje supervisado.

El término “Extreme” se refiere al objetivo de llevar al límite los recursos computacionales para obtener mejores resultados.

El modelo se basa en la teoría del aumento, por lo que las predicciones de varios aprendices "débiles" (modelos cuyas predicciones son ligeramente mejores que las suposiciones aleatorias), se combinan para desarrollar un aprendiz "fuerte". Estos aprendices "débiles" se combinan siguiendo una estrategia de aprendizaje gradual. Al comienzo del proceso de calibración, un aprendiz "débil" se ajusta a todo el espacio de datos, y luego, un segundo aprendiz se ajusta a los residuos de la primera. Este proceso de ajuste de un modelo a los residuos del anterior continúa hasta que se alcanza algún criterio de detención. El resultado es un tipo de media ponderada de las predicciones individuales de cada alumno débil. Tradicionalmente, los árboles de regresión se seleccionan como aprendices "débiles". Bajo este contexto el modelo XGBoost se basa en la siguiente función objetivo: *pérdida* + *regularizador:*

$$Obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{i=1}^{t} \Omega\ (f_i) \qquad (12)$$

donde l es el término predictivo y Ω el término de regularización. La función de pérdida para el término predictivo puede ser especificada por el usuario. El término de regularización se obtiene con una expresión analítica basada en el número de hojas del árbol y las puntuaciones de cada hoja. El punto clave del proceso de calibración de XGBoost es que ambos términos se reordenan en última instancia en la siguiente expresión:

$$Obj^{(t)} = -\frac{1}{2}\sum_{i=1}^{T}\frac{G_j^2}{H_j+\lambda} + \gamma \mathrm{T} \quad (13)$$

donde *G* y *H* se obtienen de la expansión de las series de Taylor de la función de pérdida, λ es el parámetro de regularización L2 y *T*, el número de hojas. Esta expresión analítica de la función objetivo permite un rápido escaneo de izquierda a derecha de las posibles divisiones del árbol, pero siempre teniendo en cuenta la complejidad.

XGBoost tiene una amplia gama de parámetros de ajuste. Además, la flexibilidad del algoritmo se mejora al dar la oportunidad al usuario de incluir algunos parámetros autodefinidos, como la función de pérdida o la métrica utilizada para la validación y prueba (Urraca *et al*., 2017).

Entre algunos de los ejemplos de problemas que se han solucionado y en los que se han encontrado buenos resultados se puede mencionar a la predicción de ventas en una tienda; la clasificación de eventos de física de alta energía; la clasificación de texto web; la predicción del comportamiento del cliente; la detección de movimiento; la clasificación de malware; la categorización de productos, etc. el hecho de que XGBoost sea la elección de consenso de muchos investigadores, muestra el impacto y la importancia de este sistema y la potenciación de los árboles (Chen y Guestrin, 2016)

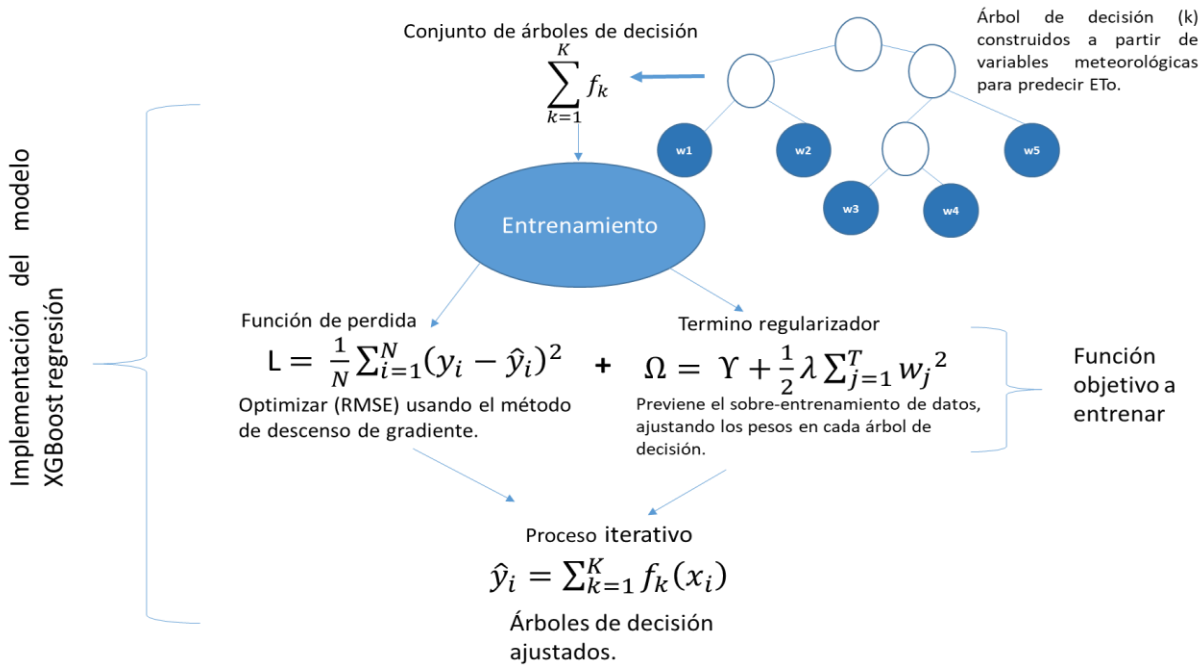


Figura 6. Esquema de funcionamiento del modelo XGBoost

2.3 Factores climáticos que afectan la ETo.

La ETo es afectada únicamente por factores climáticos, por lo que su determinación se puede llevar a cabo tomando en cuenta datos meteorológicos. Estos datos son obtenidos de las variables climáticas y nos proporcionan la energía necesaria para la transpiración del cultivo y la evaporización del agua de una superficie evaporante.

2.3.1 Temperatura

La temperatura del aire es el factor climático con el que más registros y disponibilidad se cuentan en las estaciones meteorológicas. Esta variable explica una parte significativa en el cálculo de la ETo. La temperatura depende directamente de la cantidad de radiación solar absorbida por la atmosfera y del calor emitido por la tierra, transfiriendo energía al cultivo y controlando la tasa de evapotranspiración. Traducido a otros términos; en un

día soleado y cálido, la cantidad de agua perdida por evapotranspiración será mayor que en un día que presenta nubosidades y tenga un clima fresco (Ahmad *et al*., 2017; FAO, 2006).

2.3.2 Radiación solar

La radiación solar (*H*) es la fuente de energía más importante en nuestro planeta y con ella se pueden cambiar grandes cantidades de agua en estado líquido a vapor de agua. Su intensidad estará determinada por el ángulo existente entre la dirección de los rayos del sol y la atmosfera, este ángulo cambia durante el transcurso del día y es diferente en diferentes latitudes del planeta en diferentes estaciones del año, debido a la posición en la que se encuentre el planeta y a su respectivo movimiento alrededor del sol. Dicho esto, la evapotranspiración medida en cualquier parte del planeta, estará definida por la cantidad de energía que se tenga disponible para evaporar agua, tomando en cuenta el grado de transparencia y presencia de nubosidad en dicha región (Ahmad *et al*., 2017; FAO, 2006)..

2.3.3 Humedad relativa

La humedad relativa (*HR*) expresa el grado de saturación del aire, y se puede estimar como la relación entre la presión de vapor actual (e_a) y la presión de vapor a saturación (e ° (T)) a la misma temperatura (T). La evapotranspiración dependerá en gran medida del tipo de región en la cual se está determinando; en regiones tropicales donde a pesar de que el ingreso de energía es considerado alto, la alta humedad presente en el aire reduce la demanda de evapotranspiración, esto se debe a que el aire se encuentra cerca del punto de saturación, provocando que el aire absorba menos agua y por consecuencia la tasa de evapotranspiración es baja. Caso contrario ocurre en regiones áridas, donde se consumen grandes cantidades de agua por la existencia de una gran disponibilidad de energía y el mayor poder de extracción de vapor de la atmosfera (Ahmad *et al*., 2017; FAO, 2006).

2.3.4 Velocidad del viento

Bajo condiciones de humedad relativa elevada, la velocidad del viento afectará la tasa de evapotranspiración en un grado de menor importancia que bajo condiciones de climas áridos. Esto se debe principalmente cuando el aire sobre la superficie evaporante se satura gradualmente con la evaporación del agua. Si en su momento no se sustituye este aire saturado por aires más secos, la remoción de vapor y la tasa de evapotranspiración disminuirán gradualmente.

Cuando se combinan los efectos de los factores meteorológicos de diferente manera pueden presentarse varios escenarios. Si las condiciones atmosféricas caliente y seco debido a la sequedad del aire y de la cantidad de energía disponible como radiación solar directa, la demanda de evapotranspiración será alta, en consecuencia, una mayor cantidad de vapor de agua puede ser almacenado en el aire, el viento puede promover el transporte del agua, permitiendo la pérdida de una mayor cantidad de vapor. Sin embargo; bajo condiciones atmosféricas húmedas, la demanda de evapotranspiración será más baja, debido a la alta humedad del aire y a la presencia de nubosidad. En otros términos, las variaciones en la velocidad del viento, por más pequeña que sea, pueden dar lugar a importantes variaciones en la tasa de evapotranspiración de una determinada región (Ahmad *et al*., 2017; FAO, 2006).

2.4 Estaciones Meteorológicas

2.4.1 Estaciones Meteorológicas Automatizadas (EMAs)

Una estación meteorológica automática (EMAs) se define como una instalación que transmite o registra automáticamente las observaciones obtenidas de los instrumentos de medición. Está conformada por un grupo de sensores que registran y transmiten información meteorológica de forma automática de los sitios donde están estratégicamente colocadas. Su función principal es la recopilación y monitoreo de algunas variables meteorológicas para generar archivos del promedio de cada 10 minutos de todas las variables. Esta información es enviada vía satélite en intervalos de 1 o 3 horas por estación. En una EMAs las medidas de los elementos meteorológicos se

convierten en señales eléctricas a través de sensores. Las señales son procesadas y transformadas en datos meteorológicos. La información resultante se transmite finalmente por cable o radio o se almacena automáticamente en un medio de grabación. Las EMAs se pueden dividir en estaciones en tiempo real, que transmiten automáticamente datos observados a horas fijas, y estaciones fuera de línea, que registran datos en dispositivos de almacenamiento. Las ventajas de contar con una EMAs son el poder realizar una observación continua de las variables meteorológicas. Los datos de observación en las estaciones tripuladas pueden obtenerse incluso cuando no se encuentra el personal presente en la estación. Todos los sistemas totalmente automatizados también se pueden instalar en sitios inaccesibles, se puede reducir tanto el número de observadores como los costos operativos, además se eliminan los errores del observador en la lectura debido a que los datos meteorológicos se toman como señales eléctricas. Las técnicas de observación estandarizadas permiten la homogeneización de los datos observados en regiones donde se adopta la observación automática del clima, entre otras (Ahmad *et al*., 2017)

2.4.2 Componentes de una EMAs

Existen dos tipos de estructuras donde se instalan las EMAs, las de tipo andamio (figura 7) y las de torre triangular (figura 8).

Ambas estructuras constan de unidades principales:

*Unidad de campo para el registro de datos; recopila y almacena toda la información de los datos meteorológicos de cada sensor y los almacena en su propia tarjeta de memoria; además, se encarga del procesamiento de los datos meteorológicos asignado valores promedios, mínimos y máximos de los parámetros registrados. *Terminal de registro de los datos, indica a la unidad de campo qué sensor usar, a qué canal están conectados estos sensores, cuándo almacenar los datos y cómo etiquetar y organizar los datos. También ayuda a ver los datos en la pantalla, verifica la duración de la batería y la memoria restante en el paquete de almacenamiento de datos.

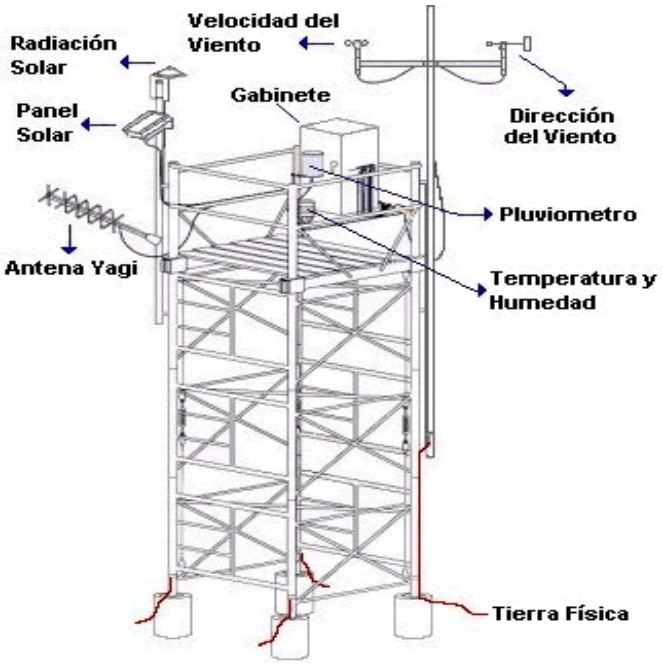


Figura 7. Estructura tipo andamio

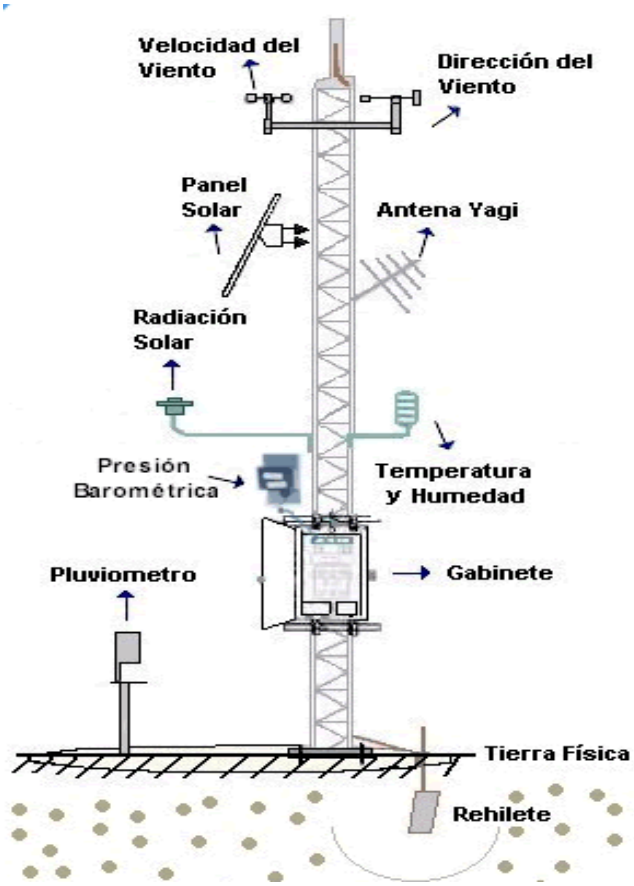


Figura 8. Estructura tipo Torre triangular

*El paquete de almacenamiento de datos, es el lugar donde se almacenan los datos tabulados de las unidades prácticas de medición, este módulo se conecta a la unidad de campo de registro de datos para poder descargar los datos almacenados, y posteriormente transferirlos a una computadora personal o almacenarlos en discos duros para su registro permanente. *Los paneles solares, proporcionan el poder para ejecutar las estaciones meteorológicas. Están equipados con baterías de plomo-ácido recargables. *Sensores, todas las EMAs cuentan con sensores para medir variables meteorológicas, cada sensor registra información de una variable; sensor de temperatura y humedad relativa, sensor de temperatura, sensor de velocidad de viento, sensor de dirección del viento, sensor de radiación solar, sensor de precipitación (Ahmad *et al*., 2017).

2.4.3 Tiempo Universal Coordinado (TUC)

El tiempo universal coordinado, llamado UTC por siglas en inglés, “Universal Time Coordinated”, es una escala de tiempo compuesta: una basada en relojes atómicos y otra basada en la medida del ángulo de rotación de la Tierra con respecto al sol denominado UT1, con una diferencia entre escala de 1 segundo acumulado por año. Es utilizado como zona horaria de referencia respecto a las diferentes zonas horarias divididas en todo el mundo. México cuenta con tres zonas horarias; tiempo del centro, tiempo del pacífico y tiempo del noreste, sumando al horario de verano, lo cual presenta complicaciones al analizar la información obtenida de todas las estaciones meteorológicas establecidas en todo el país (Cruz, 2013).

2.5 Calidad de los datos meteorológicos

2.5.1 Técnicas de interpolación

Una interpolación “es la búsqueda de una función que pasa por todos los puntos deseados” con la finalidad de realizar estimaciones a partir de una muestra cualquiera, que son los valores de Q en un conjunto de puntos (X, Y). También es una técnica para

agregar nuevos puntos de datos dentro de un rango de un conjunto de puntos de datos conocidos. La interpolación se puede clasificar por el tipo de interpoladores que operan a partir de los puntos en un conjunto de datos. La interpolación exacta conserva los valores originales de los puntos en un muestreo; la interpolación no exacta no conserva dichos valores. La interpolación global se encuentra basada en todos los puntos del conjunto de datos y utiliza todo el conjunto de datos para estimar el valor de cada nuevo punto; la interpolación local se basa en los puntos de un subconjunto del conjunto de datos originales y utiliza solo los puntos de muestreo más cercano.

La interpolación también puede clasificarse dependiendo del tipo de problema a resolver o de la función que se utiliza; la interpolación Polinómica es de las más utilizadas, además de la interpolación trigonométrica, exponencial o logarítmica, por mencionar algunas, todas con el objetivo de completar datos faltantes, suavizar los datos existentes, hacer predicciones, entre otras funciones (Chica Jiménez y Bosch, 2018; Fallas, 2007).

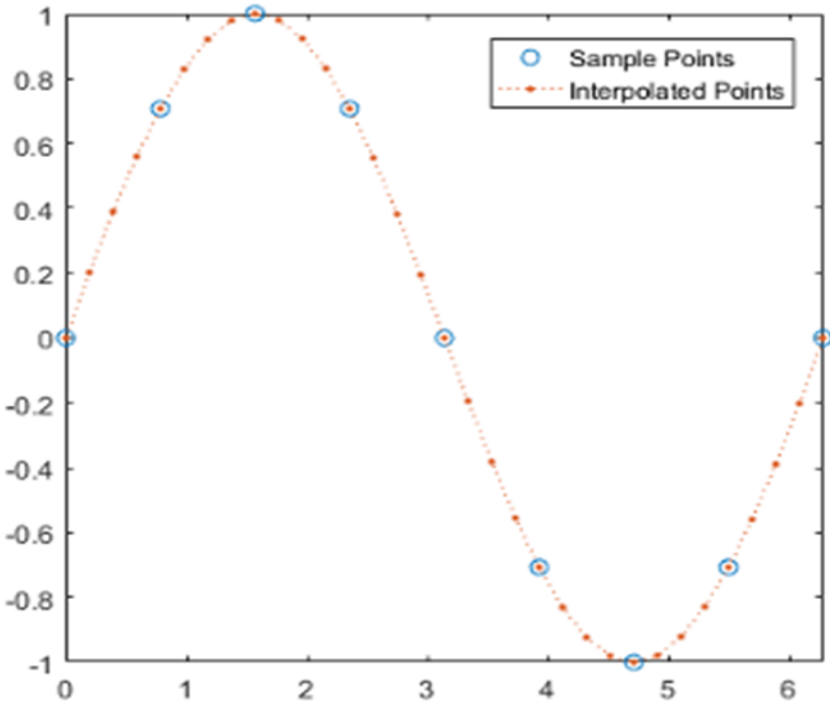


Figura 9. Representación gráfica de una interpolación de datos

2.5.1.1 Interpolación Polinómica de Hermite (PCHIP)

La interpolación Polinómica, como su nombre lo indica, emplea polinomios para interpolar un conjunto de datos; donde, si la interpolación Polinómica es local, es posible determinar la función que identifica la pareja de puntos teniendo en cuenta la información suministrada por el par de puntos. Si la interpolación es global, la función resultante que une un par de puntos, no solo toma en cuenta la información dada por el par de puntos, sino toma en cuenta en su totalidad al conjunto de puntos en la muestra.

La interpolación Polinómica de Hermite, consiste en encontrar el polinomio del grado requerido para satisfacer cada punto x_i. Dada las condiciones $p^{(k)}(x_i) = f^{(k)}(x_i), \forall\, k = 1, \ldots, n_i,$ donde el número de las derivadas consecutivas consideradas puede depender del punto x_i, considerando un número diferente de derivadas de cada punto de muestreo. PCHIP realiza interpolaciones usando trozos de polinomios cúbicos *P(x)*. Cada subintervalo $x_k \leq x \leq x_{k+1}$, el polinomio *P(x)* es una interpolación PCHIP cúbica para cada punto de los datos con derivados específicos en los puntos interpolados, es decir $P(x_j) = y_i$ donde la primera derivada es continua $\frac{dp}{dx}$ y la segunda derivada posiblemente no lo sea $\frac{d^2p}{dx^2}$ (Quintero *et al.*, 2010; Torrente Cantó *et al.*, 2018).

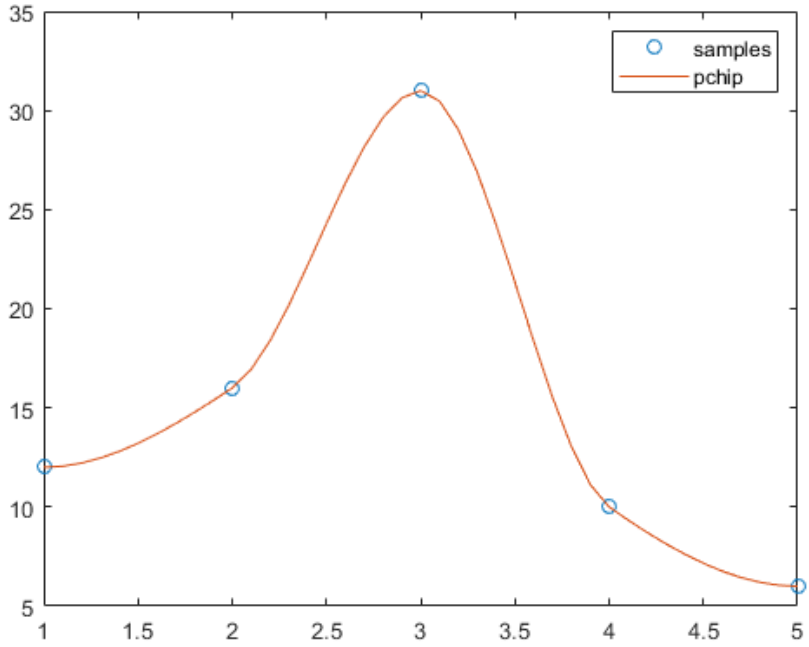


Figura 10. Representación gráfica de interpolación por PCHIP

2.5.1.2 Interpolación "SPLINE"

Los "Splines", son un conjunto de varios polinomios definidos, uno sobre su propio sub-intervalo, que se unen entre sí obedeciendo a ciertas condiciones de continuidad, describiendo tanto la tendencia como la magnitud de una línea, donde la forma de la curva es definida de manera local mediante un subconjunto de datos. Para poder utilizar este método de interpolación, es necesario contar con un conjunto de valores asociados (y_i) y con un conjunto de puntos bases (x_i):

x_i	y_i
x_0	y_0
:	:
x_n	y_n

Tomando en cuenta que siempre se tiene que *n + 1* puntos de muestra como $x_0, x_1, \ldots . x_n$ son los puntos base. Entonces, una función "Spline" interpolante de cualquier grado, con puntos base $x_0, x_1, \ldots . x_n$ es una función *S(X)* formada por varios polinomios,

cada uno definido sobre un intervalo, unido entre sí bajo condiciones de continuidad (Chica Jiménez y Bosch, 2018; Duque Martínez, 2015; Fallas, 2007; Quintero *et al*., 2010).

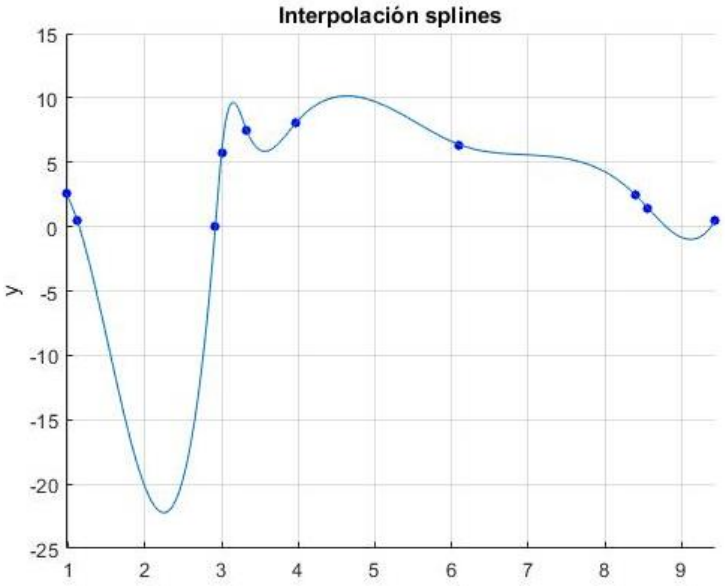


Figura 11. Representación gráfica de interpolación por Spline

2.5.2 Datos atípicos

Se le llaman datos o valores atípicos a las observaciones que parecen desviarse de otras en una muestra y su detección es importante porque nos pueden indicar la presencia de datos erróneos, ya sea por un incorrecto registro o una mala ejecución de un experimento. La presencia de los valores atípicos se debe principalmente a una variación aleatoria. Determinar si un valor atípico es catalogado erróneamente es de importancia, de eso depende si se elimina del análisis o se corrige si es posible. Si los datos presentan valores significativos, es posible considerar el uso de métodos estadísticos para su análisis.

2.5.2.1 Diagrama de caja y “bigotes”; cuartiles)

El diagrama de caja; llamada también de caja y “bigotes”, es la herramienta más utilizada para realizar representaciones de distribución de datos, así como para la identificación de observaciones atípicas en un determinado conjunto de datos. En el diagrama de caja y “bigote” se considera una observación atípica, cuando su valor se encuentra por encima del bigote superior o por debajo del bigote inferior, donde los bigotes son las líneas que sobresalen de la caja; existen casos donde el diagrama presenta asimetría, provocando que el porcentaje de los valores reconocidos como atípicos se tornan excesivos. El diagrama está integrado principalmente por la mediana, el cuartil inferior (Q1) y el cuartil superior (Q3). Consiste en una caja dividida por un segmento vertical que indica donde se encuentra la mediana, donde Q1 y Q3 se definen como los percentiles 25 y 75 e IQR se denomina rango intercuartílico (IQR = Q3-Q1), este es un rango robusto para la interpretación porque la región central del 50% no se ve afectada por valores atípicos o extremos, y ofrece una visualización menos sesgada de la extensión de las curvas (Bruffaerts *et al.*, 2014; Garcia, 2012)

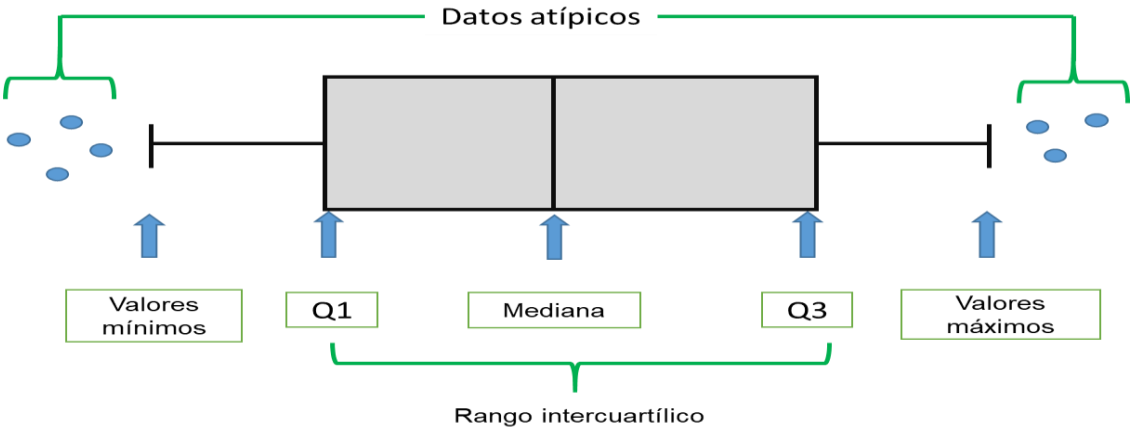


Figura 12. Diagrama de caja y “bigotes”

2.5.2.2 Método de “Mean” (método de la media)

En este método de detección, la media y la desviación estándar de los residuales son calculados y comparados. Si un valor se halla a cierta desviación estándar de la media,

el dato puede ser marcado como un valor atípico. El número específico de desviaciones estándar se llama umbral, en este trabajo se fijaron 3 desviaciones estándar.

$\bar{x} \pm \sigma^2$ Donde $\bar{x}$ es la media y σ^2 es la desviación estándar de la muestra, respectivamente.

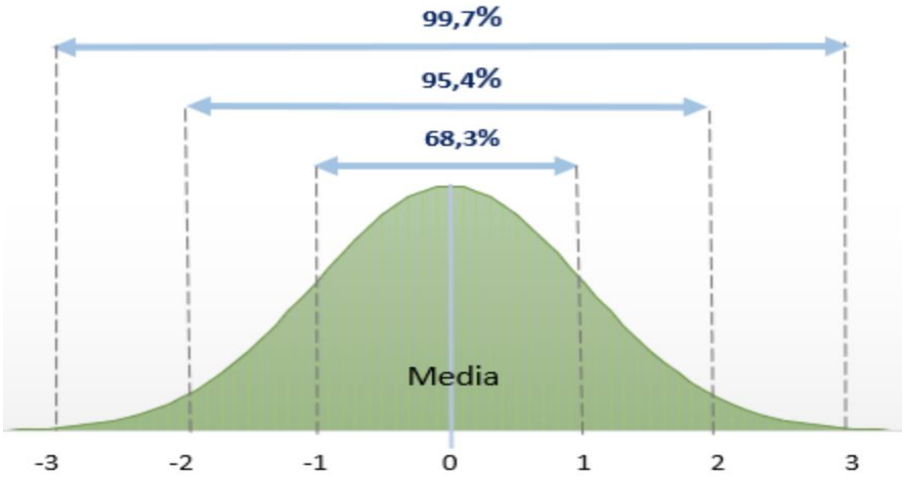


Figura 13. Representación de la técnica Mean

2.5.2.3 Método de Grubbs

La prueba de Grubbs detecta un valor atípico a la vez, asumiendo una distribución normal; a diferencia del método mean (media), donde el número de desviaciones estándar usados depende del tipo de problema a resolver, la prueba Grubbs solo usa una distribución estándar. Este valor atípico se elimina del conjunto de datos y la prueba se repite hasta que no se detectan valores atípicos. Sin embargo, las iteraciones múltiples cambian las probabilidades de detección, y la prueba no se debe usar para tamaños de muestra de seis o menos, ya que con frecuencia marca la mayoría de los puntos como valores atípicos (Garcia, 2012)

CAPÍTULO III. PLANTEAMIENTO DEL PROBLEMA

Para realizar una planeación eficiente del riego, una alternativa a implementar consiste en la determinación de la evapotranspiración del cultivo (ETc) que se puede obtener de manera directa utilizando lisímetros (hidrológico), torres de Eddy Covarianza (micrometeorológico) o método de flujo de savia (fisiología de la planta) por mencionar algunos. La principal desventaja que presentan estos métodos son los altos costos de instalación y mantenimiento, por lo que su uso muchas veces se limita a centros de investigación o empresas que cuenten con financiamiento para implementar estos métodos. Por otra parte, la ETc se puede determinar de manera indirecta mediante el uso de la evapotranspiración de referencia (ETo), que al multiplicarla por un coeficiente de cultivo (Kc) (integra los efectos de las características que distinguen una superficie cultivada de una superficie de referencia) se puede realizar una estimación de la ETc (ETc = ETo X Kc). La ecuación de la FAO56 Penman-Monteith ha mostrados buenos resultados en su implementación para determinar la ETo, además de ser la ecuación estándar para calibrar otras ecuaciones empíricas o semi-empíricas, sin embargo, la implementación de este método requiere de datos de las variables meteorológicas temperatura, humedad relativa, radiación solar y velocidad de viento, que en muchas ocasiones, las estaciones meteorológicas de algunas regiones no miden, siendo la radiación y la velocidad del viento las menos medidas, lo cual hace que la utilización de este método no sea la más viable para la estimación de la ETo. Ante estas condiciones, para realizar estimaciones de ETo, se utilizan ecuaciones empíricas que funcionan con menos variables meteorológicas y su implementación en lugares donde no se cuenta con todas las variables es viable para determinar la ETo. Sin embargo, estas ecuaciones presentan la desventaja de que requieren de calibración para obtener buenos resultados (FAO, 2006). En el contexto anterior, en el presente estudio se propone estimar con precisión valores de ETo usando datos de temperatura y radiación solar extraterrestre (Ho) como valores de entrada, mediante ecuaciones empíricas y métodos de soft-computing.

CAPÍTULO IV. JUSTIFICACIÓN

En el manejo sustentable del agua y estudios hidrológicos, la determinación precisa de la ETo es de suma importancia. Su cálculo exacto se realiza comúnmente usando la ecuación FAO56-PM. Sin embargo, esta ecuación requiere de la determinación de cuatro variables meteorológicas: radiación solar, humedad relativa, velocidad y temperatura del viento, que son medidas únicamente por las estaciones meteorológicas automatizadas. En el contexto anterior, con el objetivo de estimar la ETo con precisión como un insumo para la determinación de la evapotranspiración de los cultivos ETc, en este trabajo se propone hacer uso de ecuaciones empíricas y métodos de inteligencia artificial basados únicamente en datos de temperatura. La temperatura es la variable más medida en las estaciones meteorológicas convencionales, siendo que en el estado de Campeche la relación de estaciones meteorológicas automatizadas sobre las convencionales es de 1:160. El uso de las ecuaciones empíricas basadas en temperatura para la estimación de la ETo radica en la facilidad para su implementación, ya que, con solo mediciones de temperatura y la fórmula calibrada es posible implementarla mediante el uso de alguna calculadora de bolsillo u hoja de cálculo. Por otra parte, el presente estudio también sugiere el uso de técnicas modernas de “soft-computing” o de inteligencia artificial para estimar la ETo, esto debido a la enorme capacidad de procesamiento de información de las computadoras modernas.

El uso de los métodos de inteligencia artificial puede considerarse como una herramienta adecuada para su predicción (Mehdizadeh, 2018), ya que presentan algunas ventajas, como la posibilidad de obtener soluciones más simples para problemas multivariables y el manejo de grandes cantidades de información para el entrenamiento de datos y el autoaprendizaje. Estas técnicas han sido ampliamente usadas en la modelación hidrológica y en la estimación de la ETo han mostrado superioridad sobre las ecuaciones empíricas, debido a que incrementan la precisión de las estimaciones utilizando pocas variables.

CAPÍTULO V. HIPÓTESIS

Las ecuaciones empíricas y los métodos de soft-computing basados en datos de temperatura estiman con precisión valores de evapotranspiración de referencia.

CAPÍTULO VI. OBJETIVOS

6.1. Objetivo General

- Evaluar la capacidad de dos ecuaciones empíricas y tres métodos de soft-computing basados en temperatura para estimar evapotranspiración de referencia (ETo).

6.2. Objetivos específicos

- Evaluar dos técnicas de interpolación para el relleno de información faltante en bases de datos meteorológicos.
- Evaluar tres técnicas para la detección de valores atípicos en bases de datos meteorológicos.
- Calibrar y evaluar dos ecuaciones empíricas basadas en temperatura para estimar la ETo
- Evaluar la precisión de tres métodos de soft-computing basados en parámetros meteorológicos para estimar la ETo.

CAPÍTULO VII. MATERIALES Y MÉTODOS

7.1 Descripción del sitio de estudio

La presente investigación se realizó utilizando datos de estaciones meteorológicas automatizadas ubicadas en el estado de Campeche, México. El clima predominante es cálido subhúmedo, que se presenta en el 92% de su territorio, el 7.75% presenta clima cálido húmedo localizado en la parte este del estado y en la parte norte, un pequeño porcentaje del 0.05% con clima semi seco. La temperatura más alta es mayor a 30°C y la mínima de 18°C. La temperatura media anual es de 26 a 27°C. Las lluvias son de abundantes a muy abundantes durante el verano. La precipitación total anual varía entre 1 200 y 2 000 mm, y en la región norte, de clima semi seco, es alrededor de 800 mm anuales (INEGI, 2017).

7.1.1 Ubicación de las EMAs

Las bases de datos meteorológicos utilizadas se obtuvieron de las estaciones meteorológicas automatizadas que se muestran en la Cuadro 1.

Cuadro 1. Ubicación geográfica en coordenadas de las estaciones meteorológicas automatizadas EMAs utilizadas en el presente trabajo.

Estación	Latitud (Norte)	Longitud (Oeste)	Altitud
Calakmul	18º 21' 54"	89º 53' 33"	28
Campeche	19° 50' 10"	90° 30' 26"	3
Cd del Carmen	18º 39' 29"	91º 45' 55"	4
Escárcega	18º 36' 30"	90º 45' 14"	60
Los Petenes	19º 56' 36"	90º 22' 26"	2
Monclova	18º 03' 25"	90º 49' 15"	100

Fuente: SNM.CONAGUA.2018.

7.2 Datos y análisis de calidad

La información de las variables meteorológicas a estudiar en la presente investigación fue obtenida de los registros de la Comisión Nacional del Agua (CONAGUA).

Cuadro 2. Información histórica de las estaciones meteorológicas automatizadas (EMAs)

Estación meteorológica	Años de registro
Calakmul	2000 al 2018
Campeche	2003 al 2018
Cd del Carmen	2000 al 2018
Escárcega *	2004 al 2018
Monclova	2008 al 2018
Los Petenes	2012 al 2018

Fuente: SNM.CONAGUA.2018.
*= El número de datos meteorológicos de esta estación no corresponde al número de años de registro, a partir del año 2011-2018.

En el cuadro 2 se encuentra la información registrada por cada estación meteorológica, agrupada en carpetas por cada año de registro y el número de años depende directamente del año de operación de cada EMAs. La hora utilizada en el registro de los datos es el Tiempo Universal Coordinado (TUC o UTC). Las variables meteorológicas que se registran son: promedio de dirección del viento medido en grados (0º = norte), velocidad del viento medido en km/h, dirección del viento más fuerte con duración mayor a 5 segundos, medido en grados, y velocidad del viento máximo con duración mayor a 5 segundos en intervalos de 10 minutos (Km h^{-1}), temperatura promedio del aire (°C), humedad de relativa promedio en porcentaje (%), presión barométrica (milibar), precipitación de lluvia acumulada (mm), y radiación solar (watt m^{-2}). El intervalo de tiempo de registro de la información fue de 10 minutos, con excepción del año 2015 que los datos fueron registrados en intervalos de 1 hora.

7.2.1 Estado de las bases de datos meteorológicos

La revisión de cada base de datos de cada año de registro por estación se realizó de forma manual para conocer el estado en el que se encontraban dichas bases de datos.

7.2.1.1 Adecuación de los datos meteorológicos

Para procesar de una manera adecuada la información de las estaciones meteorológicas y obtener resultados consistentes, es necesario que esta información esté completa y con los formatos adecuados para su procesamiento con el fin de evitar posibles sesgos en los resultados. De las bases de datos se seleccionaron las variables meteorológicas: Velocidad del viento (Km/h), temperatura (°C), humedad relativa (%), precipitación pluvial (mm) y radiación solar (W/m^2), eliminando el resto de las variables en la base de datos.

7.2.1.2 Ajuste del formato de fecha y de hora

Para homogenizar el formato de fecha y hora en todas las estaciones meteorológicas en estudio, se implementaron códigos macro en Microsoft Excel. Esto con la finalidad de convertir la hora de registro UTC a hora local, recordando que las estaciones registran la información con Tiempo Universal Coordinado (TUC o UTC).

7.2.1.3 Arreglo de series consecutivas

Posterior al ajuste del formato de hora y fecha, se procedió a localizar series de tiempo faltantes en las bases de datos. Para llevar a cabo esto, se utilizó el código macro (1) “consecutivo”, el cual se implementó en Microsoft Excel, localizando y añadiendo series de tiempo faltantes.

7.2.2 Manejo de datos faltantes

Los datos faltantes se completaron mediante las siguientes técnicas de interpolación; Interpolación Polinómica de Hermite PCHIP y SPLINE.

Los códigos 2 y 3 (sección de anexo) muestran las líneas de comandos implementadas en el software Matlab.

Para evaluar la exactitud de las técnicas de interpolación (PCHIP y SPLINE), de las bases de datos, se eligieron series consecutivas al azar de entre 20 y 25 días en diferentes años, y se eliminaron intencionalmente series consecutivas de entre 10 y 15 días para después aplicar los métodos de interpolación.

Con el objetivo de evaluar el ajuste, se procedió a realizar un análisis comparativo de lo observado sobre lo estimado mediante regresión lineal para cada una de las variables meteorológicas y considerando el valor del coeficiente de determinación R^2.

7.2.3 Detección de valores atípicos

Para la detección de valores atípicos se utilizaron las técnicas Grubbs, mean y Cuartiles, los cuales se implementaron utilizando el software Matlab en conjunto con los códigos 4, 5, y 6 encontrados en la sección de anexos.

7.3 Estimación de la ETo

7.3.1 Fórmula de la FAO-56 PM

Para calcular la ETo mediante la fórmula de la FAO-56 PM, que se utilizó como referencia para evaluar la precisión de los demás métodos, se utilizó una hoja de cálculo diseñada en Microsoft Excel, la cual está estructurada por 41 columnas de las cuales 7 son columnas para los valores de entradas.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Fecha	Dia Juliano	Precipitacion	Tmax	Tmin	Viento	Vel V	R. Global	R.Global	Humedad	Tmed	Prueba logica
2		Dia	mm	°C	°C	Km/h	m/s	WM^{-2}	(MJM^{-2} dia^{-1})	%	°C	Rad Solar
3	27/04/2000	118	0.25	31.40	26.70	12.90	3.58	10133.00	6.08	76.48	29.05	0.00
4	28/04/2000	119	0.00	36.80	23.30	11.81	3.28	42310.00	25.39	69.91	30.05	0.00
5	29/04/2000	120	0.00	35.30	23.50	12.83	3.56	43776.00	26.27	69.06	29.40	0.00
6	30/04/2000	121	0.00	37.40	22.60	11.98	3.33	44509.00	26.71	69.24	30.00	0.00
7	01/05/2000	122	0.00	38.30	22.50	11.94	3.32	44810.00	26.89	66.79	30.40	0.00
8	02/05/2000	123	0.00	38.40	22.60	13.03	3.62	43267.93	25.96	63.29	30.50	0.00
9	03/05/2000	124	0.00	37.50	22.00	12.65	3.51	40220.00	24.13	63.92	29.75	0.00
10	04/05/2000	125	0.00	36.90	23.40	12.01	3.34	39926.00	23.96	66.07	30.15	0.00

Figura 14. Valores de entrada de la fórmula de la FAO-56 PM en la hoja de cálculo de Excel

En la figura 14 se puede observar la distribución de los valores de entradas de fecha, precipitación, temperatura máxima, temperatura mínima, velocidad de viento, radiación solar y humedad relativa. Para el caso de algunos de los valores de entrada, la hoja de cálculo tiene incluidos conversores que, para los valores de fecha los convierte a día juliano, la velocidad del viento en km h^{-1} la convierte a m s^{-1}, la radiación solar medida en W m^{-2} la convierte a MJ m^{-2}día^{-1} (megajoules metro cuadrado día). También se anexó una prueba lógica para la radiación solar con la finalidad de encontrar valores atípicos.

La hoja de cálculo de la fórmula de la FAO-56 PM también tiene un apartado de parámetros donde se incluyen principalmente los factores de conversión de viento y radiación solar (de WM^{-2} a MJM^{-2} día^{-1}), una constante solar, y los datos de ubicación de cada estación meteorológica automatizada como se observa en la figura 15.

N	O
Parámetros	
Datos	Valor
F.Conversión Viento	0.2777778
F.Conversión Solar	0.0006
Constante solar (Gsc)	118.5495
Estacion	Campeche
Estado	Campeche
Municipio	Campeche
Latitud	19.8361111
Longitud	90.5072222
Altitud	3
Lat.radianes	0.3462054

Figura 15. Parámetros de la hoja de cálculo de la FAO-56 PM

El resto de las columnas de la hoja de cálculo son valores de salida, encontrando el valor de la ETo estimada por la fórmula de la FAO-56 PM en la columna 40.

7.3.2 Ecuaciones Empíricas

7.3.2.1 Calibración

La calibración de las ecuaciones empíricas se llevó a cabo utilizando el software MATLAB v2018, para tal propósito se utilizaron los códigos 7 y 8 encontrados en la sección de anexos.

7.3.3 Métodos de inteligencia artificial o "*Soft-computing*"

7.3.3.1 Máquinas de Soporte Vectorial (MSV)

La estimación de la ETo con el método de Maquinas de Soporte Vectorial (SVM por sus siglas en inglés) se realizó mediante el software R (RDevelopment, 2009). Las variables de entrada fueron los valores de los datos meteorológicos de T_{max}, T_{min} y Ho, y

ETo - FAO56-PM como variable objetivo o respuesta. El modelo SVM se construyó utilizando el paquete LIBSVM 3.1 (Chang *et al*., 2013), el cual se ejecutó en el software R. Para la redimensión de los datos se utilizó el Kernel de Función de Base Radial (RBF). Por otra parte para evitar un sobreajuste del modelo SVM se llevó a cabo un ajuste previo de los parámetros de la SVM y del Kernel, esto se realizó mediante el uso del algoritmo genético (GA) contenido en la librería e1079 para el software R, el paquete "Caret", y la técnica de validación cruzada (Cross validation) (Quej *et al*., 2017; Shrestha & Shukla, 2015).

Los parámetros de SVM a optimizar fueron el costo (C) y épsilon (ε) y el parámetro gamma (γ) del kernel RBF. Las líneas de comandos en la implementación del GA se hallan en el código (9), esta técnica frecuentemente es utilizada en la optimización de parámetros, que está basada en principios de selección natural y la genética en los sistemas biológicos y combina operadores de selección, generación, cruzamiento y mutación con la finalidad de identificar la mejor solución al problema de optimización que se esté tratando (Antonanzas-Torres *et al*., 2015; Quej *et al*., 2017).

Posterior a la optimización de los parámetros se procedió a construir el modelo SVM utilizando el código (10), el cual divide el conjunto de datos meteorológicos para el entrenamiento y para la validación, haciendo una modificación a lo usado por Quej *et al*., (2017). El 60% de los datos se usaron en la etapa de entrenamiento y el 40% para la validación.

7.3.3.2 Programación de Expresión Genética (GEP)

La implementación de este método en la estimación de la ETo se llevó a cabo mediante el programa computacional llamado GenexprooTols 5.0. Las variables de entrada fueron los valores de los datos meteorológicos de T_{max}, T_{min} y Ho con la ETo estimada con el método de la FAO56-PM como variable respuesta. Los operadores aritméticos y funciones matemáticas implementadas dentro del programa fueron $\{+, -, \times, \div, \sqrt{x}, \sqrt[3]{x}, x^2, x^3, In(x), e^x, \sin(x), \cos(x), Arctan(x)\}$ (Mattar, 2018; Mehdizadeh *et al.*, 2017; Shiri, 2017). Para la etapa de entrenamiento se utilizaron el 70% de los datos

meteorológicos y el 30% para la validación. No es necesario realizar un ajuste preliminar de los parámetros del modelo, ya que el mismo programa realiza este proceso. Los parámetros GEP que se utilizaron en el presente estudio se muestran en el Cuadro 3 (Shiri *et al*., 2014).

Cuadro 3. Parámetros del modelo GEP

Parámetro	***Valor***
Numero de cromosomas	30
Tamaño de cabeza	8
Número de genes	3
Función de enlace	Adición
Tipo de error en la función fitness	RMSE
Tasa de mutación	0.044
Tasa de inversión	0.1
Tasa de recombinación primer punto	0.3
Tasa de recombinación segundo punto	0.3
Tasa de recombinación de genes	0.1
Tasa de transposición de genes	0.1
Tasa de transposición de la secuencia de inserción	0.1
Raíz Inserción Secuencia Transposición	0.1
Herramienta de penalización	Pp*

*Presión de Parsimonia (Parsimony pressure)

7.3.3.3 Máquina de aumento del Gradiente Extremo (XGBoost)

Para la implementación del método XgBoost, primero se realizó el ajuste de los parámetros del modelo XGBoost utilizando el software R en conjunto con el paquete llamado “xgboost” (código 11), realizando validación cruzada con el fin de evitar el sobreajuste de los datos durante la etapa de entrenamiento. Las variables de entrada fueron los valores de los datos meteorológicos de $T_{max,}$ T_{min} y Ho, y la ETo FAO56-PM como variable respuesta. Los parámetros del modelo XGBoost a calibrar son nrounds, max_depth, eta, gamma, colsample bytree, min_child_weight, y subsample (Fan *et al.*,

2018). El ajuste de los parámetros antes mencionados nos permite conocer el valor de exacto de cada uno de ellos, lo que impide un sobre entrenamiento posterior del modelo.

Una vez ajustados los parámetros del modelo XGBoost, se procedió a entrenar y validar el modelo mediante el código (12) el cual nos permitió realizar las estimaciones de la ETo. Durante la ejecución del software, se utilizó el 60% de los datos para el entrenamiento y 40% para la etapa de validación.

7.4 SOFTWARE

7.4.1 MATLAB

La plataforma MATLAB está diseñada para resolver problemas de ingeniería y científicos. El lenguaje MATLAB basado en matrices es la forma más natural del mundo para expresar matemáticas computacionales. Los gráficos incorporados facilitan la visualización y el conocimiento de los datos. Una vasta biblioteca de cajas de herramientas pre-construidas le permite comenzar de inmediato con algoritmos esenciales para su dominio. El entorno de escritorio invita a la experimentación, exploración y descubrimiento. Estas herramientas y capacidades de MATLAB están todas rigurosamente probadas y diseñadas para trabajar juntas. Se utiliza para aprendizaje automático, procesamiento de señales, procesamiento de imágenes, visión artificial, comunicaciones, finanzas computacionales, diseño de control, robótica y mucho más (MATLAB, 2018)

Las técnicas de interpolación (PCHIP y SPLINE), la detección de valores atípicos (Grubbs, Mean, Cuartiles) y el ajuste de los coeficientes de las ecuaciones empíricas (Hargreaves-Samani y Camargo) fueron implementados usando este programa.

7.4.2 The R Project

R es un conjunto integrado de herramientas de software para la manipulación, el cálculo y la visualización gráfica de datos. Es un entorno de software libre para computación estadística y gráficos. Compila y se ejecuta en una amplia variedad de plataformas UNIX, Windows y MacOS

Entre otras cosas tiene:

• Manejo efectivo de datos y fácil almacenamiento
• Conjunto de operadores para cálculos de arreglos, en particular matrices
• Colección grande, coherente e integrada de herramientas intermedias para el análisis de datos.
• Facilidades gráficas para el análisis y visualización de datos ya sea directamente en la computadora o en una copia impresa.
• Lenguaje de programación bien desarrollado, simple y efectivo (llamado "S") que incluye condicionales, bucles, funciones recursivas definidas por el usuario y facilidades de entrada y salida. De hecho, la mayoría de las funciones proporcionadas por el sistema están escritas en el lenguaje S (Venables y Smith, 2013).

7.4.3 GeneXproTools 5.0

Es una herramienta de modelado extremadamente flexible diseñada para regresión, regresión logística, clasificación, predicción de series de tiempo y síntesis lógica. Puede procesar conjuntos de datos con decenas de miles de variables y extraer sin esfuerzo las características más significativas y sus relaciones. Es una aplicación muy fácil de usar que simplifica el acceso a todos los tipos de almacenes de datos, desde archivos de texto en bruto a bases de datos y hojas de cálculo de Excel. No es necesario conocer ningún lenguaje de programación para crear modelos potentes y precisos.

GeneXproTools proporciona todas las herramientas necesarias para limpiar y analizar sus datos, manejar sus conjuntos de datos, generar modelos, analizarlos y luego aplicarlos inmediatamente a cualquier base de datos nueva utilizando su motor de puntuación flexible.

Puede implementar modelos individuales o conjuntos de modelos en Excel, con la generación automática de los modelos de voto mayoritario y probabilidad media para clasificación y regresión logística, y los modelos promedio y mediana para la regresión y la predicción de series de tiempo. Genera modelos que son al mismo tiempo muy precisos y tienen una alta capacidad de generalización y se puede comprender y analizar

completamente los modelos evolucionados, ya que GeneXproTools traduce automáticamente los modelos que genera utilizando su código Karva nativo en un amplio conjunto de lenguajes de programación (Ada, C, C ++, C #, Fortran, Java, Java Script, Matlab, Pascal, Perl, PHP, Python, Visual Basic, VB.Net, Verilog y VHDL) mediante el uso de gramáticas integradas.

7.5 Índices estadísticos para la evaluación de modelos de predicción de la ETo

El desempeño de los modelos, empíricos y de soft-computing, se evaluó mediante los siguientes índices estadísticos:

Coeficiente de determinación (R^2)

Indica el grado de ajuste de la recta de regresión a los valores de la muestra y es definida como el porcentaje de variabilidad total de la variable dependiente (*Y*), explicada por la recta de regresión. Dado que R^2 es una cantidad adimensional que solo puede tomar valores entre [0,1], cuanto más cercano a la unidad se encuentre este valor, mejor será el ajuste, indicando que la fuerza de asociación entre las variables es mayor; sin embargo, si el valor se encuentra cercano a cero, es un indicativo de que no existe asociación entre las variables.

El R^2 está definido de la siguiente manera:

$$R^2 = \frac{\left[\sum_{i=1}^{n}(P_i - P_{prom})(O_i - O_{prom})\right]^2}{\sum_{i=1}^{n}(P_i - P_{prom})^2 \sum_{i=1}^{n}(O_i - O_{prom})^2} \quad (14)$$

Donde:

n = Número de datos

P_i = Valores estimados

P_{prom} = Promedio de los valores estimados

O_i = Valores observados

O_{prom} = Promedio de los valores observados

Raíz Cuadrada Media del Error (RMSE)

Llamada RMSE por sus siglas en inglés (Root Mean Squared Error), consiste en la raíz cuadrada de la sumatoria de los errores cuadráticos. Amplifica y penaliza con mayor fuerza aquellos errores de mayor magnitud. Realiza una diferencia entre los valores estimados y los valores reales, estas diferencias se elevan al cuadrado y se calcula el promedio de todas. Al promedio se le debe calcular su raíz cuadrada. La fórmula para calcularlo es la siguiente:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2} \quad (15)$$

Donde:

n = Número de datos

P_i = Valores estimados

O_i = Valores observados

Error Absoluto Medio (MAE)

Indica que tan cercana es la predicción realizada (valores estimados) al resultado real (valores observados), haciendo una diferencia entre el valor obtenido y el valor real en valores absolutos, después se calcula el promedio. La ecuación es la siguiente:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(|P_i - O_i|) \quad (16)$$

Donde:

n = Número de datos

P_i = Valores estimados

O_i = Valores observados

CAPÍTULO VIII. RESULTADOS Y DISCUSIÓN

En la presente investigación, de manera inicial se realizó el análisis de la calidad de los datos, con el objetivo de contar con información precisa y confiable que nos permitiera obtener resultados y conclusiones consistentes. Posteriormente se eligieron las ecuaciones empíricas de Hargreaves y Samani, así como la de Camargo para estimar la ETo usando solamente datos de temperatura, no sin antes realizar una calibración de los parámetros de las mismas para después evaluar su desempeño y capacidad para estimar con precisión valores de ETo. Por otra parte, aprovechando la capacidad de procesamiento de las computadoras actuales, se evaluó la capacidad de tres técnicas de soft computing o inteligencia artificial (SVM, GEP y XGBoost) para estimar valores de ETo. Al final se realizó una comparación de resultados de las ecuaciones empíricas sobre las técnicas de inteligencia artificial.

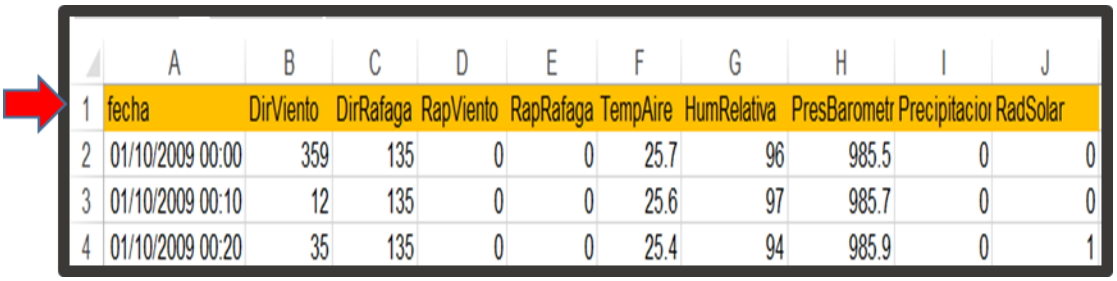
8.1 Análisis de la calidad de datos

8.1.1 Estado de la calidad de datos

Después del análisis realizado a la información contenida en los archivos de cada estación meteorológica, se encontraron bases de datos en diferentes años en una misma estación con diferentes formatos de hora-fecha y nombre de los encabezados. Tomando como ejemplo la estación de Calakmul, en el año 2004 se observa que el registro de la hora y la fecha se encuentran en columnas separadas, los encabezados de las variables medidas tienen abreviaturas de dichas variables en su respectiva columna (figura 16). En cambio, en el año 2009 en la misma estación meteorológica se encontró que los registros tanto de hora como de fecha se encuentran en la misma columna y los encabezados presentan el nombre completo de la variable climática a registrar (figura 17). Esta des-uniformidad en las bases de datos fue observada en todas las bases de datos proporcionadas por la CONAGUA.

	A	B	C	D	E	F	G	H	I	J	K
1	Date	Time	Dir	WSMDir	WSK	WSMK	AvgTemp	AvgRh	AvgBP	Rain	AvgSR
2			deg	deg	kph	kph	C	%	mbar	mm	W/m^2
3	2003 Feb 04	19:30	124	111	5.8	12	29.7	0	985.6	0	106
4	2003 Feb 04	19:40	130	144	3.7	7.4	29.5	0		0	

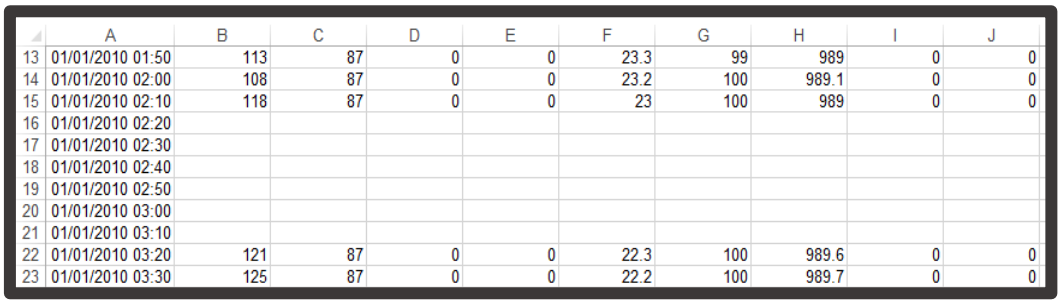
Figura 16. Hora y fecha en columnas diferentes (febrero 2004). Estación Calakmul.



	A	B	C	D	E	F	G	H	I	J
1	fecha	DirViento	DirRafaga	RapViento	RapRafaga	TempAire	HumRelativa	PresBarometr	Precipitacior	RadSolar
2	01/10/2009 00:00	359	135	0	0	25.7	96	985.5	0	0
3	01/10/2009 00:10	12	135	0	0	25.6	97	985.7	0	0
4	01/10/2009 00:20	35	135	0	0	25.4	94	985.9	0	1

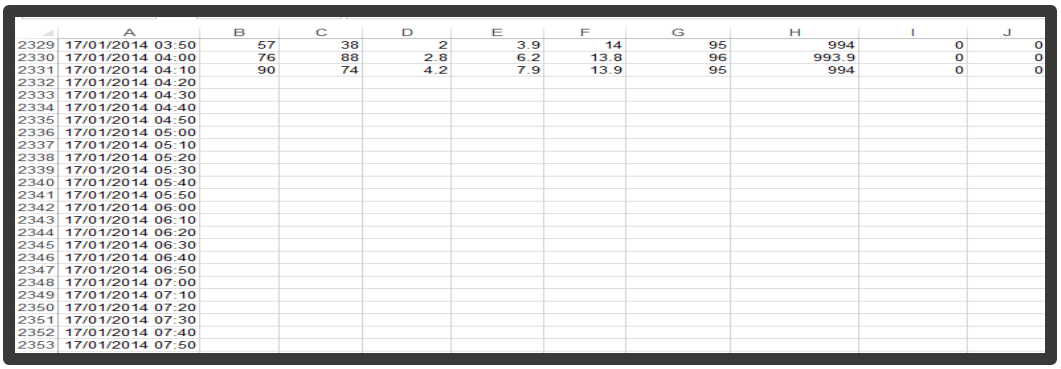
Figura 17. Hora y fecha en la misma columna (octubre 2009). Estación Calakmul.

De igual manera, se encontró que en las bases de datos también existían datos faltantes en ciertos periodos de tiempo, cortos o largos, tal y como se aprecia en las figuras 18 y 19.



	A	B	C	D	E	F	G	H	I	J
13	01/01/2010 01:50	113	87	0	0	23.3	99	989	0	0
14	01/01/2010 02:00	108	87	0	0	23.2	100	989.1	0	0
15	01/01/2010 02:10	118	87	0	0	23	100	989	0	0
16	01/01/2010 02:20									
17	01/01/2010 02:30									
18	01/01/2010 02:40									
19	01/01/2010 02:50									
20	01/01/2010 03:00									
21	01/01/2010 03:10									
22	01/01/2010 03:20	121	87	0	0	22.3	100	989.6	0	0
23	01/01/2010 03:30	125	87	0	0	22.2	100	989.7	0	0

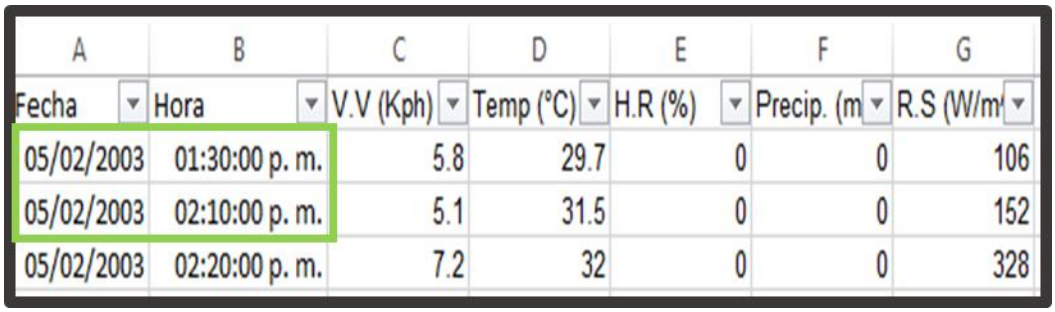
Figura 18. Periodo corto sin registro de datos meteorológicos.



	A	B	C	D	E	F	G	H	I	J
2329	17/01/2014 03:50	57	38	2	3.9	14	95	994	0	0
2330	17/01/2014 04:00	76	88	2.8	6.2	13.8	96	993.9	0	0
2331	17/01/2014 04:10	90	74	4.2	7.9	13.9	95	994	0	0
2332	17/01/2014 04:20									
2333	17/01/2014 04:30									
2334	17/01/2014 04:40									
2335	17/01/2014 04:50									
2336	17/01/2014 05:00									
2337	17/01/2014 05:10									
2338	17/01/2014 05:20									
2339	17/01/2014 05:30									
2340	17/01/2014 05:40									
2341	17/01/2014 05:50									
2342	17/01/2014 06:00									
2343	17/01/2014 06:10									
2344	17/01/2014 06:20									
2345	17/01/2014 06:30									
2346	17/01/2014 06:40									
2347	17/01/2014 06:50									
2348	17/01/2014 07:00									
2349	17/01/2014 07:10									
2350	17/01/2014 07:20									
2351	17/01/2014 07:30									
2352	17/01/2014 07:40									
2353	17/01/2014 07:50									

Figura 19. Periodo largo sin registro de datos meteorológicos.

Existieron casos donde las bases de datos registraban “saltos de tiempo”, por lo tanto, tampoco registraban los datos meteorológicos en esos intervalos, como se observa en la figura 20.



A	B	C	D	E	F	G
Fecha	Hora	V.V (Kph)	Temp (°C)	H.R (%)	Precip. (m	R.S (W/m
05/02/2003	01:30:00 p. m.	5.8	29.7	0	0	106
05/02/2003	02:10:00 p. m.	5.1	31.5	0	0	152
05/02/2003	02:20:00 p. m.	7.2	32	0	0	328

Figura 20. Ausencia de periodos de registro en las bases de datos.

La estación meteorológica registraba información a cierta hora y posteriormente hacia un salto de tiempo a otra hora del día. Estos saltos sin registros estaban dados en rangos cortos de minutos o rangos largos de horas, recordando que las estaciones realizan registro de información cada 10 minutos.

8.1.1.1 Ajuste del formato de fecha y hora

Para reparar esas inconsistencias en las bases de datos, el formato de fecha se ajustó mediante una macro ejecutada en Microsoft Excel. El código modifica el formato no

viable, separando las fechas de los valores horarios como se observa en la parte derecha de la figura 21.

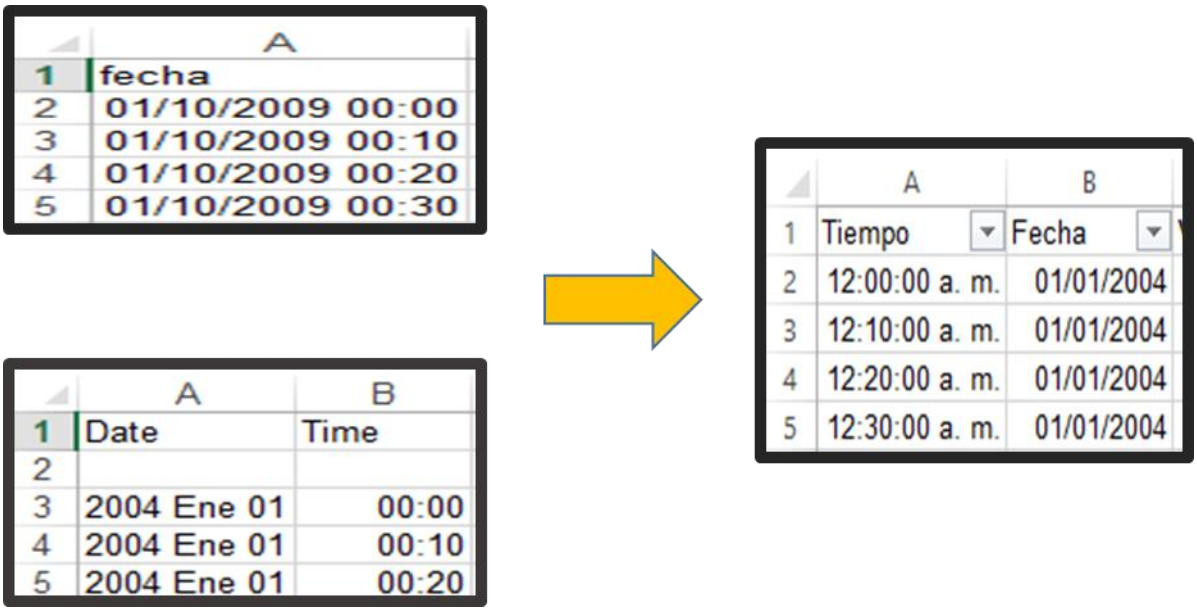


Figura 21. Ajuste en los formatos de hora y fecha mediante códigos macros implementados en Microsoft Excel.

8.1.1.2 Arreglo en las series consecutivas

Para completar las series consecutivas de tiempo se utilizó el código 1 (sección de anexos), el cual busca los “saltos de tiempo” e inserta las filas correspondientes como se aprecia en la figura 22.

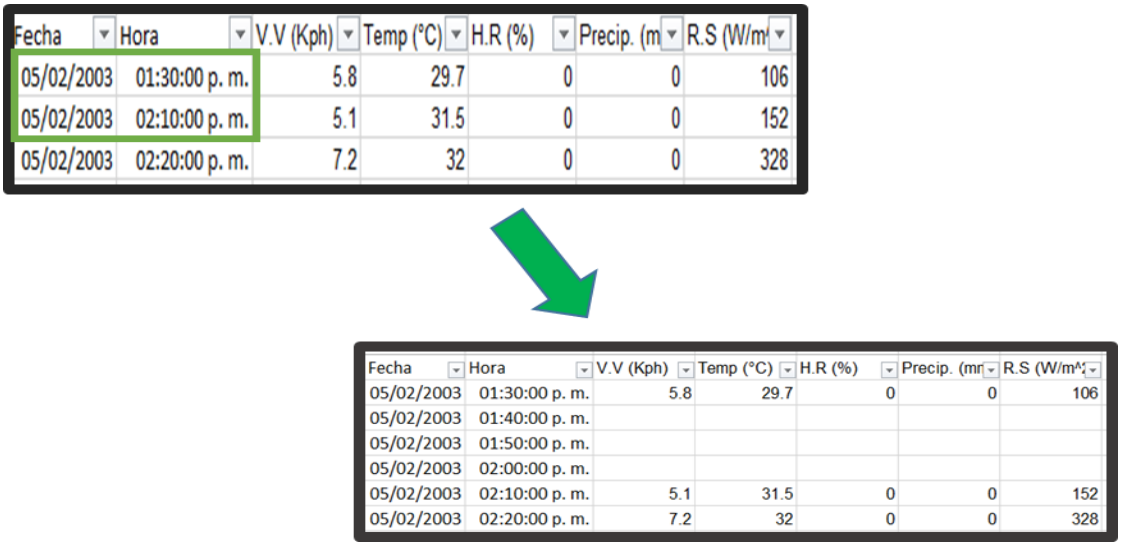


Figura 22. Series consecutivas completadas (sin datos meteorológicos).

Una vez que se identificaron los saltos de tiempo y se añadieron las filas correspondientes, se procedió a verificar si realmente se contaba con las 144 filas (registros por día), toda vez que los datos están registrados en intervalos de cada 10 minutos. La figura 23 muestra una tabla con filas incompletas debido a los saltos de tiempo, y la figura 24 muestra el mismo fragmento de la base de datos una vez que se ha aplicado el código macro 1 (sección de anexo), donde se observa una base de datos ya completa.

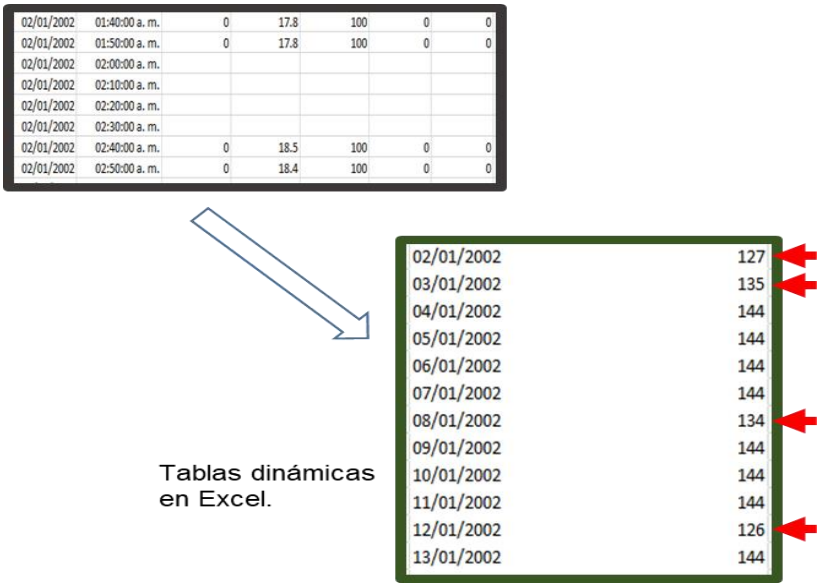


Figura 23. Series incompletas

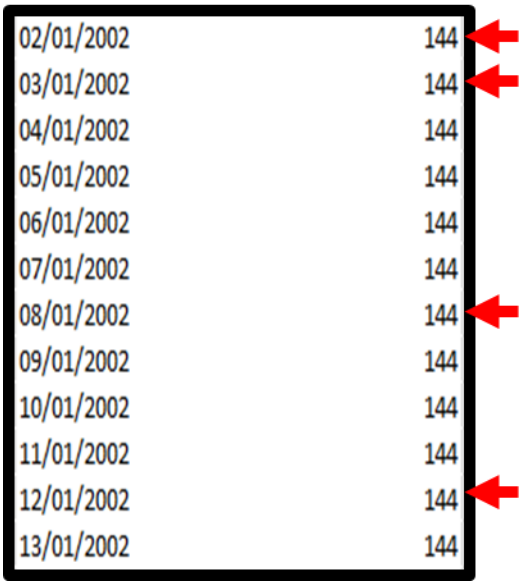
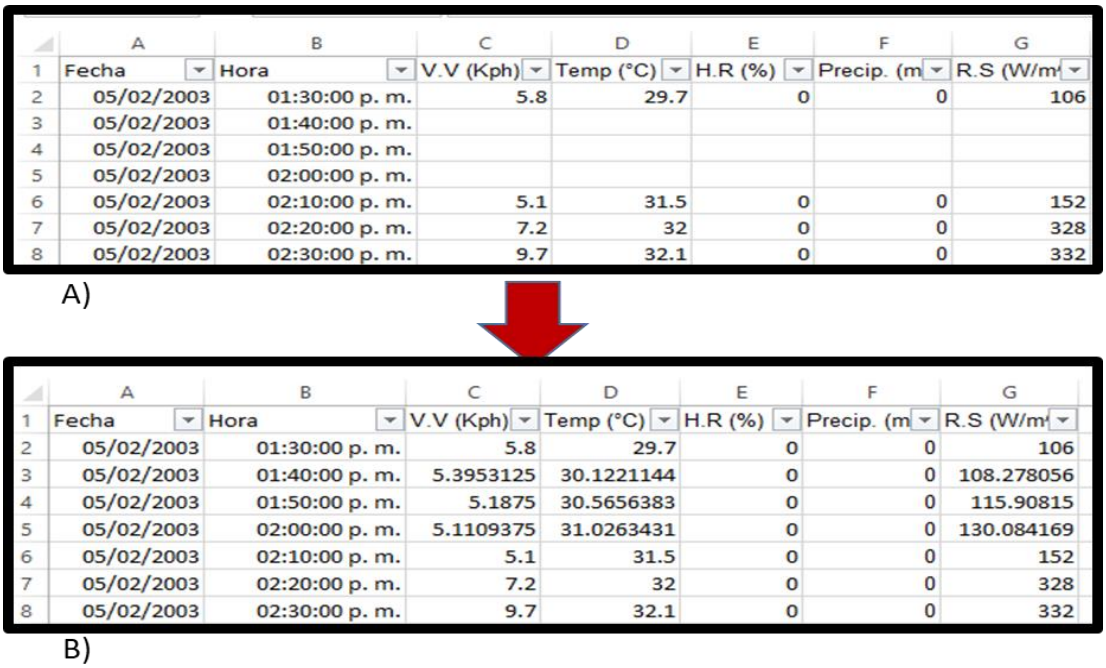


Figura 24. Series completas.

8.1.2 Evaluación de métodos de relleno de datos meteorológicos

El relleno de datos meteorológicos faltantes se realizó mediante las técnicas de interpolación PCHIP y SPLINE las cuales fueron implementadas usando el software MATLAB.

Para el rellenado de datos, se utilizaron los códigos 2 y 3 (sección de anexos). La figura 25B muestra una base de datos después de aplicar las técnicas de interpolación, nótese los valores interpolados desde la 01:40 p.m a 02:00 p.m.



	A	B	C	D	E	F	G
1	Fecha	Hora	V.V (Kph)	Temp (°C)	H.R (%)	Precip. (m	R.S (W/m
2	05/02/2003	01:30:00 p. m.	5.8	29.7	0	0	106
3	05/02/2003	01:40:00 p. m.					
4	05/02/2003	01:50:00 p. m.					
5	05/02/2003	02:00:00 p. m.					
6	05/02/2003	02:10:00 p. m.	5.1	31.5	0	0	152
7	05/02/2003	02:20:00 p. m.	7.2	32	0	0	328
8	05/02/2003	02:30:00 p. m.	9.7	32.1	0	0	332

A)

	A	B	C	D	E	F	G
1	Fecha	Hora	V.V (Kph)	Temp (°C)	H.R (%)	Precip. (m	R.S (W/m
2	05/02/2003	01:30:00 p. m.	5.8	29.7	0	0	106
3	05/02/2003	01:40:00 p. m.	5.3953125	30.1221144	0	0	108.278056
4	05/02/2003	01:50:00 p. m.	5.1875	30.5656383	0	0	115.90815
5	05/02/2003	02:00:00 p. m.	5.1109375	31.0263431	0	0	130.084169
6	05/02/2003	02:10:00 p. m.	5.1	31.5	0	0	152
7	05/02/2003	02:20:00 p. m.	7.2	32	0	0	328
8	05/02/2003	02:30:00 p. m.	9.7	32.1	0	0	332

B)

Figura 25. Llenado de filas vacías mediante técnicas de interpolación.

Los ajustes de las diferentes técnicas de interpolación fueron evaluados utilizando el coeficiente de determinación R^2, como se muestra las figuras 26, 27, 28 y 29. Se eligieron series consecutivas al azar de entre 20 y 25 días en diferentes años, y se eliminaron intencionalmente series consecutivas de entre 10 y 15 días para después aplicar los métodos de interpolación.

Pchip	Spline	Observado
114.800613	115.159315	117
146.684148	147.88366	153
178.725609	180.166175	184
559.278073	558.704805	568
586.178033	586.357865	594
613.338956	614.558521	619
639.399919	640.906119	641
655.525968	735.65633	609
615.697903	713.027261	625
566.906854	641.370026	582
523.543871	557.941861	535
107.813685	80.3353381	67
98.0524549	74.3529858	48
89.7649977	81.6941407	107

pchip	spline	obser
279.972028	280.405598	280
316.395299	316.919304	317
352.370921	352.723359	352
826.056	728.799371	774
711.728	475.888432	758
575.272	249.477807	461
460.944	183.778122	658
575.625	792.524695	619
401	625.838555	509
226.375	346.733138	235
106.093775	102.940336	93
71.6519137	68.7555964	69
41.8840953	40.6930589	45

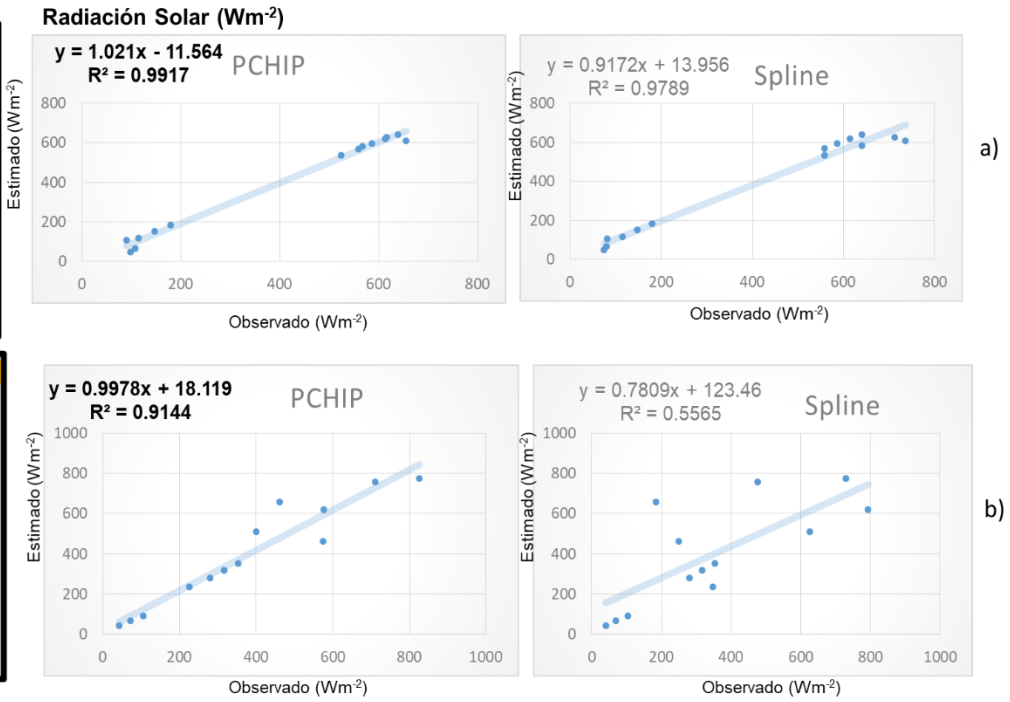


Figura 26. Interpolación de datos en radiación solar.

La Figura 26 muestra que para la variable radiación solar la técnica PCHIP supera a la técnica de SPLINE en las muestras evaluadas aleatoriamente. Para la muestra 26a se observa un R^2 de 0.991 para PCHIP y una R^2 de 0.978 para SLPINE. En la muestra 26b, se observa que PCHIP presenta un R^2 de 0.914, mientras que SPLINE obtuvo un R^2 de 0.556. Lo anterior demuestra que la técnica PCHIP es mejor para estimar valores de radiación solar.

En la Figura 27a y b, se muestran los R^2 de PCHIP y SPLINE de la variable velocidad del viento. En ambas técnicas el R^2 es muy bajo, estos se deben principalmente a que se usaron periodos de tiempos muy cortos. Sin embargo, PCHIP presenta en ambas muestras (a y b) los R^2 más altos. Razón por la cual se utilizó esta técnica para interpolar datos faltantes de velocidad de viento

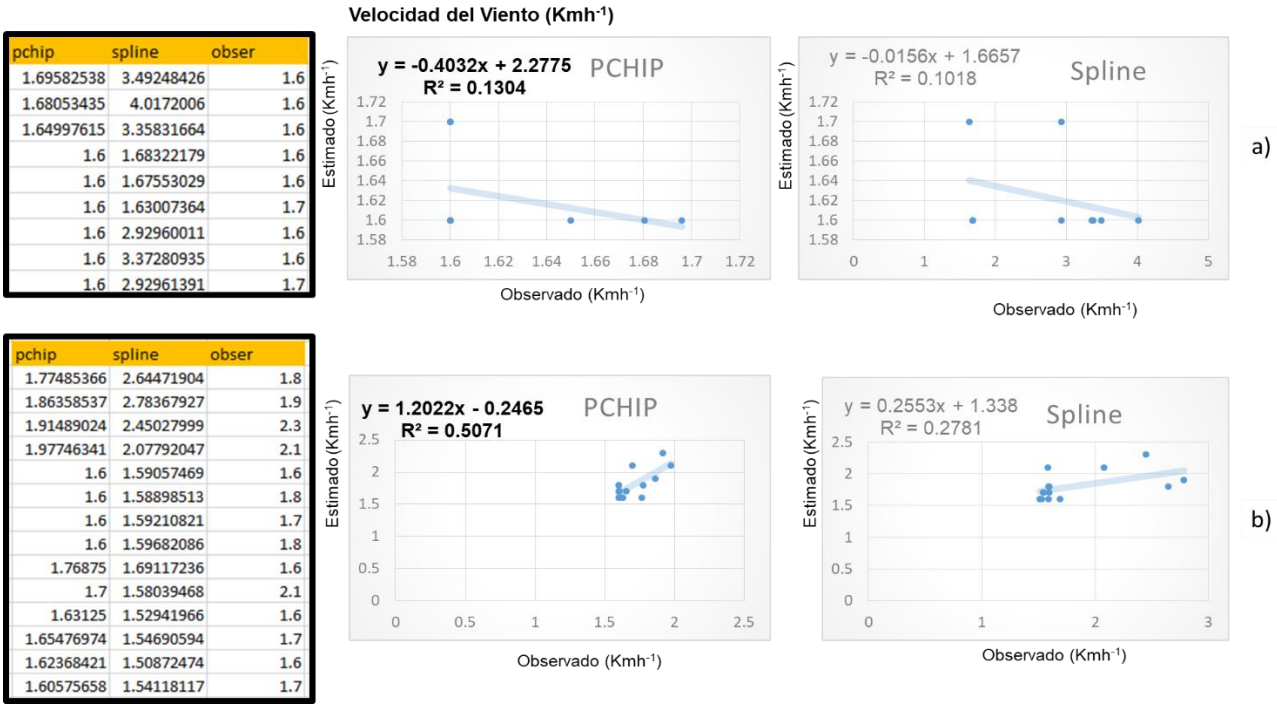
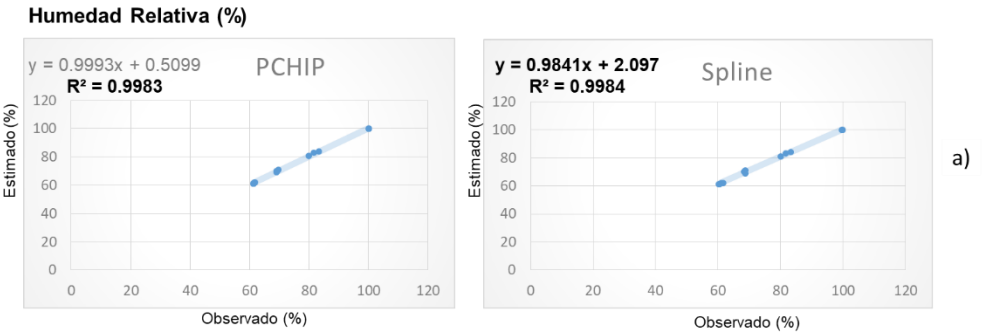


Figura 27. Interpolación de datos en velocidad de viento

Para las variables de humedad relativa (Figura 28a y b) y temperatura (Figura 29a a y b) no se presentó una diferencia significativa de acuerdo al estadístico R^2 entre PCHIP y SPLINE. Los R^2 fueron superiores al 0.971 los cuales son muy aceptables para estimar este tipo de datos. Lo anterior indica que ambas técnicas de interpolación de datos pueden ser utilizadas para estimar valores de humedad relativa y temperatura.

pchip	spline	obser
100	99.8578772	100
100	99.6845518	100
100	99.6689504	100
61.896	61.6545994	62
61.648	60.9841058	62
61.352	60.3863126	61
61.104	60.2590129	61
69.0328889	68.7677628	69
69.1386667	68.4205898	70
69.328	68.2895354	70
69.6115556	68.705654	71
79.78125	79.9929772	81
81.5	81.6292592	83
83.21875	83.2009115	84



pchip	spline	obser
99.4225543	99.3536478	100
98.0434783	98.1137931	99
96.392663	96.5670418	98
75.3877895	74.902211	74
74.8412632	73.4433054	73
74.4008421	72.3332942	74
74.1069474	72.2821887	75
71	71.2377707	67
71	70.4446865	66
71	69.9292591	68
70.15625	69.9200372	71
70.5	70.2528423	71
70.84375	70.7092263	72

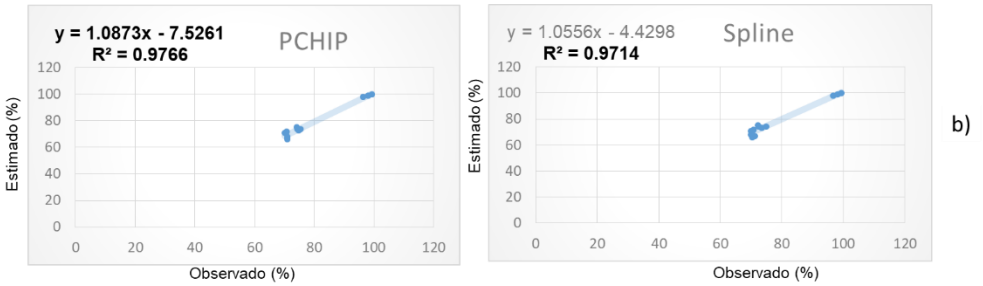
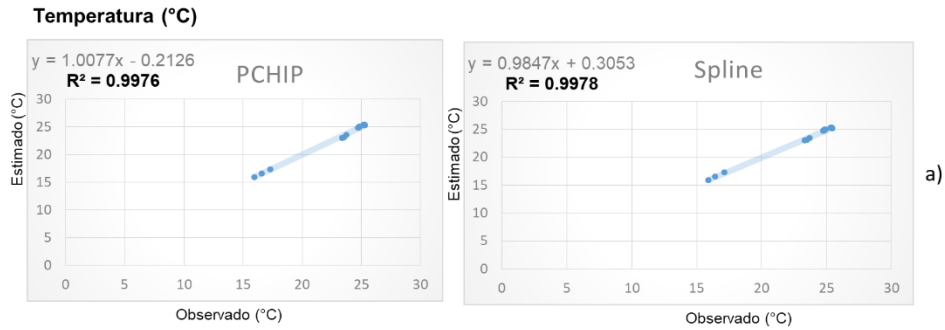


Figura 28. Interpolación de datos en humedad relativa

pchip	spline	obse
15.9598024	15.8748099	15.9
16.5604536	16.4063312	16.5
17.2808779	17.1346868	17.3
24.7208	24.7131372	24.8
24.7704	24.7741548	24.8
24.8296	24.8486037	25
24.8792	24.9020351	25
25.2917217	25.3891402	25.3
25.2671652	25.4228113	25.2
25.2267478	25.3919123	25.3
25.170887	25.2873422	25.3
23.7068354	23.6819308	23.5
23.5157502	23.4841034	23.1
23.34179	23.3192243	23



pchip	spline	obser
24.575	24.5855388	24.5
25	25.0061874	24.9
25.425	25.4237423	25.4
30.4272	30.5440686	30.6
30.2536	30.3066912	30.4
30.0464	29.9772794	30.1
29.8728	29.7452451	30.1
30.8722654	31.0391824	31.6
30.7217306	30.9497466	31.9
30.5353306	30.7104375	31.2
30.1582589	30.2213349	30.1
29.9803571	30.0705063	29.9
29.7622768	29.8344246	29.7

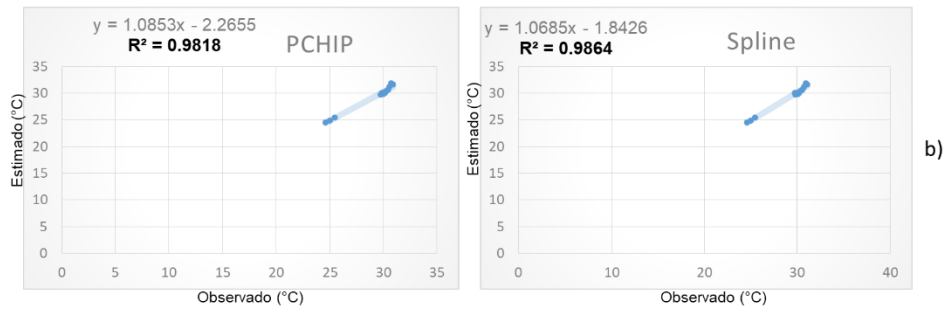
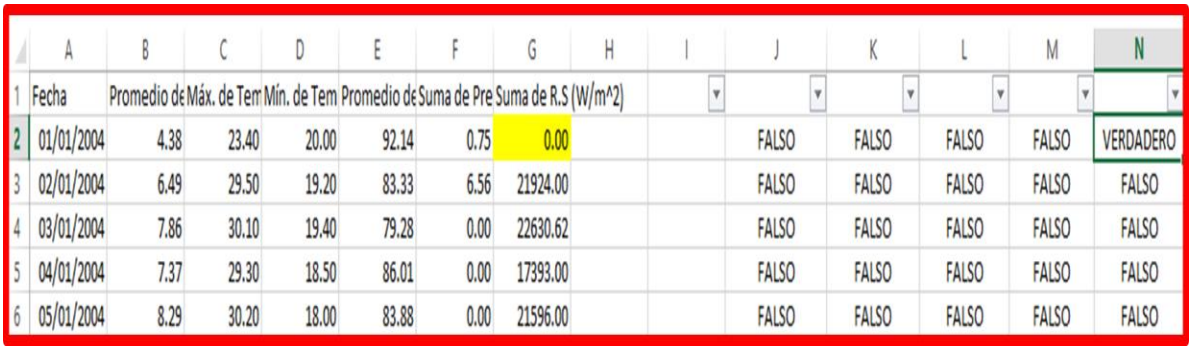


Figura 29. Interpolación de datos en temperatura

8.1.3 Detección de valores atípicos

La detección de valores atípicos en las bases de datos se llevó a cabo utilizando las técnicas de Grubbs, Mean y Cuartiles utilizando el software MATLAB, y en cada una de las estaciones meteorológicas en estudio. Los códigos 4, 5 y 6 (sección de anexos) se utilizaron para tal propósito.

A manera de ejemplo, la figura 30 muestra el resultado obtenido con la técnica cuartiles donde se observa en color amarillo un valor atípico señalado como verdadero.

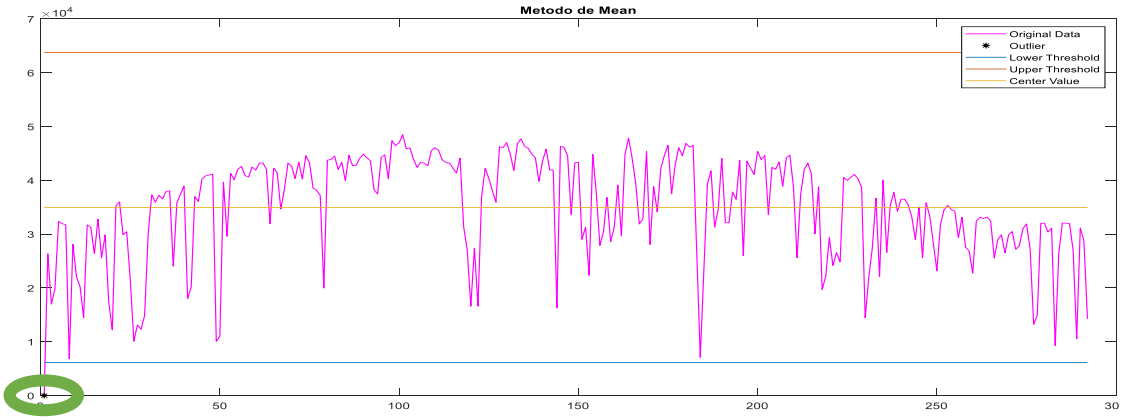


	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Fecha	Promedio de	Máx. de Tem	Mín. de Tem	Promedio de	Suma de Pre	Suma de R.S (W/m^2)							
2	01/01/2004	4.38	23.40	20.00	92.14	0.75	0.00			FALSO	FALSO	FALSO	FALSO	VERDADERO
3	02/01/2004	6.49	29.50	19.20	83.33	6.56	21924.00			FALSO	FALSO	FALSO	FALSO	FALSO
4	03/01/2004	7.86	30.10	19.40	79.28	0.00	22630.62			FALSO	FALSO	FALSO	FALSO	FALSO
5	04/01/2004	7.37	29.30	18.50	86.01	0.00	17393.00			FALSO	FALSO	FALSO	FALSO	FALSO
6	05/01/2004	8.29	30.20	18.00	83.88	0.00	21596.00			FALSO	FALSO	FALSO	FALSO	FALSO

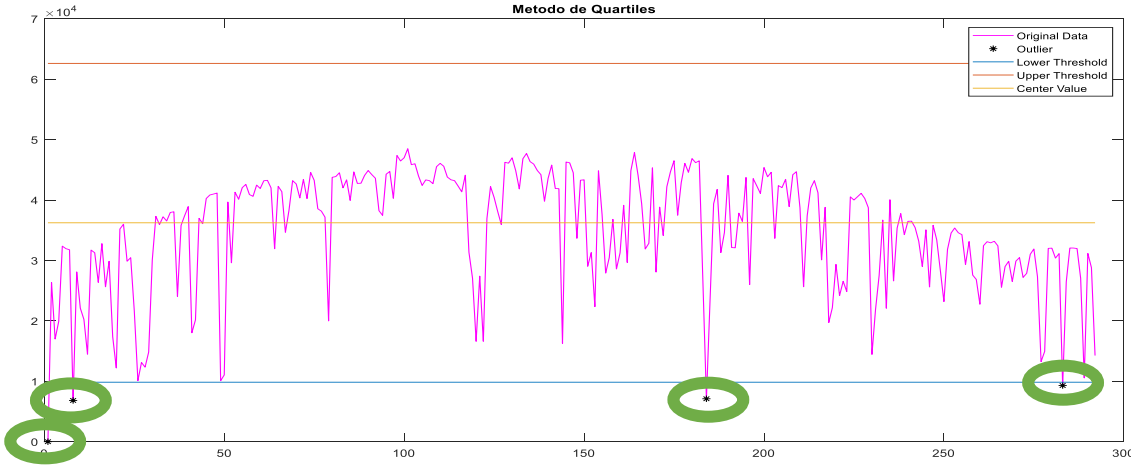
Figura 30. Detección de datos atípicos.

Un análisis comparativo entre las técnicas evaluadas muestra que Mean obtuvo el mejor resultado, ya que presentó mayor sensibilidad al momento de detectar valores atípicos, tomando en cuenta que existían valores “sospechosos” provocados por eventos de lluvia y/o nubosidad, los cuales al final no se consideraron como tales. Por otra parte, la técnica de Cuartiles expresaba mayor cantidad de valores atípicos aun cuando no estaban asociados a eventos de lluvia y/o nubosidad. Por otra parte, el método de Grubbs, en la mayoría de los casos, no presentó detección de valores atípicos. La figura 24 muestra a manera de ejemplo, el comportamiento de cada una de las técnicas, donde en un círculo verde se señalan los valores atípicos detectados.

A)



B)



C)

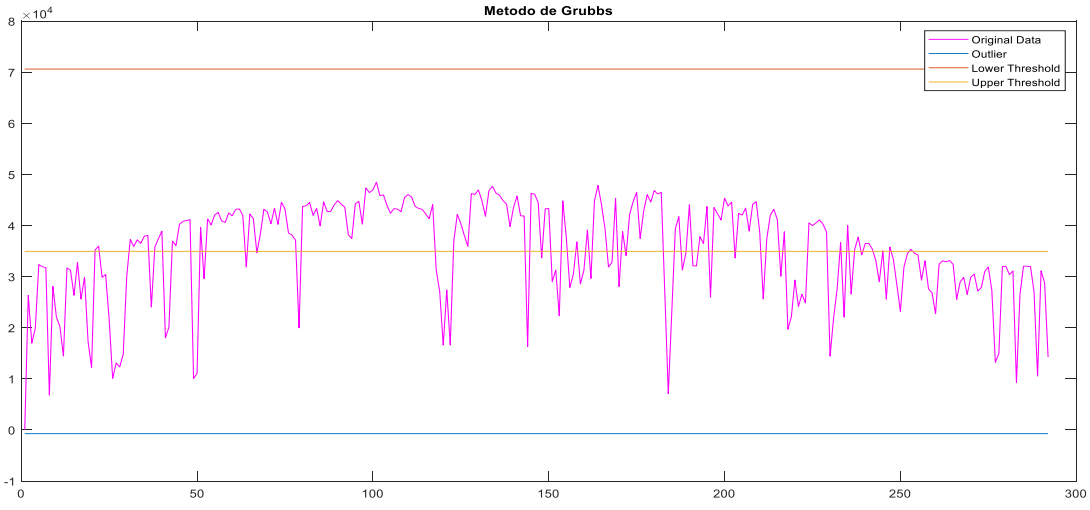
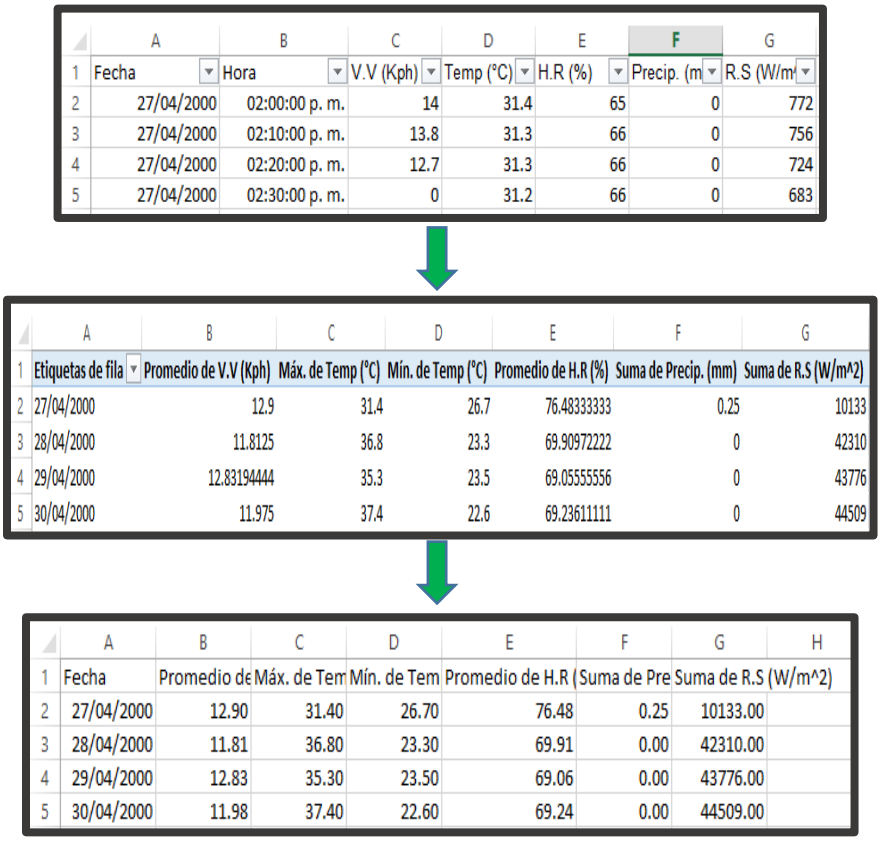


Figura 31. Representación gráfica de la detección de datos atípicos. A) Mean, B) Quartiles, C) Grubbs, en una misma serie de datos.

8.2 Evaluación de los modelos empíricos y de soft-computing en la estimación de la ETo.

Para calcular los valores diarios de ETo mediante la fórmula de la FAO-56 PM, se utilizó una plantilla diseñada en Microsoft Excel. Las bases de datos de las estaciones meteorológicas deberán presentar un formato adecuado con la información necesaria para el funcionamiento de la fórmula. Dentro de las bases de datos por estación, se insertaron tablas dinámicas, las cuales nos proporcionaron el promedio de la velocidad del viento, la temperatura máxima y mínima, el porcentaje de humedad relativa, y la suma de valores de la precipitación y radiación por día (Figura 32).



	A	B	C	D	E	F	G
1	Fecha	Hora	V.V (Kph)	Temp (°C)	H.R (%)	Precip. (m	R.S (W/m
2	27/04/2000	02:00:00 p. m.	14	31.4	65	0	772
3	27/04/2000	02:10:00 p. m.	13.8	31.3	66	0	756
4	27/04/2000	02:20:00 p. m.	12.7	31.3	66	0	724
5	27/04/2000	02:30:00 p. m.	0	31.2	66	0	683

	A	B	C	D	E	F	G
1	Etiquetas de fila	Promedio de V.V (Kph)	Máx. de Temp (°C)	Mín. de Temp (°C)	Promedio de H.R (%)	Suma de Precip. (mm)	Suma de R.S (W/m^2)
2	27/04/2000	12.9	31.4	26.7	76.48333333	0.25	10133
3	28/04/2000	11.8125	36.8	23.3	69.90972222	0	42310
4	29/04/2000	12.83194444	35.3	23.5	69.05555556	0	43776
5	30/04/2000	11.975	37.4	22.6	69.23611111	0	44509

	A	B	C	D	E	F	G	H
1	Fecha	Promedio de	Máx. de Tem	Mín. de Tem	Promedio de H.R (	Suma de Pre	Suma de R.S (W/m^2)	
2	27/04/2000	12.90	31.40	26.70	76.48	0.25	10133.00	
3	28/04/2000	11.81	36.80	23.30	69.91	0.00	42310.00	
4	29/04/2000	12.83	35.30	23.50	69.06	0.00	43776.00	
5	30/04/2000	11.98	37.40	22.60	69.24	0.00	44509.00	

Figura 32. Adecuación de los valores de entrada en la fórmula de la FAO-56 PM.

El Cuadro 4 muestra los índices estadísticos resultado del análisis comparativo de las ecuaciones empíricas (HS y Camargo), de soft-computing (SVM, GEP y XGBoost)

evaluadas sobre valores proporcionados por la ecuación de la FAO56-PM para estimar la ETo.

Cuadro 4. Índices estadísticos (MAE, RMSE y R2) de los modelos empíricos y de soft-computing usados para la estimación de la ETo de cada estación meteorológica.

Estación/modelo	R^2	MAE	RMSE	K_{HS}	K_{CA1}	K_{CA2}
Calakmul						
HS	0.670	0.569	0.729	0.0015		
Camargo	0.706	0.534	0.688		36.071	0.200
SVM	0.740	0.487	0.646			
GEP	0.695	0.544	0.719			
XGBoost	0.740	0.492	0.648			
Campeche						
HS	0.702	0.550	0.709	0.0020		
Camargo	0.623	0.623	0.797		40.256	0.218
SVM	0.730	0.519	0.680			
GEP	0.695	0.561	0.727			
XGBoost	0.695	0.543	0.721			
Cd del Carmen						
HS	0.694	0.633	0.821	0.0027		
Camargo	0.701	0.627	0.811		44.476	0.240
SVM	0.741	0.585	0.779			
GEP	0.720	0.638	0.810			
XGBoost	0.703	0.611	0.803			
Escárcega						
HS	0.692	0.654	0.825	0.0018		
Camargo	0.775	0.553	0.705		38.581	0.211
SVM	0.828	0.471	0.608			
GEP	0.773	0.561	0.714			
XGBoost	0.814	0.500	0.642			
Monclova						
HS	0.787	0.523	0.651	0.0020		
Camargo	0.815	0.488	0.608		39.391	0.214
SVM	0.853	0.426	0.531			
GEP	0.816	0.500	0.618			
XGBoost	0.842	0.442	0.569			
Los Petenes						
HS	0.673	0.599	0.767	0.0014		
Camargo	0.715	0.555	0.716		34.922	0.195
SVM	0.824	0.392	0.578			
GEP	0.811	0.406	0.576			
XGBoost	0.804	0.414	0.586			
Todas						

HS	0.703	0.588	0.750
Camargo	0.723	0.563	0.721
SVM	0.786	0.480	0.637
GEP	0.752	0.535	0.694
XGBoost	0.766	0.500	0.662

* K_{HS}, K_{CA1}, K_{CA2} son los coeficientes empíricos ajustados de las ecuaciones de HS y Camargo, respectivamente.

* Los valores de los estadísticos R^2, MAE y RMSE de los modelos de soft-computing corresponden a la etapa de validación.

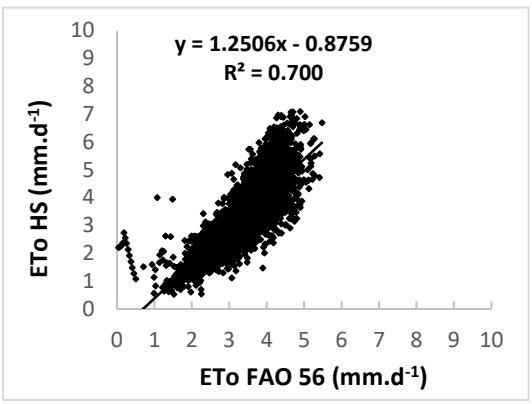
Como se observa en el Cuadro 4, los valores del coeficiente de calibración (K_{HS}) de la ecuación HS, señalan que el ajuste se realizó de manera correcta, ya que estos valores se encuentran cercanos al valor del coeficiente original (0.0023), con un rango de 0.0014 a 0.0027, siendo la estación de Los Petenes en la que se encontró el valor más alejado con 0.0014. Sin embargo, aun mostrando buenos resultados en el ajuste de su coeficiente empírico. De manera global, la ecuación de Camargo mostró mejores resultados (R^2 = 0.723; MAE = 0.563; RMSE = 0.721) sobre la ecuación de HS (R^2 = 0.703; MAE = 0.588; RMSE = 0.750) en la estimación de la ETo. Respecto a los métodos de soft-computing, de manera global, el modelo SVM obtuvo el mejor desempeño con relación a los otros modelos, obteniendo valores de R^2 = 0.786, MAE= 0.480 y RMSE= 0.637, siendo la estación de Monclova donde presento mejor desempeño (R^2 = 0.853; MAE= 0.426; RMSE= 0.531); seguido del modelo XGBoost con resultados de R^2 = 0.766, MAE= 0.500 y RMSE= 0.662, modelo que presento mejor desempeño en la estación de Monclova (R^2 = 0.842; MAE= 0.442; RMSE= 0.569). El modelo GEP fue el de menor rendimiento comparado con los otros modelos de soft-computing, obteniendo resultados de R^2 = 0.752, MAE= 0.535 y RMSE= 0.694; sin embargo, su rendimiento fue superior a lo obtenido por las ecuaciones empíricas. El modelo que obtuvo el menor desempeño fue la ecuación de HS.

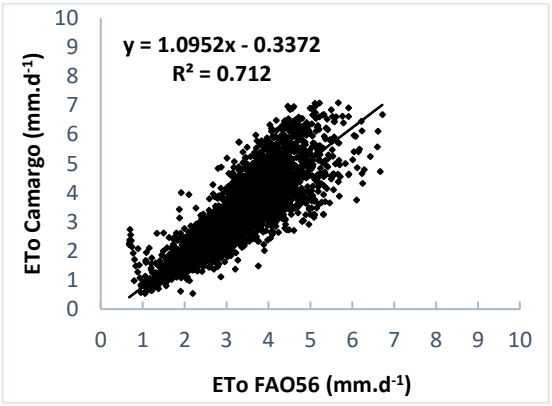
8.2.1 Evaluación de ecuaciones empíricas en la estimación de la ETo

En el cuadro 4 se observa, que entre las ecuaciones empíricas evaluadas, el modelo de Camargo mostró mejores resultados globales (R^2 = 0.723, MAE =0.563, RMSE = 0.721) respecto a la ecuación de HS (R^2 = 0.703, MAE =0.588, RMSE = 0.767), siendo

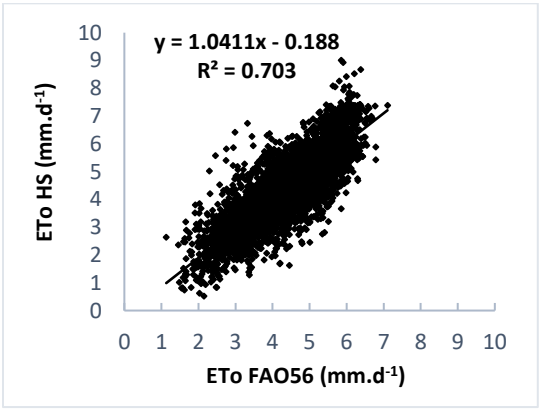
la estación de Monclova donde presentó mejor desempeño (R^2 = 0.815; MAE= 0.488; RMSE= 0.608). Respecto a la ecuación de Camargo el coeficiente calibrado K_{CA1} varió de 34.922 a 44.476, y el coeficiente K_{CA2} varió de 0.195 y 0.290. En la ecuación de HS el coeficiente K_{HS} calibrado varió 0.0015 a 0.0027.

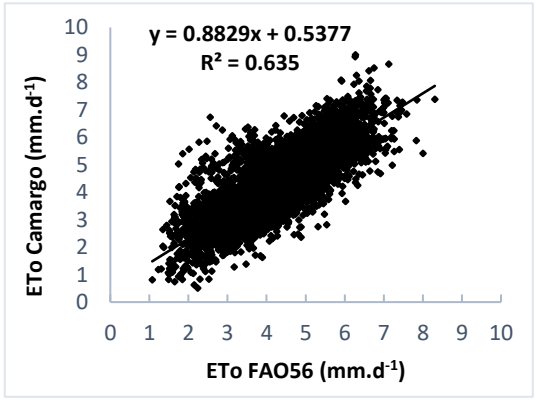
La figura 33 muestra los gráficos de comparación entre valores de ETo estimados por las ecuaciones empíricas y los calculados con la fórmula de la FAO 56 PM para validar los ajustes de las ecuaciones en cada estación meteorológica.



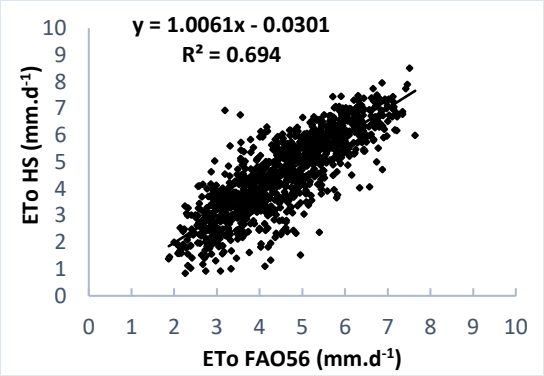


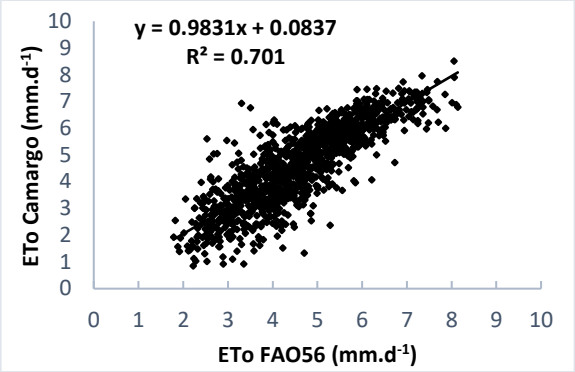
A)



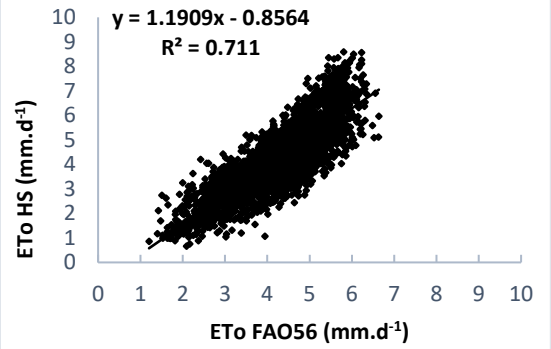


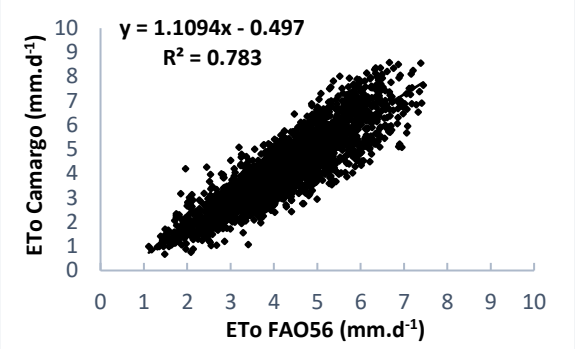
B)


y = 1.0061x - 0.0301
R² = 0.694
ETo HS (mm.d-1)
ETo FAO56 (mm.d-1)

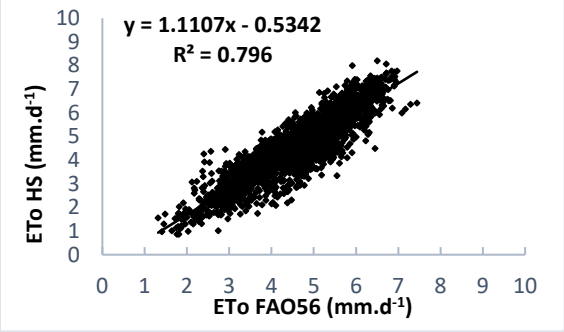

y = 0.9831x + 0.0837
R² = 0.701
ETo Camargo (mm.d-1)
ETo FAO56 (mm.d-1)

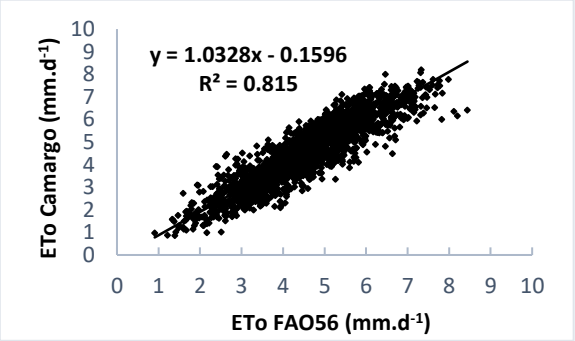
C)


y = 1.1909x - 0.8564
R² = 0.711
ETo HS (mm.d-1)
ETo FAO56 (mm.d-1)

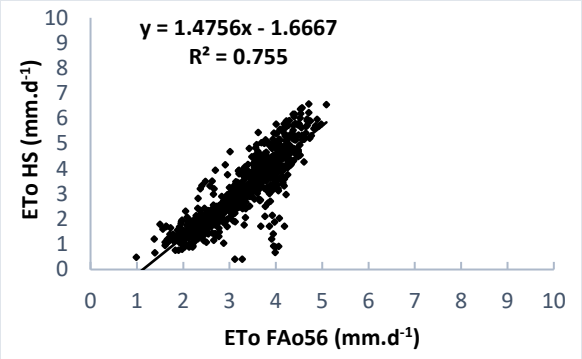

y = 1.1094x - 0.497
R² = 0.783
ETo Camargo (mm.d-1)
ETo FAO56 (mm.d-1)

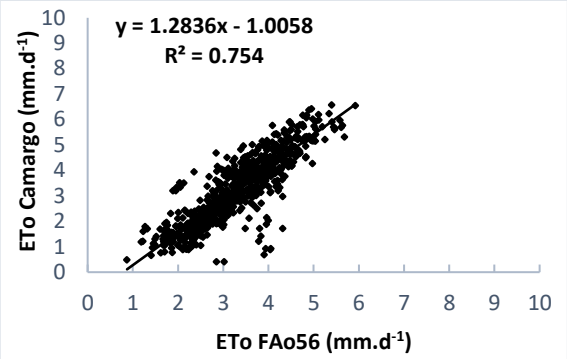
D)


y = 1.1107x - 0.5342
R² = 0.796
ETo HS (mm.d-1)
ETo FAO56 (mm.d-1)


y = 1.0328x - 0.1596
R² = 0.815
ETo Camargo (mm.d-1)
ETo FAO56 (mm.d-1)

E)





F)

Figura 33. Regresión lineal, valores estimados (Ecuaciones empíricas) vs valores observados (FAO 56 PM). Estación meteorológica A) Calakmul, B) Campeche, C) Cd del Carmen, D) Escárcega, E) Monclova, F) Los Petenes.

8.2.2 Evaluación de los métodos de inteligencia artificial (soft-computing) en la estimación de la ETo.

Respecto a los modelos de soft-computing para estimar la ETo, el cuadro 4 muestra, que el modelo SVM obtuvo el mejor desempeño con relación a los otros modelos evaluados según los estadísticos globales de R^2 = 0.786, MAE= 0.480 y RMSE= 0.637, siendo la estación de Monclova donde presentó mejor desempeño (R^2 = 0.853; MAE= 0.426; RMSE= 0.531); seguido del modelo XGBoost con resultados de R^2 = 0.766, MAE= 0.500 y RMSE= 0.662, donde presento mejor desempeño en la estación de Monclova (R^2 = 0.842; MAE= 0.442; RMSE= 0.569); el modelo GEP fue el de menor rendimiento comparado con los otros modelos de soft-computing, obteniendo resultados de R^2 = 0.752, MAE= 0.535 y RMSE= 0.694, sin embargo, su rendimiento fue superior a lo obtenido por las ecuaciones empíricas. En el cuadro 3 se observa un equilibrio entre los índices estadísticos obtenidos durante la etapa de entrenamiento y validación de los modelos de soft-computing, lo que indica que no hubo un sobre entrenamiento de los modelos.

Es importante presentar los resultados de los índices estadísticos obtenidos por los modelos basados en soft-computing durante la fase de entrenamiento y validación, ya que nos permiten observar si hubo algún caso de sobre entrenamiento que pudiera

afectar en las estimaciones. Para detectar un sobre entrenamiento todos los modelos se validan sobre una muestra de datos seleccionados aleatoriamente; de esta manera, los modelos SVM, GEP y XGBoost se evaluaron en función de su desempeño en las etapas de entrenamiento y validación.

Cuadro 5. Índices estadísticos (R^2, MAE y RMSE) de los modelos SVM, GEP y XGBoost durante la etapa de entrenamiento y validación, de cada estación meteorológica.

EMAs	Entrenamiento			Validación		
	MAE	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**
SVM						
Calakmul	0.460	0.611	0.769	0.487	0.646	0.740
Campeche	0.500	0.654	0.744	0.519	0.680	0.731
Cd del Carmen	0.576	0.743	0.742	0.585	0.779	0.742
Escárcega	0.475	0.609	0.838	0.471	0.608	0.828
Monclova	0.426	0.549	0.852	0.426	0.531	0.853
Los Petenes	0.375	0.573	0.802	0.392	0.578	0.824
Media	**0.469**	**0.623**	**0.791**	**0.480**	**0.637**	**0.786**
GEP						
Calakmul	0.544	0.710	0.708	0.544	0.719	0.695
Campeche	0.557	0.719	0.692	0.561	0.727	0.695
Cd del Carmen	0.633	0.824	0.683	0.638	0.810	0.720
Escárcega	0.509	0.659	0.802	0.561	0.714	0.773
Monclova	0.483	0.606	0.815	0.500	0.618	0.816
Los Petenes	0.471	0.654	0.766	0.406	0.576	0.811
Media	**0.533**	**0.695**	**0.744**	**0.535**	**0.694**	**0.752**
XGBoost						
Calakmul	0.492	0.607	0.771	0.492	0.648	0.740
Campeche	0.543	0.609	0.780	0.543	0.721	0.695
Cd del Carmen	0.465	0.592	0.845	0.611	0.803	0.703
Escárcega	0.426	0.547	0.864	0.500	0.642	0.815
Monclova	0.367	0.467	0.890	0.442	0.569	0.842
Los Petenes	0.343	0.468	0.879	0.414	0.586	0.804
Media	**0.439**	**0.548**	**0.838**	**0.500**	**0.662**	**0.767**

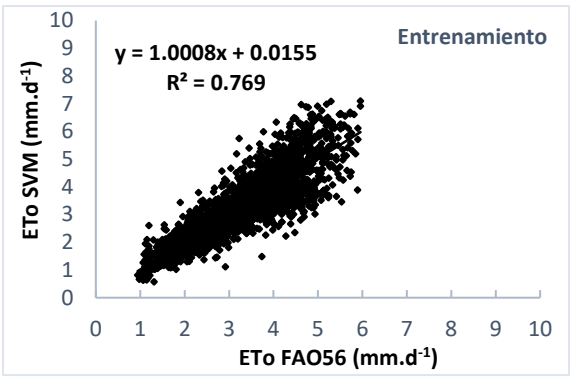
El sobreajuste de los datos se observa cuando los valores de los índices estadísticos en el entrenamiento de un modelo cualesquiera, está muy alejado de los índices obtenidos durante la validación. Los resultados obtenidos del entrenamiento y validación de los modelos, mostrados en el Cuadro 5, indican que no hubo sobreajuste en la estimación de la ETo, ya que tanto los valores del R^2 del entrenamiento como para los de la validación son relativamente similares, de igual forma para los valores del MAE y el RMSE, valores por debajo de lo obtenido con el R^2.

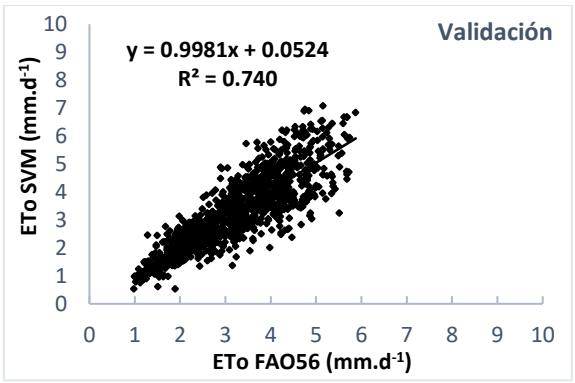
Respecto a los modelos SVM y XGBoost, con la finalidad de obtener modelos optimizados que estimen con precisión valores de ETo, es necesario realizar el ajuste de algunos parámetros internos (ya mencionados en el apartado 7.3.3 de esta tesis). En el caso de la SVM el ajuste se llevó a cabo mediante el algoritmo genético. El Cuadro 6 muestra los resultados del ajuste de los parámetros de los modelos SVM en cada estación meteorológica.

Cuadro 6. Parámetros ajustados del modelo SVM por estación meteorológica utilizando el algoritmo genético.

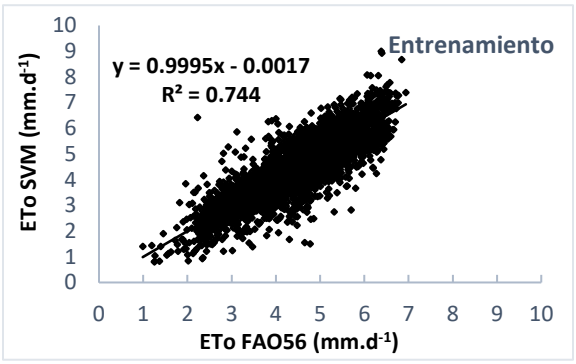
EMAs	**Parámetros ajustados**		
	Costo (C)	**Gamma (ɣ)**	**Épsilon (ϵ)**
Calakmul	1.471	0.334	0.147
Campeche	3.752	0.535	0.344
Cd del Carmen	3.547	0.147	0.410
Escárcega	5.995	0.285	0.229
Monclova	8.223	0.069	0.255
Los Petenes	7.837	0.269	0.147

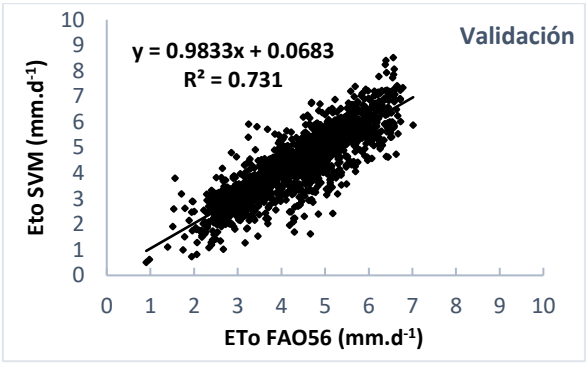
Los datos estimados por el modelo SVM usando el código (10), se compararon con los valores de ETo calculados con la fórmula FAO-56 PM mediante regresión lineal, realizada para cada estación en estudio. En la figura 34 se pueden observar los gráficos comparativos durante la etapa de entrenamiento y validación para cada estación.


Entrenamiento
y = 1.0008x + 0.0155
R² = 0.769
ETo SVM (mm.d-1)
ETo FAO56 (mm.d-1)

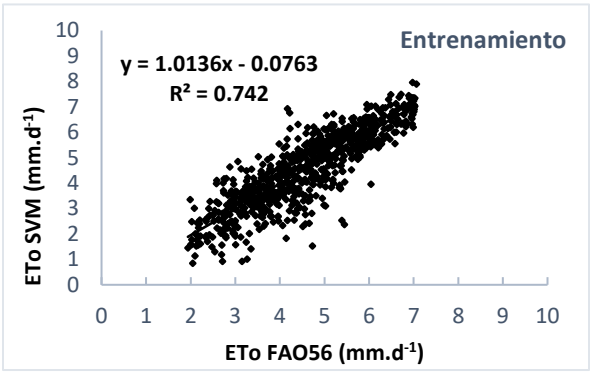

Validación
y = 0.9981x + 0.0524
R² = 0.740
ETo SVM (mm.d-1)
ETo FAO56 (mm.d-1)

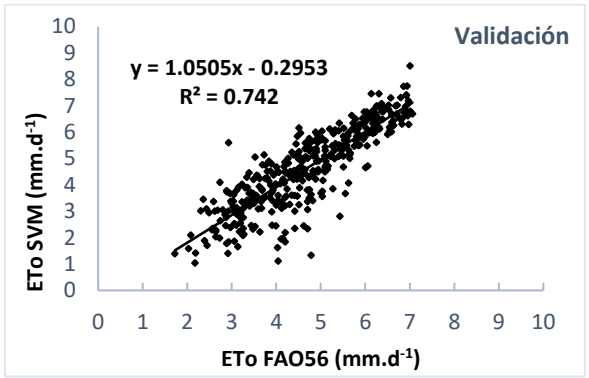
A)


Entrenamiento
y = 0.9995x - 0.0017
R² = 0.744
ETo SVM (mm.d-1)
ETo FAO56 (mm.d-1)

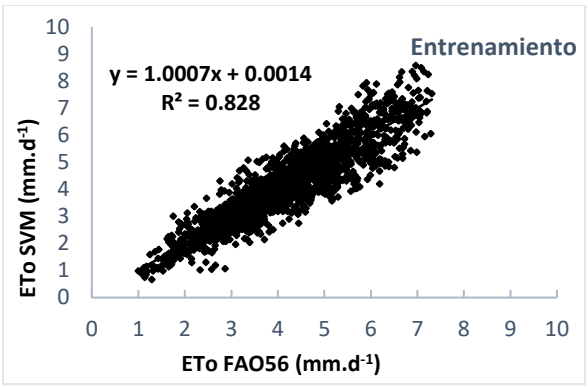

Validación
y = 0.9833x + 0.0683
R² = 0.731
Eto SVM (mm.d-1)
ETo FAO56 (mm.d-1)

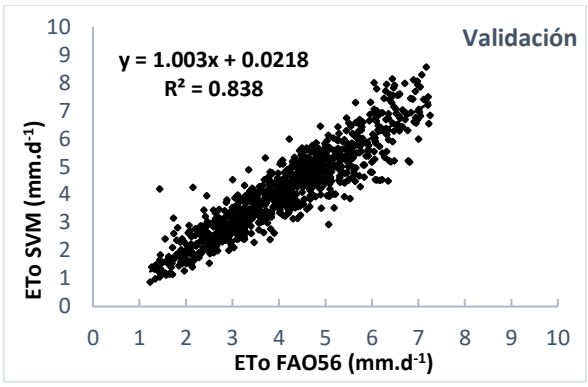
B)


Entrenamiento
y = 1.0136x - 0.0763
R² = 0.742
ETo SVM (mm.d-1)
ETo FAO56 (mm.d-1)

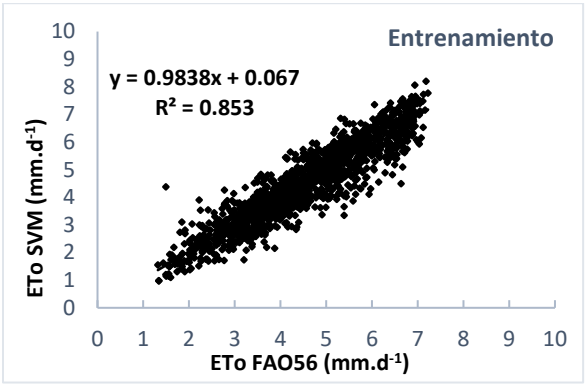

Validación
y = 1.0505x - 0.2953
R² = 0.742
ETo SVM (mm.d-1)
ETo FAO56 (mm.d-1)

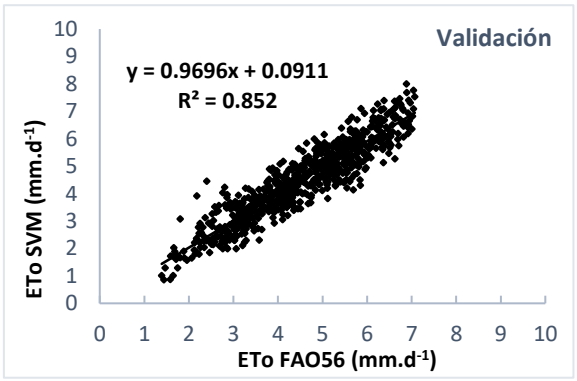
C)



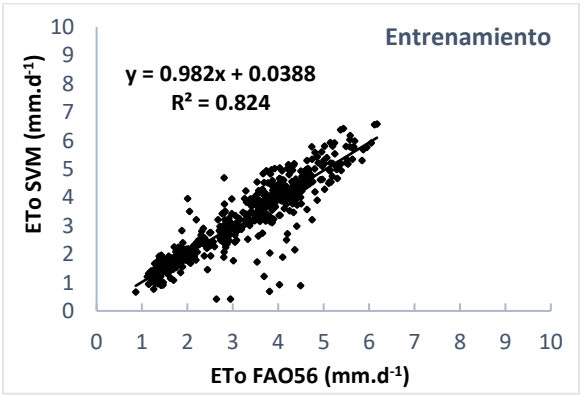


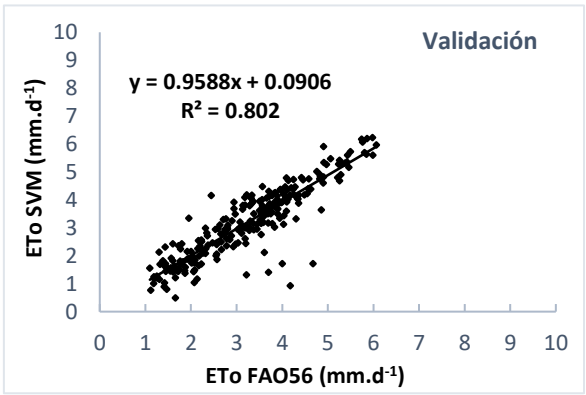
D)





E)

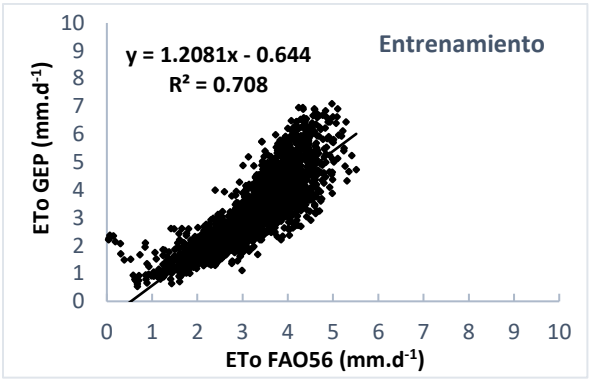


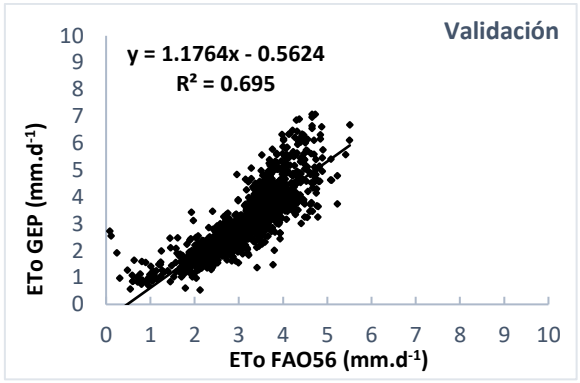


F)

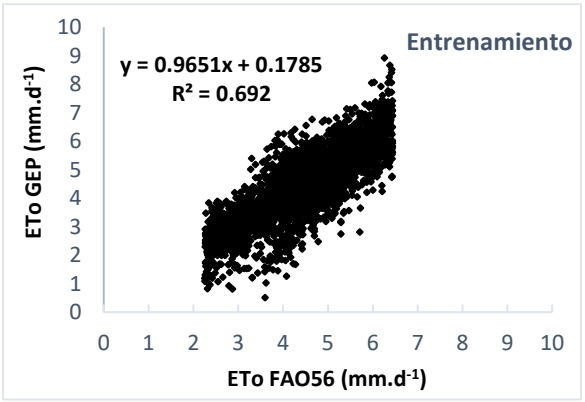
Figura 34. Gráficos de regresión lineal de valores estimados de ETo en etapa de entrenamiento y validación del modelo SVM vs valores calculados de ETo con la fórmula FAO56-PM, por estación meteorológica A) Calakmul, B) Campeche, C) Cd del Carmen, D) Escárcega, E) Monclova, F) Los Petenes.

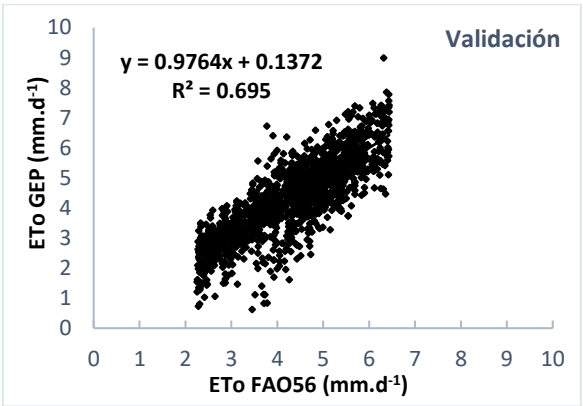
En el caso de los modelos GEP, la figura 35 muestra los gráficos de regresión lineal de los valores estimados sobre los calculados con la Fórmula de la FAO-56 PM durante la etapa de entrenamiento y validación para cada estación.



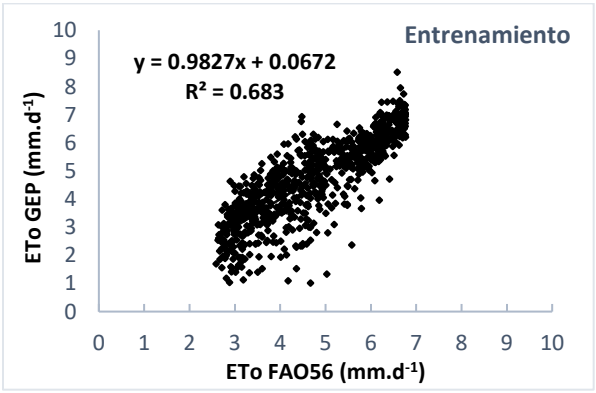


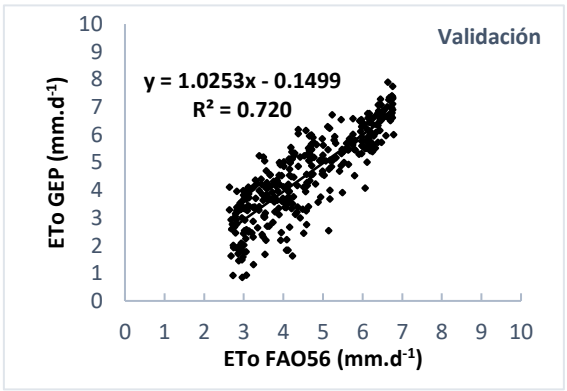
A)



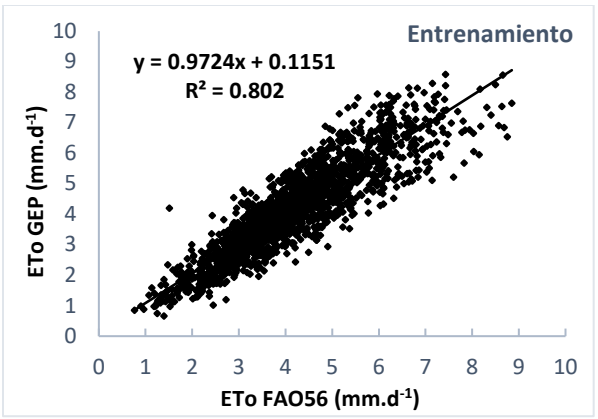


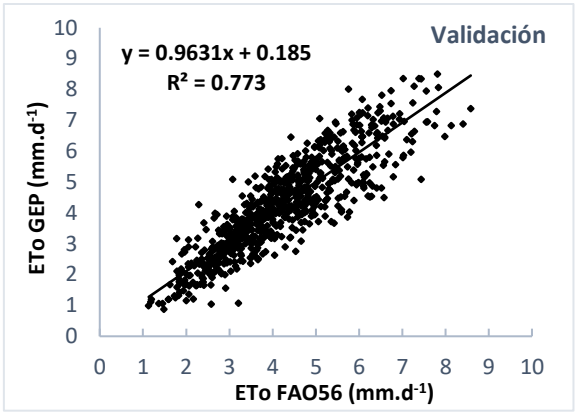
B)



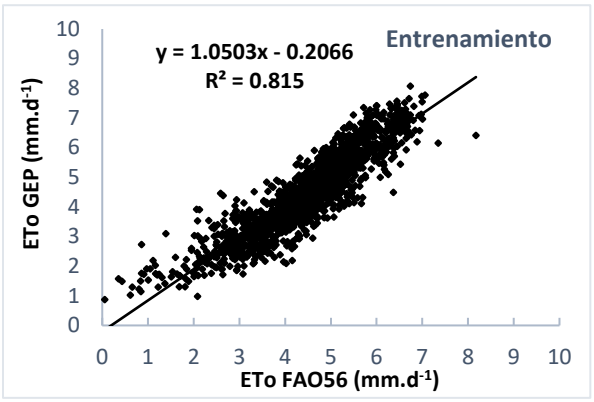


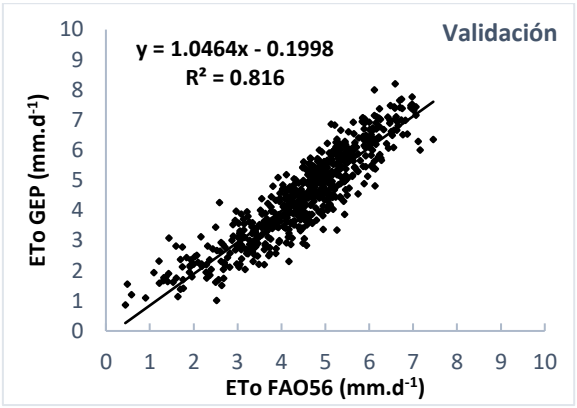
C)



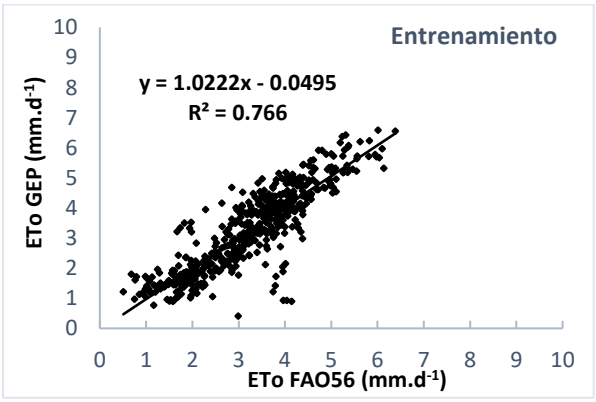


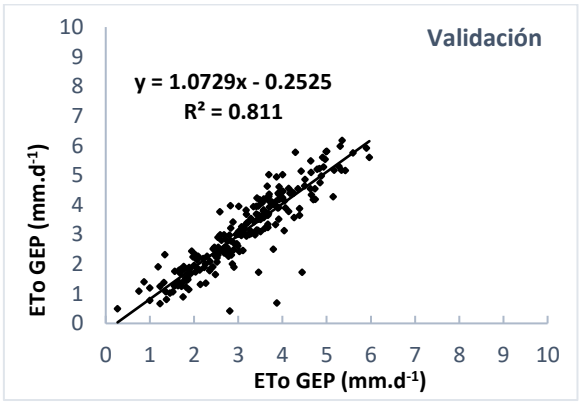
D)





E)





F)

Figura 35. Regresión lineal, entre los valores estimados de ETo en las etapas de entrenamiento y validación del modelo GEP vs valores calculados de ETo con la fórmula FAO56 por estación meteorológica: A) Calakmul, B) Campeche, C) Cd del Carmen, D) Escárcega, E) Monclova, F) Los Petenes.

Una de las características interesantes del modelo GEP es la generación de expresiones matemáticas a partir de las variables de entradas para estimar la ETo, que puede programarse en una hoja de cálculo, software R, Matlab o Python.

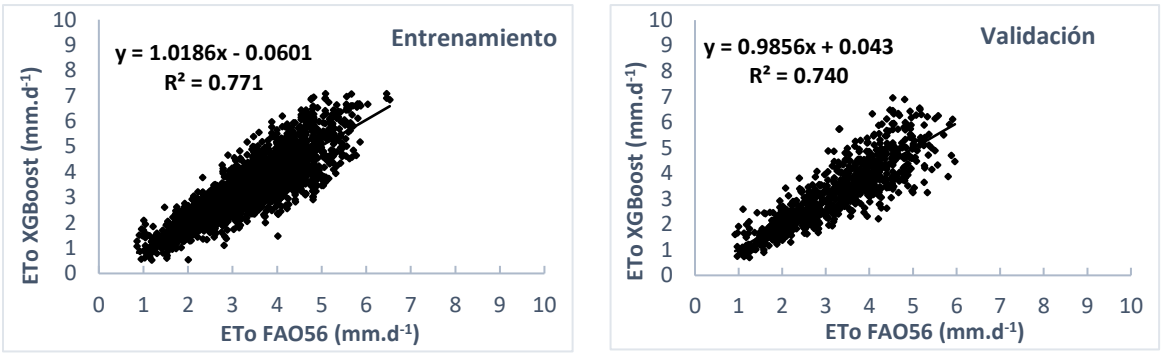
El Cuadro 7 presenta las expresiones algebraicas obtenidas por el modelo para cada estación meteorológica.

Cuadro 7. Expresiones matemáticas obtenidas por el modelo GEP para cada estación.

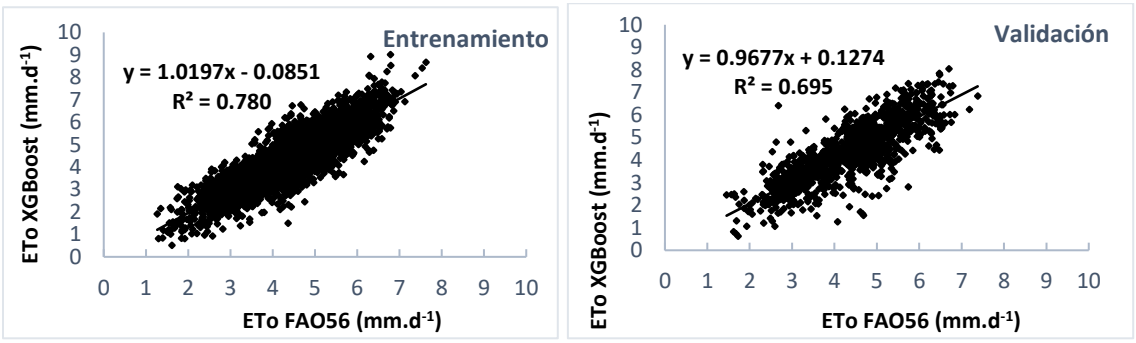
Estación	Expresión matemática
Calakmul	$ETo = \frac{T_{max}}{\sqrt[3]{(0.815 * T_{max}}} + \frac{\text{Arctan}(T_{max}) * (-9.223)}{\text{Arctan}(Ho - 5.126)} + \frac{Ho}{\text{Arctan}(T_{max} - T_{min} - \log(Ho - T_{min})}$
Campeche	$ETo = exp[\cos(\sqrt{T_{max}})^9] + exp\left[\sqrt{\cos\left(\sqrt{\frac{1.707}{T_{min}} + Ho}\right)^3}\right] + \cos\left[\frac{\left(\frac{1.707}{T_{min}}\right) + T_{max}}{\sqrt{(T_{max})^3}}\right]$
Carmen	$ETo = \sqrt{\left(\frac{Ho - (T_{max} + T_{min}) * sin(T_{max})}{Ho}\right)} + Arctan[(-3.589 * 8.418) + T_{max} - \sqrt{T_{min}} + log(T_{max} - 7.291)] + sin\left[\frac{(Ho * -3.589) - T_{min}}{T_{max} - \sqrt[3]{7.291 + T_{min}}}\right]$
Escárcega	$ETo = \log\left(\frac{7.347}{\sqrt[27]{T_{min}}}\right) + \left(\frac{Ho * \sqrt[3]{6.801}}{7.347 - 9.023 + T_{max}}\right) + \left(\frac{Ho * \sqrt{T_{max} - T_{min}}}{7.347^2 - T_{max}}\right)$
Monclova	$ETo = -13.981 + \sqrt[3]{2Ho} + \frac{T_{max}}{\sqrt[3]{\frac{Tmax}{T_{min}}}} + \sqrt[3]{\sqrt[6]{Ho} + \sqrt[3]{-3.443 + Ho}}$
Los Petenes	$ETo = \log(\log(\log(4.244 + T_{max}) + T_{max}^{27} + \left(\frac{\sqrt[3]{2Ho}}{\log\left(\frac{T_{max}}{T_{min}}\right)}\right) + \log\left(\frac{Ho}{4.244^3} * Ho\right)$

Dónde: ETo = Evapotranspiración de referencia; T_{max}= temperatura máxima; T_{min}= temperatura mínima; Ho = Radiación extraterrestre; Arctan= arcotangente; Cos= coseno; Sin= seno; Exp= Exponente; Log= logaritmo.

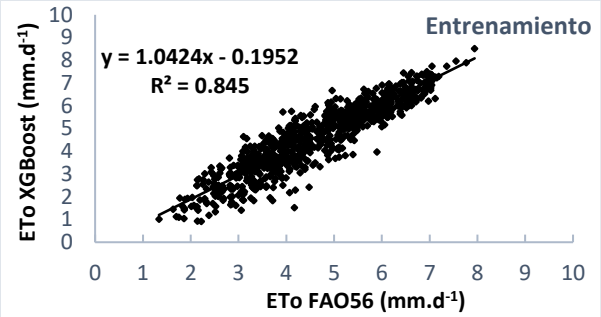
Respecto al modelo XGBoost, la figura 36 presenta la comparación de las estimaciones de ETo por el modelo XGBoost y la ETO calculada con la fórmula de la FAO56-PM, en las etapas de entrenamiento y validación para cada estación meteorológica.

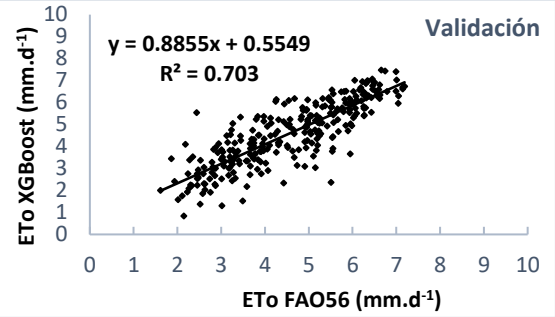


A)

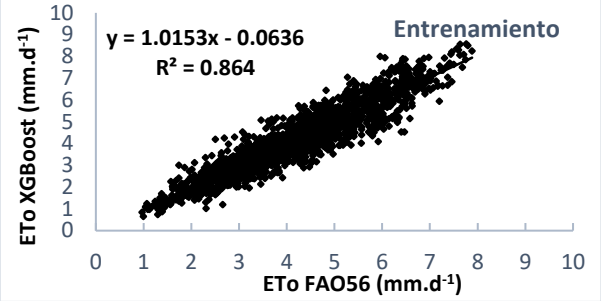


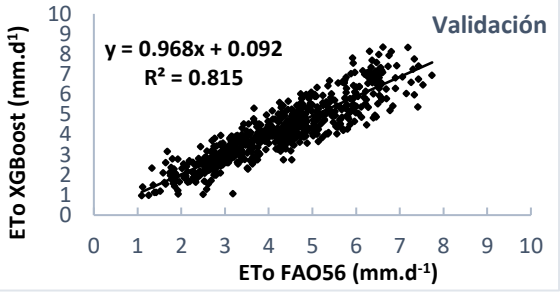
B)


Entrenamiento
y = 1.0424x - 0.1952
R² = 0.845
ETo XGBoost (mm.d-1)
ETo FAO56 (mm.d-1)

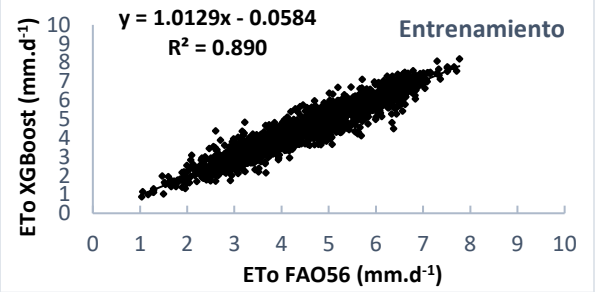

Validación
y = 0.8855x + 0.5549
R² = 0.703
ETo XGBoost (mm.d-1)
ETo FAO56 (mm.d-1)

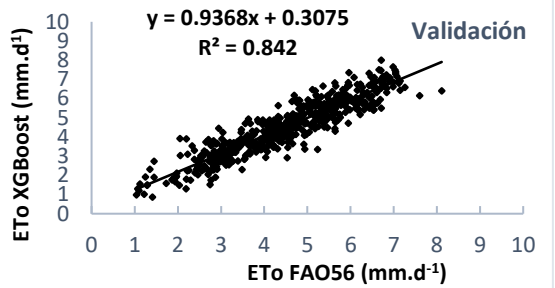
C)


Entrenamiento
y = 1.0153x - 0.0636
R² = 0.864
ETo XGBoost (mm.d-1)
ETo FAO56 (mm.d-1)

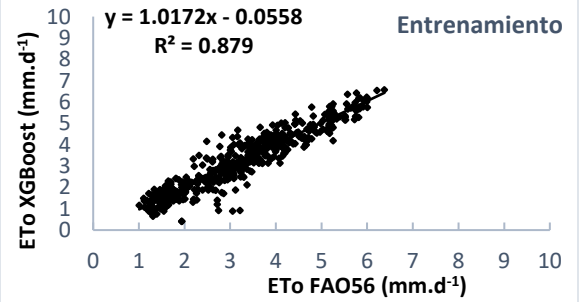

Validación
y = 0.968x + 0.092
R² = 0.815
ETo XGBoost (mm.d1)
ETo FAO56 (mm.d-1)

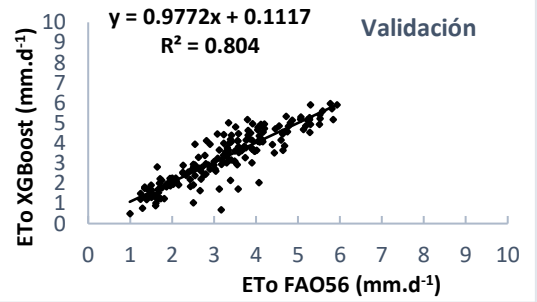
D)


Entrenamiento
y = 1.0129x - 0.0584
R² = 0.890
ETo XGBoost (mm.d-1)
ETo FAO56 (mm.d-1)


Validación
y = 0.9368x + 0.3075
R² = 0.842
ETo XGBoost (mm.d1)
ETo FAO56 (mm.d-1)

E)


Entrenamiento
y = 1.0172x - 0.0558
R² = 0.879
ETo XGBoost (mm.d-1)
ETo FAO56 (mm.d-1)


Validación
y = 0.9772x + 0.1117
R² = 0.804
ETo XGBoost (mm.d-1)
ETo FAO56 (mm.d-1)

F)

Figura 36. Regresión lineal entre los valores estimados de ETo en las etapas de entrenamiento y validación con el modelo XGBoost vs valores calculados de ETo con la fórmula de la FAO56 por estación meteorológica A) Calakmul, B) Campeche, C) Cd del Carmen, D) Escárcega, E) Monclova, F) Los Petenes.

Para climas áridos y súper húmedos donde existe una amplitud térmica mayor, Camargo *et al*., 1999 presentaron una modificación al método de Thornthwaite usando el término "temperatura efectiva" Tef = 0.36 (3 T_{max} – T_{min}), obteniendo excelentes resultados para regiones súper húmedas de Brasil. En el presente estudio, la ecuación empírica de Camargo obtuvo mejores estimaciones de ETo en comparación con la ecuación de HS, esta última comúnmente utilizada en la Península de Yucatán para estimar la ETo cuando solo existen datos de temperatura. Asimismo, la calibración del coeficiente K_{HS} de la ecuación de HS coincide con el obtenido por (Bautista *et al.*, 2009) para algunos sitios de la Península de Yucatán, donde el valor más alto del coeficiente K_{HS} = 0.0027 se observó en la estación de Carmen rodeada por aguas del golfo de México, y los valores más bajos se observaron en regiones rodeadas por abundante vegetación como en los casos de las reservas de la biosfera de los Petenes (K_{HS} = 0.0014) y Calakmul (K_{HS} = 0.0015).

Por otra parte, entre las técnicas de soft-computing evaluadas para estimar la ETo, el modelo SVM utilizando el kernel de base radial presento mejores resultados, GEP obtuvo el rendimiento más bajo de las tres técnicas en ambas etapas, esto coincide con los resultados obtenidos por (Mehdizadeh et al., 2017) en regiones áridas y semi-áridas de Irán donde implementaron la técnica GEP, dos modelos SVM de base radial y polinomial; y MARS (Spline de Regresión Adaptativa Multivariada) comparándolo con 16 ecuaciones empíricas basadas en transferencia de masa, radiación y parámetros meteorológicos; los resultados revelaron que, tanto MARS como SVM de base radial, obtuvieron mejores estimaciones que el resto de las técnicas de soft-computing y que las ecuaciones empíricas.

CAPÍTULO IX. CONCLUSIONES

En los últimos años, se ha incrementado el número de parcelas agrícolas que cuentan con sistemas de riego y pocas veces se le da la importancia adecuada al manejo de los recursos hidráulicos; es por esto que es necesario contar con herramientas que contribuyan en la planificación y manejo del agua de riego. La estimación de la evapotranspiración de referencia (ETo) juega un papel de gran importancia en muchos campos de estudio; en el manejo de recursos hidráulicos, diseño de sistemas de riego y/o en estudios hidrológicos y agrícolas. En la presente investigación, se evaluó el potencial de tres modelos de soft-computing y dos ecuaciones empíricas basadas en variables de temperatura y radiación solar extraterrestre para estimar la ETo en seis estaciones meteorológicas automatizadas ubicadas en el estado de Campeche. Se realizó un análisis de calidad a las bases de datos por cada estación meteorológica estudiada, proporcionada por la CONAGUA, donde en términos generales, se puede concluir que:

El estado de las estaciones meteorológicas es deficiente, encontrando muchas irregularidades en las bases de datos, por lo que algunas de las estaciones no fueron consideradas en el estudio. Otras estaciones presentaban años con series de tiempos incompletos, y para lo cual, se llevó a cabo el relleno de dichas series mediante técnicas de interpolación. La mejor técnica de interpolación para estimar valores faltantes en series históricas de datos meteorológicos es la técnica PCHIP.

En la detección de datos atípicos el método de Mean fue elegido como el mejor, ya que permite mayor tolerancia a valores atípicos causados por eventos de lluvias y/o nubosidades.

En la evaluación del desempeño de los modelos usados en este estudio para la estimación de la ETo se puede concluir que:

De las ecuaciones empíricas evaluadas basadas en temperatura, la ecuación propuesta por Camargo obtuvo mejor desempeño en la estimación de la ETo, por lo que se recomienda su uso para climas cálidos - subhúmedos como el caso de la Península de Yucatán. Sin embargo, hay que considerar que su implementación requiere un mayor

número de valores de entrada y cálculos, en comparación con la ecuación de HS. En ambos casos el estudio provee de coeficientes calibrados tanto para estaciones que se localizan en sitios cercanos al mar y en sitios tierra adentro.

Respecto a los modelos de soft-computing, el modelo SVM de base radial se recomienda para realizar estimaciones de ETo.

El ajuste previo de los parámetros de los modelos de soft-computing mediante algoritmos, es fundamental para evitar un sobre entrenamiento que afectaría a futuras estimaciones utilizando otras series de datos.

Por otra parte, es importante destacar que los modelos GEP también son una buena opción al momento de realizar estimaciones de la ETo, ya que el modelo algebraico proporcionado por la técnica se podría programar en una hoja de cálculo u otro software, y de este modo realizar predicciones, y como se comprobó en el presente estudio, el modelo GEP superó ligeramente a los modelos empíricos.

Los modelos de soft-computing son una excelente opción para estimar valores de ETo al superar a las ecuaciones empíricas, sin embargo, para su implementación se requiere de un conocimiento especializado en el uso de software y ejecución de códigos de programación.

CAPÍTULO X. LITERATURA CITADA

Ahmad, L., Habib Kanth, R., Parvaze, S., & Sheraz Mahdi, S. (2017). *Experimental Agrometeorology: A Practical Manual*. *Experimental Agrometeorology: A Practical Manual*. https://doi.org/10.1007/978-3-319-69185-5

Alberto, M. C. R., Quilty, J. R., Buresh, R. J., Wassmann, R., Haidar, S., Correa, T. Q., & Sandro, J. M. (2014). Actual evapotranspiration and dual crop coefficients for dry-seeded rice and hybrid maize grown with overhead sprinkler irrigation. *Agricultural Water Management*, *136*, 1–12. https://doi.org/10.1016/j.agwat.2014.01.005

Allen, R. G., Pereira, L. S., Raes, D., Smith, M., & Ab, W. (1998). Crop evapotranspiration - Guidelines for computing crop water requirements - FAO Irrigation and drainage paper 56 By, 1–15.

Almorox, J. (2016). Worldwide assessment of the Penman – Monteith temperature approach for the estimation of monthly reference evapotranspiration. *Theoretical and Applied Climatology*. https://doi.org/10.1007/s00704-016-1996-2

Antonanzas-Torres, F., Urraca, R., Antonanzas, J., Fernandez-Ceniceros, J., & Martinez-De-Pison, F. J. (2015). Generation of daily global solar irradiation with support vector machines for regression. *Energy Conversion and Management*, *96*, 277–286. https://doi.org/10.1016/j.enconman.2015.02.086

Ayyoub, A., Er-raki, S., Khabba, S., Merlin, O., Ezzahar, J., Rodriguez, J. C., & Bahlaoui, A. (2017). A simple and alternative approach based on reference evapotranspiration and leaf area index for estimating tree transpiration in semi-arid regions. *Agricultural Water Management*, *188*, 61–68. https://doi.org/10.1016/j.agwat.2017.04.005

Bautista, F., Bautista, D., Investigaciones, C. De, Nacional, U., & México, A. De. (2009). Calibration of the equations of Hargreaves and Thornthwaite to estimate the potential evapotranspiration in semi-arid and subhumid tropical climates for regional applications, *22*(4), 331–348.

Bruffaerts, C., Verardi, V., & Vermandele, C. (2014). A generalized boxplot for skewed and heavy-tailed distributions. *Statistics and Probability Letters*, *95*, 110–117. https://doi.org/10.1016/j.spl.2014.08.016

Čadro, S., Uzunović, M., Žurovec, J., & Žurovec, O. (2017). Validation and calibration of various reference evapotranspiration alternative methods under the climate conditions of Bosnia and Herzegovina. *International Soil and Water Conservation Research*, *5*(4), 309–324. https://doi.org/10.1016/j.iswcr.2017.07.002

CAMARGO, Â. P. DE, & CAMARGO, M. B. P. DE. (2000). Uma revisão analítica da evapotranspiração potencial. *Bragantia*, *59*(2), 125–137. https://doi.org/10.1590/s0006-87052000000200002

Cascone, S., Coma, J., Gagliano, A., & Pérez, G. (2019). The evapotranspiration process in green roofs: A review. *Building and Environment*, *147*, 337–355. https://doi.org/10.1016/j.buildenv.2018.10.024

Chang, C., Lin, C., & Tieleman, T. (2013). LIBSVM : A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *307*, 1–39. https://doi.org/10.1145/1961189.1961199

Chen, T., & Guestrin, C. (2016). XGBoost: AScalable Tre Boosting System, *19*(6). https://doi.org/10.1145/2939672.2939785

Chica Jiménez, J. A., & Bosch, M. (2018). Interpolación spline y aplicación a las curvas de nivel.

Cruz, M. H. (2013). Estimación de la evapotranspiración de referencia en regiones con datos climaticos limitados.

Duque Martínez, J. S. R. R. (2015). *Comparación estadística de métodos interpolación determinísticos y estocásticos para la generación de Modelos Digitales del Terreno a partir de datos LIDAR, en la parroquia de Tumbabiro, cantón de San Miguel de Urcuquí, provincia de Imbabura*.

Fallas, J. (2007). Modelos digitales de elevación: Teoría, métodos de interpolación y aplicaciones. *Mapealo. Com*, 83.

Fan, J., Wang, X., Wu, L., Zhou, H., Zhang, F., Yu, X., … Xiang, Y. (2018). Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China. *Energy Conversion and Management*, *164*(January), 102–111. https://doi.org/10.1016/j.enconman.2018.02.087

Fan, J., Yue, W., Wu, L., Zhang, F., Cai, H., Wang, X., … Xiang, Y. (2018). Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China. *Agricultural and Forest Meteorology*, *263*(August), 225–241. https://doi.org/10.1016/j.agrformet.2018.08.019

FAO. (2006). Evapotranspiración del cultivo.

Feng, Y., Cui, N., Zhao, L., Hu, X., & Gong, D. (2016). Comparison of ELM, GANN, WNN and empirical models for estimating reference evapotranspiration in humid region of Southwest China. *Journal of Hydrology*, *536*, 376–383. https://doi.org/10.1016/j.jhydrol.2016.02.053

Ferreira, C. (2001). Gene Expression algo, 1–22.

Garcia, F. (2012). Tests to identify outliers in data series. *Pontifical Catholic University of Rio de Janeiro, …*, 1–16. Retrieved from http://habcam.whoi.edu/HabCamData/HAB/processed/Outlier Methods_external.pdf

Gong, D., Feng, Y., Jia, Y., Cui, N., Li, C., & Zhao, L. (2016). Calibration of Hargreaves model for reference evapotranspiration estimation in Sichuan basin of southwest China. *Agricultural Water Management*, *181*, 1–9. https://doi.org/10.1016/j.agwat.2016.11.010

He, Z., Wen, X., Liu, H., & Du, J. (2014). A comparative study of artificial neural network,

adaptive neuro fuzzy inference system and support vector machine for forecasting river flow in the semiarid mountain region. *Journal of Hydrology*, *509*, 379–386. https://doi.org/10.1016/j.jhydrol.2013.11.054

Hernández, R. M. (2009). Estimación de la evapotranspiración de cultivo y requerimientos hídricos del tomate (Salanum lycopersicum Mill. cv. El Cid) en invernadero.

INEGI. (2017). Anuario estadístico y geográfico de Campeche 2017. https://doi.org/10.1111/j.1469-8749.2009.03468.x

Jovic, S., Nedeljkovic, B., Golubovic, Z., & Kostic, N. (2018). Evolutionary algorithm for reference evapotranspiration analysis. *Computers and Electronics in Agriculture*, *150*(April), 1–4. https://doi.org/10.1016/j.compag.2018.04.003

Leszek, Ł., Kanecka-geszke, E., Bak, B., & Slowinska, S. (2011). Estimation of Reference Evapotranspiration using the FAO Penman-Monteith Method for Climatic Conditions of Poland.

Mahmoud, S. H., & Gan, T. Y. (2019). Irrigation water management in arid regions of Middle East: Assessing spatio-temporal variation of actual evapotranspiration through remote sensing techniques and meteorological data. *Agricultural Water Management*, *212*(August 2018), 35–47. https://doi.org/10.1016/j.agwat.2018.08.040

Martel, M., Glenn, A., Wilson, H., & Kröbel, R. (2018). Journal of Hydrology : Regional Studies Simulation of actual evapotranspiration from agricultural landscapes in the Canadian Prairies. *Journal of Hydrology: Regional Studies*, *15*(November 2017), 105–118. https://doi.org/10.1016/j.ejrh.2017.11.010

Mathworks, C. (2018). User ' s Guide R 2018 b.

Mattar, M. A. (2018). Using gene expression programming in monthly reference evapotranspiration modeling: A case study in Egypt. *Agricultural Water Management*, *198*, 28–38. https://doi.org/S0378377417304092

Mehdizadeh, S. (2018). Estimation of daily reference evapotranspiration (ETo) using artificial intelligence methods: Offering a new approach for lagged ETodata-based modeling. *Journal of Hydrology*, *559*, 794–812. https://doi.org/10.1016/j.jhydrol.2018.02.060

Mehdizadeh, S., Behmanesh, J., & Khalili, K. (2017). Using MARS, SVM, GEP and empirical equations for estimation of monthly mean reference evapotranspiration. *Computers and Electronics in Agriculture*, *139*, 103–114. https://doi.org/10.1016/j.compag.2017.05.002

Moeletsi, M. E., Walker, S., & Hamandawana, H. (2013). Comparison of the Hargreaves and Samani equation and the Thornthwaite equation for estimating dekadal evapotranspiration in the Free State Province, South Africa. *Physics and Chemistry of the Earth*, *66*, 4–15. https://doi.org/10.1016/j.pce.2013.08.003

Negm, A., Jabro, J., & Provenzano, G. (2017). Assessing the suitability of American

National Aeronautics and Space Administration (NASA) agro-climatology archive to predict daily meteorological variables and reference evapotranspiration in Sicily, Italy. *Agricultural and Forest Meteorology*, *244–245*(October 2016), 111–121. https://doi.org/10.1016/j.agrformet.2017.05.022

Quej, V. H., Almorox, J., Arnaldo, J. A., & Saito, L. (2017). ANFIS, SVM and ANN soft-computing techniques to estimate daily global solar radiation in a warm sub-humid environment. *Journal of Atmospheric and Solar-Terrestrial Physics*, *155*(September 2016), 62–70. https://doi.org/10.1016/j.jastp.2017.02.002

Quintero, S., Andrés, E., Urueña, A., & Armando, H. (2010). CÚBICOS GUI for Data Interpolation by Cubic Spline, (44), 195–200.

RDevelopment, C. (2009). TEAM 2009: R: A Language and Environment for Statistical Computing. Vienna, Austria.

Sentelhas, P. C., Gillespie, T. J., & Santos, E. A. (2010). Evaluation of FAO Penman-Monteith and alternative methods for estimating reference evapotranspiration with missing data in Southern Ontario, Canada. *Agricultural Water Management*, *97*(5), 635–644. https://doi.org/10.1016/j.agwat.2009.12.001

Shiri, J. (2017). Evaluation of FAO56-PM, empirical, semi-empirical and gene expression programming approaches for estimating daily reference evapotranspiration in hyper-arid regions of Iran. *Agricultural Water Management*, *188*, 101–114. https://doi.org/10.1016/j.agwat.2017.04.009

Shiri, J., Ashraf, A., Hossein, A., Marti, P., Fakheri, A., Kisi, O., & Landeras, G. (2015). Independent testing for assessing the calibration of the Hargreaves – Samani equation : New heuristic alternatives for Iran. *COMPUTERS AND ELECTRONICS IN AGRICULTURE*, *117*, 70–80. https://doi.org/10.1016/j.compag.2015.07.010

Shiri, J., Sadraddini, A. A., Nazemi, A. H., Kisi, O., Landeras, G., Fakheri Fard, A., & Marti, P. (2014). Generalizability of Gene Expression Programming-based approaches for estimating daily reference evapotranspiration in coastal stations of Iran. *Journal of Hydrology*, *508*, 1–11. https://doi.org/10.1016/j.jhydrol.2013.10.034

Shrestha, N. K., & Shukla, S. (2015). Support vector machine based modeling of evapotranspiration using hydro-climatic variables in a sub-tropical environment. *Agricultural and Forest Meteorology*, *200*, 172–184. https://doi.org/10.1016/j.agrformet.2014.09.025

Straatmann, Z., Stevens, G., Vories, E., Guinan, P., & Travlos, J. (2018). Measuring short-crop reference evapotranspiration in a humid region using electronic atmometers. *Agricultural Water Management*, *195*, 180–186. https://doi.org/10.1016/j.agwat.2017.10.007

Topi, P. K. P., & Vanita, N. (2017). Estimation of reference evapotranspiration using data driven techniques under limited data conditions. *Modeling Earth Systems and Environment*, *0*(0), 0. https://doi.org/10.1007/s40808-017-0367-z

Torrente Cantó, L., Trillo Moya, J. C., & Ruiz Álvarez, J. (2018). Reconstrucción basada

en interpolacion de Hermite aplicada a funciones discontinuas.

Urraca, R., Antonanzas, J., Antonanzas-torres, F., & B, F. J. M. (2017). Estimation of Daily Global Horizontal Irradiation Using Extreme Gradient Boosting Machines. https://doi.org/10.1007/978-3-319-47364-2

Valiantzas, J. D. (2013). Simplified forms for the standardized FAO-56 Penman-Monteith reference evapotranspiration using limited weather data. *Journal of Hydrology*, *505*, 13–23. https://doi.org/10.1016/j.jhydrol.2013.09.005

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory*.

Venables, W. N., & Smith, D. M. (2013). An Introduction to R 3.0.1, *1*. https://doi.org/10.1016/S1751-3243(07)03003-9

Webb, C. P. (2010). BUREAU OF METEOROLOGY REFERENCE EVAPOTRANSPIRATION CALCULATIONS, (February).

Wen, X., Si, J., He, Z., & Wu, J. (2015). Support-Vector-Machine-Based Models for Modeling Daily Reference Evapotranspiration With Limited Climatic Data in Extreme Arid Regions. https://doi.org/10.1007/s11269-015-0990-2

Yan, Y., Wang, L., Wang, T., Wang, X., Hu, Y., & Duan, Q. (2018). Application of soft computing techniques to multiphase flow measurement: A review. *Flow Measurement and Instrumentation*, *60*(November 2017), 30–43. https://doi.org/10.1016/j.flowmeasinst.2018.02.017

Zanetti, S. S., Dohler, R. E., Cecílio, R. A., Eduardo, J., Pezzopane, M., & Xavier, A. C. (2019). Proposal For The Use Of Daily Thermal Amplitude For The Calibration Of The Hargreaves-Samani Equation. *Journal of Hydrology*. https://doi.org/10.1016/j.jhydrol.2019.01.049

Zhang, Z., Gong, Y., & Wang, Z. (2018). Accessible remote sensing data based reference evapotranspiration estimation modelling. *Agricultural Water Management*, *210*(July), 59–69. https://doi.org/10.1016/j.agwat.2018.07.039

CAPÍTULO XI. ANEXOS

Código 1. Macro “consecutivo” en Excel

```
Sub Consecutivo ()
Dim Nro_Fila As Double
Dim Valor_Celda As Double
Dim Comienzo_Fila As Double
Nro_Fila = InputBox (¿” Ingrese fila donde empezar?”, "Bienvenido")
Comienzo_Fila = Nro_Fila
Valor_Celda = Range ("A" & Nro_Fila).Value 'si no es la columna A cambiar letra
Do While Val(Range("A" & Nro_Fila).Value) > 0 'si no es la columna A cambiar letra
If Nro_Fila > Comienzo_Fila Then
If Val (Range ("A" & Nro_Fila).Value) > Valor_Celda Then 'si no es la columna A cambiar letra
Rows(Nro_Fila & ":" & Nro_Fila).Select
Selection.Insert Shift:=xlDown
Range("A" & Nro_Fila).Value = Valor_Celda 'si no es la columna A cambiar letra
End If
End If
Nro_Fila = Nro_Fila + 1
Valor_Celda = Valor_Celda + 6.94444444525288E-03
Loop
End Sub
```

Código 2. Interpolación Polinómica de Hermite (PCHIP) en Matlab

```
## Especifica de que hoja de Excel tomara los datos
## Leer los datos
datos = xlsread('Calakmul_2004.xlsx',1);
## Rango de valores
x = (2:39925)';
```

```
## Posición de la columna a analizar
vv= datos(:,3);
## Asignación de la función PCHIP a la variable
vvregresion= pchip(x,vv,x);
## Columna donde se asignaran los valores interpolados
variable1 = {'vvregresion'};
## Se escriben los resultados en un archivo de excel
[sucess, message]=xlswrite('Calakmul_2004.xlsx',variable1,1,'I1:I1')
[sucess, message]=xlswrite('Calakmul_2004.xlsx',vvregresion,1,'I2:I39925')
## Posición de la columna a analizar
temp= datos(:,4);
## Asignación de la función PCHIP a la variable
tempregresion= pchip(x,temp,x);
## Columna donde se asignaran los valores interpolados
variable2 = {'tempregresion'};
## Se escriben los resultados en un archivo de excel
[sucess, message]=xlswrite('Calakmul_2004.xlsx',variable2,1,'J1:J1')
[sucess, message]=xlswrite('Calakmul_2004.xlsx',tempregresion,1,'J2:J39925')


## Posición de la columna a analizar
hr= datos(:,5);
## Asignación de la función PCHIP a la variable
hrregresion= pchip(x,hr,x);
## Columna donde se asignaran los valores interpolados
variable3 = {'hrregresion'};
## Se escriben los resultados en un archivo de excel
[sucess,message]=xlswrite('Calakmul_2004.xlsx',variable3,1,'K1:K1')
[sucess,message]=xlswrite('Calakmul_2004.xlsx', hrregresion,1,'K2:K39925')


## Posición de la columna a analizar
rs= datos(:,7);
## Asignación de la función PCHIP a la variable
rsregresion= pchip(x,rs,x);
```

```
## Columna donde se asignaran los valores interpolados
variable4 = {'rsregresion'};
## Se escriben los resultados en un archivo de excel
[sucess,message]=xlswrite('Calakmul_2004.xlsx',variable4,1,'L1:L1')
[sucess,message]=xlswrite('Calakmul_2004.xlsx',rsregresion,1,'L2:L39925')
```

Código 3. Interpolación SPLINE en Matlab

```
## Especifica de que hoja de Excel tomara los datos
## Leer los datos
datos = xlsread('Calakmul_2004.xlsx',1);
## Rango de valores
x = (2:39925)';
## Posición de la columna a analizar
vv= datos(:,3);
## Asignación de la función SPLINE a la variable
vvregresion= spline(x,vv,x);
## Columna donde se asignaran los valores interpolados
variable1 = {'vvregresion'};
## Se escriben los resultados en un archivo de excel
[sucess, message]=xlswrite('Calakmul_2004.xlsx',variable1,1,'P1:P1')
[sucess, message]=xlswrite('Calakmul_2004.xlsx',vvregresion,1,'P2:P39925')

## Posición de la columna a analizar
temp= datos(:,4);
## Asignación de la función SPLINE a la variable
tempregresion= spline(x,temp,x);
## Columna donde se asignaran los valores interpolados
variable2 = {'tempmaxregresion'};
## Se escriben los resultados en un archivo de excel [sucess,
message]=xlswrite('Calakmul_2004.xlsx',variable2,1,'Q1:Q1')
[sucess, message]=xlswrite('Calakmul_2004.xlsx',tempregresion,1,'Q2:Q39925')
```

```
## Posición de la columna a analizar
hr= datos(:,5);
## Asignación de la función SPLINE a la variable
hrregresion= spline(x,hr,x);
## Columna donde se asignaran los valores interpolados
variable3 = {'hrregresion'};
## Se escriben los resultados en un archivo de excel [sucess,
message]=xlswrite('Calakmul_2004.xlsx',variable3,1,'R1:R1')
[sucess, message]=xlswrite('Calakmul_2004.xlsx', hrregresion,1,'R2:R39925')


## Posición de la columna a analizar
rs= datos(:,7);
## Asignación de la función SPLINE a la variable
rsregresion= spline(x,rs,x);
## Columna donde se asignaran los valores interpolados
variable4 = {'rsregresion'};
## Se escriben los resultados en un archivo de excel [sucess,
message]=xlswrite('Calakmul_2004.xlsx',variable4,1,'S1:S1')
[sucess, message]=xlswrite('Calakmul_2004.xlsx',rsregresion,1,'S2:S39925')
```

Codigo 4. Detección de datos atípicos método Grubbs en Matlab

```
## Especifica de que hoja de Excel tomara los datos
## Leer los datos
datos = xlsread('Calakmul_2005.xlsx',3);
## Posición de la columna a analizar
vv= datos(:,2);
tmax= datos(:,3);
tmin= datos(:,4);
hr = datos(:,5);
rs= datos(:,7);
## Transponer los datos
[A] = transpose(vv);
[B] = transpose(tmax);
```

```
[C] = transpose(tmin);
[D] = transpose(hr);
[E] = transpose(rs);
## Rango de valores
x = 1:359;
## Asignación de la función SPLINE a la variable
[TF1,L1,U1,C1] = isoutlier(A, 'grubbs' );
[TF2,L2,U2,C2] = isoutlier(B, 'grubbs' );
[TF3,L3,U3,C3] = isoutlier(C, 'grubbs' );
[TF4,L4,U4,C4] = isoutlier(D, 'grubbs' );
[TF5,L5,U5,C5] = isoutlier(E, 'grubbs' );
## Columna  donde se asignaran nuevos valores
viento = TF1.';
Tmax = TF2.';
Tmin = TF3.';
Hr = TF4.';
Rs = TF5.';
## Se escriben los resultados en un archivo de Excel
[sucess,message]=xlswrite('C:/Users/user/Documents/MATLAB/Calakmul_2005.xlsx',viento,3,'J
2:J360')
[sucess,message]=xlswrite('C:/Users/user/Documents/MATLAB/Calakmul_2005.xlsx',Tmax,3,'K
2:K360')
[sucess,message]=xlswrite('C:/Users/user/Documents/MATLAB/Calakmul_2005.xlsx',Tmin,3,'L2
:L360')
[sucess,message]=xlswrite('C:/Users/user/Documents/MATLAB/Calakmul_2005.xlsx',Hr,3,'M2:
M360')
[sucess,message]=xlswrite('C:/Users/user/Documents/MATLAB/Calakmul_2005.xlsx',Rs,3,'N2:
N360')
## Se guarda el grafico
plot(x,B,x(TF6),B(TF6),'x',x,L6*ones(1,279),x,U6*ones(1,279),x,C6*ones(1,279));
legend('Valores Originales','Valor Atípico','Umbral Inferior','Umbral Superior','Valor Central')
title('Valores Atípicos Método Grubbs')
```

Código 5. Detección de valores atípicos método Mean en Matlab

```
## Especifica de que hoja de Excel tomara los datos
## Leer los datos
datos = xlsread('Calakmul_2005.xlsx',3);
## Posición de la columna a analizar
vv= datos(:,2);
tmax= datos(:,3);
tmin= datos(:,4);
hr = datos(:,5);
rs= datos(:,7);
## Transponer los datos
[A] = transpose(vv);
[B] = transpose(tmax);
[C] = transpose(tmin);
[D] = transpose(hr);
[E] = transpose(rs);
## Rango de valores
x = 1:359;

## Asignación de la función SPLINE a la variable

[TF6,L6,U6,C6] = isoutlier(A, 'mean' );
[TF7,L7,U7,C7] = isoutlier(B, 'mean' );
[TF8,L8,U8,C8] = isoutlier(C, 'mean' );
[TF9,L9,U9,C9] = isoutlier(D, 'mean' );
[TF10,L10,U10,C10] = isoutlier(E, 'mean' );
## Columna  donde se asignaran nuevos valores
viento2 = TF6.';
Tmax2 = TF7.';
Tmin2 = TF8.';
Hr2 = TF9.';
Rs2 = TF10.';
## Se escriben los resultados en un archivo de excel
```

```
[sucess,message]=xlswrite('C:/Users/user/Documents/MATLAB/Calakmul_2005.xlsx',viento2,3,'
P2:P360')
[sucess,message]=xlswrite('C:/Users/user/Documents/MATLAB/Calakmul_2005.xlsx',Tmax2,3,'
Q2:Q360')
[sucess,message]=xlswrite('C:/Users/user/Documents/MATLAB/Calakmul_2005.xlsx',Tmin2,3,'
R2:R360')
[sucess,message]=xlswrite('C:/Users/user/Documents/MATLAB/Calakmul_2005.xlsx',Hr2,3,'S2:
S360')
[sucess,message]=xlswrite('C:/Users/user/Documents/MATLAB/Calakmul_2005.xlsx',Rs2,3,'T2:
T360')
## Se guarda el grafico
plot(x,B,x(TF6),B(TF6),'x',x,L6*ones(1,279),x,U6*ones(1,279),x,C6*ones(1,279));
legend('Valores Originales','Valor Atípico','Umbral Inferior','Umbral Superior','Valor Central')
title('Valores Atípicos Método Mean')
```

Código 6. Detección de valores atípicos método Quartiles en Matlab

```
## Especifica de que hoja de Excel tomara los datos
## Leer los datos
datos = xlsread('Calakmul_2005.xlsx',3);
## Posición de la columna a analizar
vv= datos(:,2);
tmax= datos(:,3);
tmin= datos(:,4);
hr = datos(:,5);
rs= datos(:,7);
## Transponer los datos
[A] = transpose(vv);
[B] = transpose(tmax);
[C] = transpose(tmin);
[D] = transpose(hr);
[E] = transpose(rs);
## Rango de valores
x = 1:359;
```

```
## Asignación de la función SPLINE a la variable
[TF11,L11,U11,C11] = isoutlier(A, 'quartiles' );
[TF12,L12,U12,C12] = isoutlier(B, 'quartiles' );
[TF13,L13,U13,C13] = isoutlier(C, 'quartiles' );
[TF14,L14,U14,C14] = isoutlier(D, 'quartiles' );
[TF15,L15,U15,C15] = isoutlier(E, 'quartiles' );
## Columna  donde se asignaran nuevos valores
viento3 = TF11.';
Tmax3 = TF12.';
Tmin3 = TF13.';
Hr3 = TF14.';
Rs3 = TF15.';
## Se escriben los resultados en un archivo de excel
[sucess,message]=xlswrite('C:/Users/user/Documents/MATLAB/Calakmul_2005.xlsx',viento3,3,'
V2:V360')
[sucess,message]=xlswrite('C:/Users/user/Documents/MATLAB/Calakmul_2005.xlsx',Tmax3,3,'
W2:W360')
[sucess,message]=xlswrite('C:/Users/user/Documents/MATLAB/Calakmul_2005.xlsx',Tmin3,3,'
X2:X360')
[sucess,message]=xlswrite('C:/Users/user/Documents/MATLAB/Calakmul_2005.xlsx',Hr3,3,'Y2:
Y360')
[sucess,message]=xlswrite('C:/Users/user/Documents/MATLAB/Calakmul_2005.xlsx',Rs3,3,'Z2:
Z360')
## Se guarda el grafico
plot(x,B,x(TF6),B(TF6),'x',x,L6*ones(1,279),x,U6*ones(1,279),x,C6*ones(1,279));
legend('Valores Originales','Valor Atípico','Umbral Inferior','Umbral Superior','Valor Central')
title('Valores Atípicos Método Quartiles')
```

Código 7. Calibración de Hargreaves-Samani en Matlab

```
% Se leen los datos a partir de un archivo

tbl = readtable('Petenes.csv','ReadRowNames',true,'TreatAsEmpty',{'NA'});
%'TreatAsEmpty',{'N/A'}
%Tmax(1),Tmin(2),HR(3),Tmed(4),emax(5),emin(6),emean(7),D(8),Ra(9),Lamda( 0)
```

```
%Rs(11),I(12),a(13),N(14),pp(15),Eto(16)
data=table2array(tbl);
hm=data(:,6);
%Hargreaves and samani
%Se ejecuta el ajuste de regresión no lineal
beta0 = [-1];
hs = @(b,x)b(1).* 0.408* x(:,3) .* (x(:,4) + 17.8).* (x(:,1) - x(:,2)).^0.5;
mdlHS = fitnlm(tbl,hs,beta0);
[ypred_hs yci_hs] = predict(mdlHS);
r_hs = mdlHS.Residuals.Raw;
[code1,mbe1,mae1,mpe1,mape1,rmse1,ia1,mad1,iab1,leg1] = estad(hm,ypred_hs);
Hargreaves=table(code1,mbe1,mae1,mpe1,mape1,rmse1,ia1,mad1,iab1,leg1)
```

Código 8. Calibración de Camargo en Matlab

```
% Se leen los datos a partir de un archivo

tbl = readtable('Petenes.csv','ReadRowNames',true,'TreatAsEmpty',{'NA'});
%'TreatAsEmpty',{'N/A'}
%Tmax(1),Tmin(2),HR(3),Tmed(4),emax(5),emin(6),emean(7),D(8),Ra(9),Lamda(10)
%Rs(11),I(12),a(13),N(14),pp(15),Eto(16)
data=table2array(tbl);
hm=data(:,6);
%Camargo model
%Se ejecuta el ajuste de regresión no lineal
beta8 = [10;0.1];
camargo=@(b,x)((b(1).*(10.*(b(2)*(3*x(:,1)-
x(:,2))))./97.8813776).^2.13975824)/30).*(x(:,5)/12);
mdlcamargo = fitnlm(tbl,camargo,beta8);
[ypred_camargo yci_camargo] = predict(mdlcamargo); %predictions
r_camargo= mdlcamargo.Residuals.Raw;
```

```
[code9,mbe9,mae9,mpe9,mape9,rmse9,ia9,mad9,iab9,leg9] =
estad(hm,ypred_camargo);
Camargo=table(code9,mbe9,mae9,mpe9,mape9,rmse9,ia9,mad9,iab9,leg9)
```

Código 9. Ajuste de los parámetros de SVM mediante al software R utilizando el algoritmo genético

```
##Se leen los datos del archivo
Data <- read.table("ETo.txt")
#Se configuran y dividen los datos para la validación cruzada.
K = 5 # 5-partes Validación cruzada
fold_inds <- sample(1:K, nrow(Data), replace = TRUE)
lst_CV_data <- lapply(1:K, function(i) list(
train_data = Data[fold_inds != i, , drop = FALSE],
test_data = Data[fold_inds == i, , drop = FALSE]))
#Dados los valores de los parámetros 'cost', 'gamma' and 'epsilon', se evalua el RMSE del modelo
sobre los datos de la validación.
evalParams <- function(train_data, test_data, cost, gamma, epsilon) {
# Entrenamiento
model <- svm(V1 ~ ., data = train_data, cost = cost, gamma = gamma, epsilon = epsilon, type =
"eps-regression", kernel = "radial")
# Validación
rmse <- mean((predict(model, newdata = test_data) - test_data$V1) ^ 2)
   return (rmse)}
# Se define la función Fitness a ser maximizada
# El parametro vector es x: (cost, gamma, epsilon)
fitnessFunc <- function(x, Lst_CV_Data) {
# Corrección de los parámetros del SVM.
cost_val <- x[1]
gamma_val <- x[2]
```

```
epsilon_val <- x[3]
# Se usa validación cruzada para minimizar RMSE por cada juego de datos divididos.
rmse_vals <- sapply(Lst_CV_Data, function(in_data) with(in_data, evalParams(train_data, test_data, cost_val, gamma_val, epsilon_val)))

# En el cálculo del fitness, se regresa un valor mínimo de RMSE (sobre la validación cruzada),
# Así por maximización del fitness se minimiza el valor del RMSE
   return (-mean(rmse_vals))}
# Rango de los parametros a determinarse
# Parametros: (cost, gamma, epsilon)
theta_min <- c(cost = 1e-4, gamma = 1e-3, epsilon = 1e-2)
theta_max <- c(cost = 10, gamma = 2, epsilon = 2)
# Ejecución del algoritmo genetico.
results <- ga(type = "real-valued", fitness = fitnessFunc, lst_CV_data,
   names = names(theta_min),
   min = theta_min, max = theta_max,
   popSize = 60, maxiter = 10)
summary(results)
plot(results)
```

Código 10. Obtención del modelo SVM con el software R

```
##SE LEEN DATOS
data<-read.table("CampecheETo.txt")
## SE FRACCIONAN LOS DATOS PARA ENTRENAMIENTO Y PRUEBA.
index <- 1:nrow(data)
testindex <- sample(index, trunc(length(index)/3))
testset <- (data[testindex,1:4])
trainset <- (data[-testindex,1:4])
##SE ENTRENA EL MODELO USANDO LOS PARAMETROS PREVIAMENTE AJUSTADOS
```

```
svm.model <- svm(V1 ~ ., data = trainset, type= "eps-regression", kernel= "radial", cost = 3.752501, gamma =  0.5356639, epsilon =  0.3442293, scale=TRUE)
##SE VALIDA EL MODELO
svm.pred <- predict(svm.model, testset[,-1])
svm.pred2 <- predict(svm.model, trainset[,-1])
##SE DETERMINAN LOS ERRORES MAE Y RMSE
##RMSE
rmse <- function(error){
sqrt(mean(error^2))}
error <- (svm.pred - testset[,1])
error2<-(svm.pred2 - trainset[,1])
RMSEmodel <- rmse(error)
RMSEtrain <-rmse(error2)
##MAE
mae <- function(error2){
mean(abs(error2))}
error3 <- svm.pred - testset[,1]
MAEmodel <- mae(error3)
error4 <- svm.pred2 - trainset[,1]
MAEtrain <- mae(error4)
##SE ELABORA EL GRAFICO DE DISPERSION
##SE LLEVA A CABO LA REGRESION LINEAL
regresion<-lm(formula = svm.pred ~ testset[, 1])
regresion2<-lm(formula = svm.pred2 ~ trainset[, 1])
plot(svm.pred, testset[,1], pch=20, main= "Campeche",xlab="FAO56-PM ETo (mm day-1)", ylab="SVM ETo (mm day-1)", xlim=c(0.3, 8), ylim=c(0.3, 8), width = 10, height = 5)
legend("topleft", bty="n", legend=paste("R2 =", format(summary(regresion)$adj.r.squared, digits=3)))
##SE GUARDA EL GRAFICO EN ARCHIVO
```

```
dev.copy(png,"Campeche_plot.jpg",width=10,height=7,units="in",res=300)
dev.off()
write.table(trainset[,1], 'TRAIN.txt', sep='\t')
write.table(testset[,1], 'TEST.txt', sep='\t')
write.table(svm.pred, 'TEST_pred.txt', sep='\t')
write.table(svm.pred2, 'TRAIN_pred.txt', sep='\t')
RMSEmodel
RMSEtrain
MAEmodel
MAEtrain
```

Código 11. Ajuste de parámetros del algoritmo XGBoost mediante el software R.

```
#Carga de datos
require(caret)
require(Matrix)
require(xgboost)
mydata <- read.csv('calakmul.csv')
#Seleccionar las variables a utilizar
mydata <- mydata[, c('Eto', 'Tmax', 'Tmin', 'Ra')]
#Dividir mis datos para el entrenamiento y test
library (caret)
#Fijar el valor semilla para crear un juego reproducible de entrenamiento y test
set.seed(300)
#Crear muestras estratificadas aleatorias
#Referenciar mi variable objetivo
trainIndex   <- createDataPartition(mydata$Eto, p=0.75, list=FALSE, times=1)
train        <- mydata[ trainIndex, ]
test         <- mydata[-trainIndex, ]
# Se ajustan los hiperparametros de busqueda en la validación cruzada
```

```
xgb_grid = expand.grid(
  nrounds = c (2, 10 , 20),
  eta = c (0.01, 0.001, 0.0001),
  max_depth = c(5, 10, 15),
  gamma = c(1, 2, 3),
  colsample_bytree = c(0.4, 0.7, 1.0),
  min_child_weight = c(0.5, 1, 1.5),
  subsample = 1)
# empacar los parámetros de control de entrenamiento
xgb_trncontrol = trainControl (
            method = "cv",
            number = 10,
            allowParallel = FALSE)
# entrenar el modelo para cada combinación de parámetros en la cuadrícula
#   Usando validación cruzada
gbmFit4 <- train(Eto ~ ., data = train,
            trControl = xgb_trncontrol,
            method = "xgbTree" )
gbmFit1
```

Código 12. Obtención del modelo XGboost con el software R

```
#Carga de datos
require(caret)
require(Matrix)
require(xgboost)
mydata <- read.csv('Petenes.csv')
#Seleccionar las variables a utilizar
mydata <- mydata[, c('Eto', 'Tmax', 'Tmin', 'Ra')]
#Dividir mis datos para el entrenamiento y test
```

```
library (caret)
#Fijar el valor semilla para crear un juego reproducible de entrenamiento y test
set.seed(300)
#Crear muestras estratificadas aleatorias
#Referenciar mi variable objetivo
trainIndex   <- createDataPartition(mydata$Eto, p=0.75, list=FALSE, times=1)
train        <- mydata[ trainIndex, ]
test         <- mydata[-trainIndex, ]
#Cargar el paquete Matrix
library(Matrix)
#Crear vectores separados de nuestro vector objetivo para el entrenamiento y test
train.label  <- train$Eto
test.label   <- test$Eto
# Preparar las matrices para el entrenamiento
dtrain  <- sparse.model.matrix(Eto ~ .-1, data=train)
dtest   <- sparse.model.matrix(Eto ~ .-1, data=test)
#Ver el numero de columnas y caracteristicas de cada juego.
dim (dtrain)
dim (dtest)
#ENTRENAMIENTO DEL MODELO
# Cargar el paquete XGBoost
library(xgboost)
# Se fijan los hiper parámetros previamente ajustados por validación cruzada.
param <- list(objective   = "reg:linear",
              eval_metric = "rmse",
              max_depth   = 2,
              eta         = 0.3,
              gammma      = 0,
              colsample_bytree = 0.8,
```

```
          min_child_weight = 1)
set.seed(1234)

# Pasamos los hiper parámetros y se entrena el modelo
system.time(xgb <- xgboost(params  = param,
                    data    = dtrain,
                    label   = train.label,
                    nrounds = 50,
                    print_every_n = 2,
                    verbose = 1))
# Se crean las predicciones
xgboost.pred <- predict(xgb, dtest)
xgboost.pred2 <- predict(xgb, dtrain)
##SE DETERMINAN LOS ERRORES MAE Y RMSE
##RMSE
rmse <- function(error){
  sqrt(mean(error^2))}
error <- (xgboost.pred - test.label)
error2<-(xgboost.pred2 - train.label)
RMSEmodel <- rmse(error)
RMSEtrain <-rmse(error2)
##MAE
mae <- function(error2){
    mean(abs(error2))}
error3 <- xgboost.pred - test.label
MAEmodel <- mae(error3)
error4 <- xgboost.pred2 - train.label
MAEtrain <- mae(error4)
##SE ELABORA EL GRAFICO DE DISPERSION
```

```
##SE LLEVA A CABO LA REGRESION LINEAL
regresion<-lm(formula = xgboost.pred ~ test.label)
regresion2<-lm(formula = xgboost.pred2 ~ train.label)
plot(xgboost.pred, test.label, pch=20, main= "Petenes",xlab="FAO56-PM ETo (mm day-1)", ylab="Xgboost ETo (mm day-1)", xlim=c(0.3, 8), ylim=c(0.3, 8), width = 10, height = 5)
legend("topleft", bty="n", legend=paste("R2 =", format(summary(regresion)$adj.r.squared, digits=3)))
##SE SALVA el GRAFICO EN ARCHIVO
dev.copy(png,"Petenes_plot.jpg",width=10,height=7,units="in", res=300)
dev.off()
write.table(train.label, 'TRAIN.txt', sep='\t')
write.table(test.label, 'TEST.txt', sep='\t')
write.table(xgboost.pred, 'TEST_pred.txt', sep='\t')
write.table(xgboost.pred2, 'TRAIN_pred.txt', sep='\t')
##Se lleva a cabo la evaluación de los errores mediante el estadístico RMSE (Raíz Cuadrada Media del Error)  y MAE (Error Absoluto Medio).
RMSEmodel
RMSEtrain
MAEmodel
MAEtrain
```