DOI: 10.24850/j-tyca-2021-03-02

Artículos

Estimación de la evapotranspiración de referencia con datos de temperatura: una comparación entre técnicas de cálculo convencionales y de inteligencia artificial en una región cálida-subhúmeda

Estimation of reference evapotranspiration from temperature data: A comparison between conventional calculation and artificial intelligence techniques in a warm-sub-humid region

Luis Alberto Ramos-Cirilo[1], ORCID: https://orcid.org/0000-0002-0921-7738

Victor Hugo Quej-Chi[2], ORCID: https://orcid.org/0000-0002-9356-6251

Eugenio Carrillo-Ávila[3], ORCID: https://orcid.org/0000-0002-8018-7869

Everardo Aceves-Navarro[4], ORCID: https://orcid.org/0000-0003-2711-7412

Benigno Rivera-Hernández[5], ORCID: https://orcid.org/0000-0003-1713-4710

[1]Colegio de Postgraduados, Campus Campeche, Sihochac, Champotón, Campeche, México, l_ramos90@yahoo.com

[2]Colegio de Postgraduados, Campus Campeche, Sihochac, Champotón, Campeche, México, quej@colpos.mx

[3]Colegio de Postgraduados, Campus Campeche, Sihochac, Champotón, Campeche, México, ceugenio@colpos.mx

[4]Colegio de Postgraduados, Campus Campeche, Sihochac, Champotón, Campeche, México, evarardo.aceves@colpos.com

[5]Universidad Popular de la Chontalpa, Cárdenas, Tabasco, México, brivera@colpos.mx

Autor para correspondencia: Víctor Hugo Quej-Chi, quej@colpos.mx

Resumen

La evapotranspiración de referencia (*ETo*) es un parámetro agro-meteorológico de gran importancia para muchas áreas de estudio como la geotecnia, climatología e hidrología, donde su mayor importancia recae en el cálculo de la evapotranspiración de cultivo (*ETc*). En el presente estudio, utilizando solamente datos de temperatura, se evaluó el desempeño de tres modelos de inteligencia artificial y dos ecuaciones convencionales para predecir la evapotranspiración de referencia (*ETo*) en un clima cálido subhúmedo en México. Los modelos de inteligencia artificial evaluados fueron máquinas de soporte vectorial (SVM), programación de expresión genética (GEP) y XGBoost, así como los modelos convencionales de Hargreaves- Samani y Camargo. El desempeño de los modelos se evaluó de acuerdo con los índices estadísticos error absoluto medio (MAE); raíz cuadrada media del error

(RMSE); coeficiente de determinación (R^2), y el error medio de sesgo (MBE). Se construyeron intervalos de confianza para cada índice estadístico utilizando la técnica de remuestreo *bootstrap*, con el propósito de evaluar la incertidumbre de los mismos. Los resultados demuestran que entre los modelos convencionales evaluados la ecuación de Camargo obtuvo un mejor desempeño en la estimación de la *ETo* en comparación con la ecuación de Hargreaves. Respecto a los modelos de inteligencia artificial, el modelo SVM obtuvo mejor desempeño entre las técnicas evaluadas. De manera general, se recomienda utilizar el modelo SVM para estimar valores de *ETo* al superar a las demás técnicas.

Palabras clave: evapotranspiración de referencia, técnicas de inteligencia artificial, estaciones meteorológicas automatizadas, *bootstrap*.

Abstract

Reference evapotranspiration (*ETo*) is an agro-meteorological parameter of great importance for many areas of study such as geotechnics, climatology and hydrology, where its greatest importance falls in the calculation of the crop's evapotranspiration (*ETc*). In this study, using only temperature data, the performance of three artificial intelligence models and two conventional equations to predict the reference evapotranspiration (*ETo*) was evaluated in a warm sub-humid climate in México. The artificial intelligence models evaluated were: support vector machines (SVM), Gene Expression Programming (GEP) and XGBoost, and the conventional models were those by Hargreaves-Samani and Camargo.

The performance of the models was evaluated according to the statistical indexes: Mean Absolute Error (MAE); Root Mean Square Error (RMSE); Coefficient of Determination (R^2), and Mean Bias Error (MBE). Confidence intervals were constructed for each statistical index using the technique of bootstrap resampling with the purpose of evaluating their uncertainty. The results show that among the conventional models evaluated, the equation by Camargo obtained a better performance in the estimation of *ETo* compared to the equation by Hargreaves. Regarding the artificial intelligence models, the SVM model obtained the best performance among the techniques evaluated. In general, it is recommended to use the SVM model to estimate the *ETo* values since it outperforms the other techniques.

Keywords: Reference evapotranspiration, artificial intelligence techniques, automated weather stations, bootstrap.

Recibido: 26/06/2019

Aceptado: 28/06/2020

Introducción

La evapotranspiración de referencia (*ETo*) es un parámetro agro-meteorológico de uso en muchas áreas de estudio como la geotecnia, climatología e hidrología, donde su mayor importancia radica en el cálculo de la evapotranspiración de cultivo (*ETc*) en la determinación de requerimientos hídricos en los cultivos agrícolas (Čadro, Uzunović, Žurovec, & Žurovec, 2017; Jovic, Nedeljkovic, Golubovic, & Kostic, 2018; Webb, 2010; Zhang, Gong, & Wang, 2018). La *ETo* se define como la "tasa de evapotranspiración de una superficie de referencia hipotética que presentan características específicas" (Allen, Pereira, Raes, & Smith, 1998). El cálculo exacto se realiza usando la ecuación estándar de la FAO 56 Penman-Monteith (ETo-FAO56PM) (Shiri, 2017); sin embargo, la ecuación requiere de cuatro variables meteorológicas, como son la radiación solar, humedad relativa, velocidad del viento y temperatura, que muchas veces no son medidas en las estaciones meteorológicas, por lo que en muchas ocasiones se opta por la utilización de ecuaciones que emplean menos variables meteorológicas, las cuales se clasifican dependiendo de la disponibilidad de variables (Fan *et al*., 2018a; Feng, Cui, Zhao, Hu, & Gong, 2016; Shiri, 2017). Una de las principales razones de uso de las ecuaciones convencionales es que requieren un menor número de variables meteorológicas para su implementación, siendo aquellas basadas en el parámetro temperatura del aire las menos precisas. En un estudio llevado por Almorox, Senatore, Quej y Mendicino (2018), evaluaron el desempeño del método PMT (Penman-Monteith) y compararon los resultados con los obtenidos con la ecuación de Hargreaves-Samani (HS) utilizando los datos mensuales medidos a largo plazo del conjunto de datos climáticos globales de la FAO New LocClim.

Para una base de datos completa, la expresión aproximada de PMT usando únicamente la temperatura del aire produce mejores resultados que el método de la ecuación no calibrada de HS, y el desempeño del método de PMT se desempeña aún mejor adoptando correcciones dependiendo del tipo de clima para la estimación de la radiación solar, en especial en el tipo de clima tropical.

Antonopoulos y Antonopoulos (2017) emplearon la técnica de inteligencia artificial redes neurales artificiales (ANN), y los métodos de Priestley-Taylor, Makkink (MAK), Hargreaves y transferencia de masa para estimar la evapotranspiración de referencia con datos meteorológicos diarios en un periodo de cinco años (2009-2013) en el norte de Grecia. Como resultado, se observó que al utilizar variables de entrada limitada para el ajuste de los parámetros del ANNs, los datos resultan en valores de *ETo* inexactos. Por otro lado, los métodos basados en radiación solar de Priestley-Taylor y Makkink correlacionaron correctamente con el método de Penman-Monteith seguido por el método de Hargreaves. El método de transferencia de masa fue correlacionado de manera satisfactoria, pero subestimaba los valores de *ETo*.

Recientes investigaciones en la determinación de la *ETo* hacen mención de las técnicas denominadas de inteligencia artificial o *soft-computing*, basadas en el aprendizaje automático. Estas técnicas han sido ampliamente usadas en la modelación hidrológica, y en la estimación de la *ETo* han mostrado superioridad sobre las ecuaciones convencionales debido a que incrementan la precisión de las estimaciones utilizando pocas variables (Mehdizadeh, 2018).

La técnica de inteligencia artificial llamada programación de expresión genética (*Gene Expression Programming*, GEP) propone un enfoque alternativo, que genera algoritmos y/o expresiones para resolver problemas de forma automática, y donde recientemente se ha aplicado con buenos resultados en estudios hidrológicos (Mattar, 2018). Mattar y Alazba (2019) estimaron la evapotranspiración de referencia usando la programación de expresión genética (GEP) y regresión lineal múltiple (*Multiple Linear Regression*, MLR), con datos recolectados de estaciones en Egipto; los resultados demuestran que la técnica de GEP, cuando se le agregaron los datos de la variable temperatura, obtuvo mejor desempeño que el modelo MLR y otras ecuaciones convencionales (HS y MAK). En otro estudio para evaluar el desempeño de algunas técnicas de inteligencia artificial, Wen *et al*. (2015) evaluaron el uso de la máquina de soporte vectorial (SVM) para modelar la evapotranspiración de referencia diaria (*ETo*) utilizando datos climáticos limitados. Para el SVM se usaron cuatro combinaciones de temperatura máxima del aire ($T_{máx}$), temperatura mínima del aire ($T_{mín}$), velocidad del viento (U_2) y radiación solar diaria (R_s) en la región extremadamente árida de la cuenca de Ejina, China, como entradas con $T_{máx}$ y $T_{mín}$ en el conjunto de datos base. Los resultados de los modelos SVM se evaluaron comparando la salida con la *ETo* calculada con la ecuación de Penman Monteith FAO 56 (PMF-56); la precisión del método SVM se comparó con el de la red neuronal artificial (ANN) y tres modelos convencionales, incluidos Priestley-Taylor, Hargreaves y Ritchie. Los resultados mostraron que el rendimiento del método SVM fue el mejor entre estos modelos.

Actualmente 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 (Fan *et al*., 2018a). Fan *et al.* (2018b) evaluaron el potencial para estimar la *ETo* de los modelos de algoritmos de ensamblaje basados en árboles; bosques aleatorio (*RF*), modelo árbol M5, aumento de gradiente (GBDT), aumento de gradiente extremo (XGBoost), máquinas de soporte vectorial (SVM) y máquinas de aprendizaje extremo (ELM); los resultados mostraron que los modelos XGBoost y GBDT alcanzaron excelente precisión y estabilidad en comparación con los modelos SVM y ELM, pero con menos costos computacionales. Bajo tales criterios, los autores recomiendan el uso de estos modelos para estimar la *ETo*.

Considerando que la variable meteorológica temperatura del aire es la de mayor disponibilidad, el presente estudio tiene como objetivo (1) evaluar la capacidad de tres métodos de inteligencia artificial llamados XGBoost, GEP y SVM para estimar valores de *ETo* utilizando datos de temperatura del aire, y (2) comparar los resultados con dos ecuaciones convencionales basadas en temperatura llamadas Hargreaves-Samani y Camargo bajo un clima cálido-subhúmedo.

Materiales y métodos

Sitio de estudio

La presente investigación se realizó utilizando datos de estaciones meteorológicas automatizadas (EMAs) ubicadas en el estado de Campeche, México (Figura 1). El clima predominante es cálido subhúmedo, que se presenta en el 92 % de su territorio y el 7.75 % presenta clima cálido húmedo localizado en la parte este del estado y en la parte norte un 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).

Tecnología y ciencias del agua, ISSN 2007-2422, 12(3), 32-81. DOI: 10.24850/j-tyca-2021-03-02

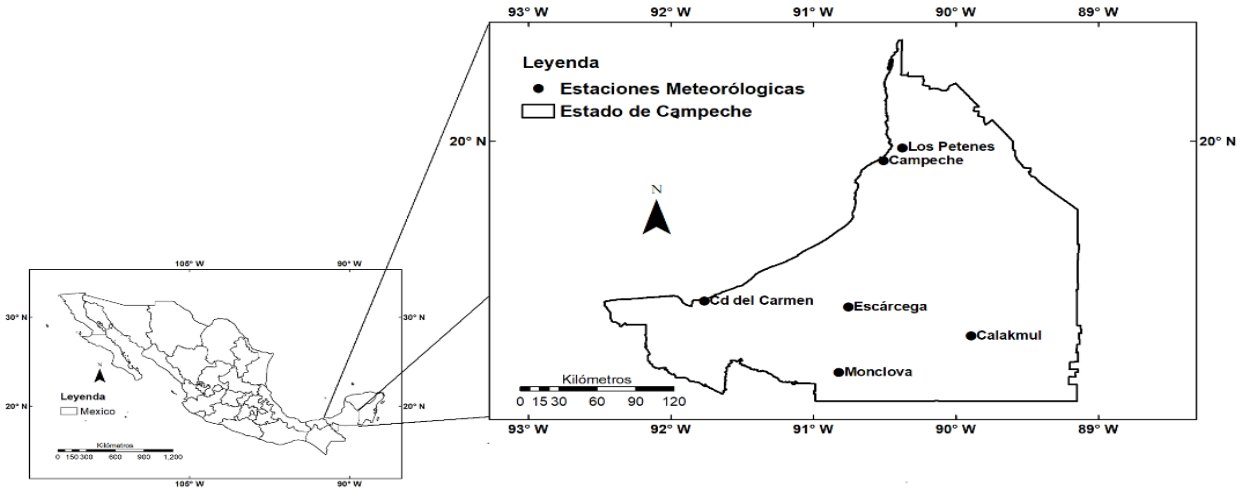

Figura 1. Ubicación de las estaciones meteorológicas en el estado de Campeche, México, usadas en este estudio.

Las bases de datos cada 10 minutos se obtuvieron de las estaciones meteorológicas automatizadas de la Comisión Nacional del Agua (Conagua) de México. La Tabla 1 muestra la información geográfica de las estaciones meteorológicas utilizadas en este estudio, así como los periodos de tiempo de registro de cada estación meteorológica.

Tabla 1. Información geográfica y condiciones anuales meteorológicas durante el periodo en estudio.

Estación	LAT (°N)	LON (°W)	ALT (msnm)	Años de registro	Promedio anual			
					T_{med} (°C)	RSG MJ $M^{-2}d^{-1}$	HR (%)	U_2 (ms^{-1})

Tecnología y ciencias del agua, ISSN 2007-2422, 12(3), 32-81. DOI: 10.24850/j-tyca-2021-03-02

Calakmul	18.365	89.893	28	2000-2018	26.20	15.66	81.00	1.16
Campeche	19.836	90.507	3	2000-2018	26.80	20.19	79.90	1.82
Cd del Carmen	18.658	91.765	4	2011-2018	27.10	19.40	75.16	2.36
Escárcega	18.608	90.754	60	2004-2018	27.17	18.74	79.2	1.51
Los Petenes	19.943	90.374	2	2012-2018	26.52	14.43	80.23	1.31
Monclova	18.057	90.821	100	2008-2018	26.70	18.77	72.59	1.63

LAT: latitud; LON: longitud; ALT: altitud; T_{med}: temperatura media; RSG: radiación solar global; HR: humedad relativa; U_2: velocidad del viento.

Manejo de datos faltantes y calidad

Las bases de datos fueron procesadas cada 10 minutos, detectando mediante algoritmos implementados en el *software Microsoft Excel*® series de tiempo faltantes, las cuales posteriormente fueron rellenadas utilizando la técnica de interpolación llamada interpolación polinómica de Hermite (PCHIP). Para una descripción más detallada, consultar Salazar, Ureña y Gallego (2010), y Torrente-Cantó (2018). Una vez que las series de datos fueron completadas, se construyeron bases de datos diarias. Asimismo, los datos se analizaron para identificar valores atípicos, donde aquellos valores por encima de tres desviaciones estándar de la media se marcaron como atípicos. Se analizaron los datos señalados como atípicos:

Tecnología y ciencias del agua, ISSN 2007-2422, 12(3), 32-81. DOI: 10.24850/j-tyca-2021-03-02

si un valor atípico extremo inferior se encontraba asociado con un evento de lluvia no se eliminaba; en caso contrario, se eliminaron; esto, con el objetivo de tener modelos funcionales incluso en época lluviosa. En el caso de los valores atípicos extremos superiores fueron eliminados. En ambos casos se utilizó la técnica de interpolación PCHIP para su relleno. La Figura 2 muestra el gráfico de caja y bigotes de las estaciones meteorológicas implicadas en los modelos de este estudio. El bigote representa el mínimo y el máximo de las variables. Respecto a la temperatura máxima, la media varía entre 33 y 34 °C, con valores máximos entre 42 y 43 °C (excepto la estación de Cd. del Carmen); el valor mínimo de la temperatura máxima varía entre 22 y 25 °C. La temperatura mínima tiene una media entre 20 y 24 °C, con valores máximos entre 25 y 28 °C, y mínimos entre 12 y 17 °C. La Figura 2 también muestra los valores atípicos, generalmente asociados con la temperatura mínima.

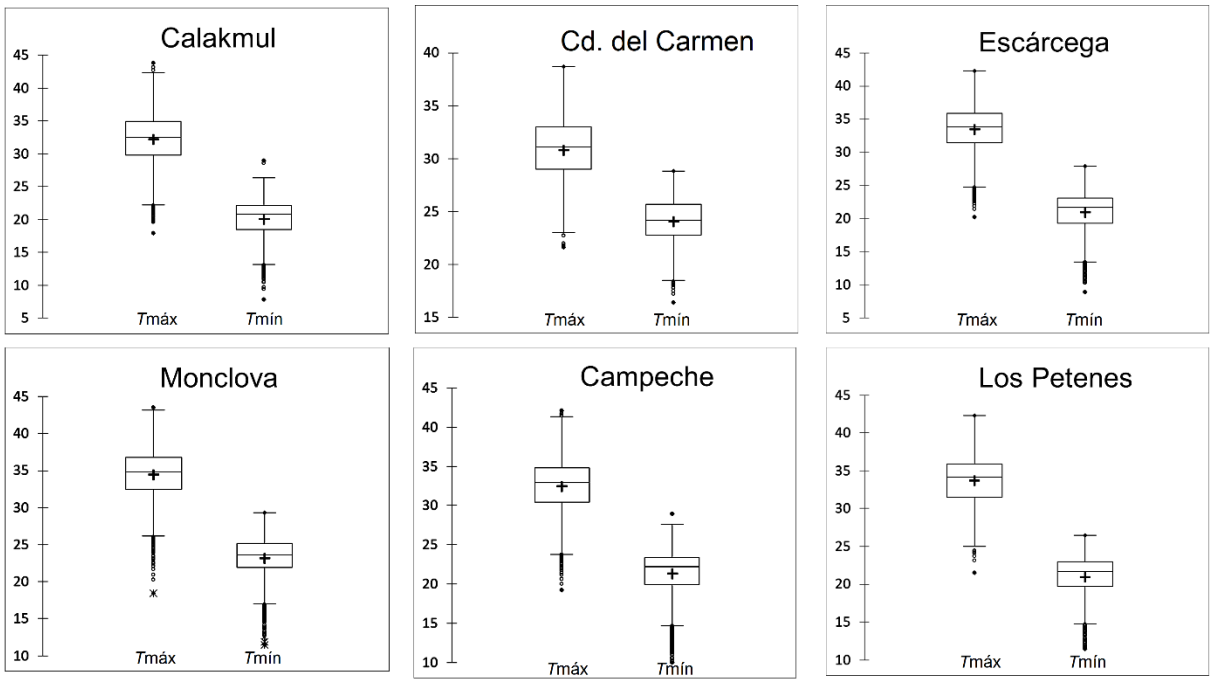

Figura 2. Gráficos de caja y bigotes de variable temperatura en las estaciones meteorológicas analizadas.

Ecuación de la FAO 56 PM (ETo-FAO56PM)

La ecuación FAO56PM es el modelo estándar usado para estimar con precisión la *ETo*, propuesto por la Organización de las Naciones Unidas para la Agricultura y la Alimentación (FAO, por sus siglas en inglés). Incorpora aspectos termodinámicos y aerodinámicos, y considera muchos parámetros meteorológicos relacionados con el proceso de

evapotranspiración, como la radiación neta, la temperatura del aire, el déficit de presión de vapor, la velocidad del viento. Ha demostrado ser un método relativamente preciso bajo diferentes condiciones o regiones (Allen *et al*., 1998). Estos aspectos se han incorporado en la siguiente ecuación:

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

donde R_n = radiación neta en la superficie (MJ m^{-2} $día^{-1}$); G = flujo del calor de suelo (MJ m^{-2} $día^{-1}$); T_{med} = temperatura media del aire a 2 m de altura (°C); u_2 = 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); Δ = pendiente de la curva de presión de vapor (kPa $°C^{-1}$); γ = constante psicrométrica (kPa $°C^{-1}$).

En este estudio, el método ETo-FAO56PM se usó para evaluar los métodos convencionales y de inteligencia artificial.

Ecuación de Hargreaves y Samani

Tecnología y ciencias del agua, ISSN 2007-2422, 12(3), 32-81. DOI: 10.24850/j-tyca-2021-03-02

El modelo de HS se considera 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). La Ecuación (2) del modelo de Hargreaves y Samani está estructurada de la siguiente manera:

$$ETo = 0.408\ K_{HG}\ (T_{med}\ +\ 17.8)(T_{max}\ -\ T_{min})^{0.5} Ra \qquad (2)$$

donde *ETo* = evapotranspiración de referencia (mm día^{-1}); K_{HG} = es un coeficiente empírico, que inicialmente fue establecido en 0.0023, pero se ha recalibrado acorde con el lugar empleado; T_{med} = temperatura media; $T_{\text{máx}}$ = temperatura máxima; $T_{\text{mín}}$ = temperatura mínima; R_a = radiación solar extraterrestre. R_a se calculó en función del día del año, latitud del sitio y ángulo solar de acuerdo con la ecuación propuesta por Allen *et al.* (1998).

Ecuación de Camargo

El modelo de Camargo es una modificación a la ecuación de Thornthwaite (TH). 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 media efectiva (T_{ef}) (Camargo, Marin, Sentelhas, & Picini, 1999). La Ecuación (3) del modelo de Camargo está estructurada de la siguiente manera:

$$Eto = K_{CA1} * (10 * (K_{CA2} * (3T_{max} - T_{min}))/I)^{a} * N/360 \quad (3)$$

donde *ETo* = evapotranspiración de referencia (mm día^{-1}); K_{CA1} y K_{CA2} = coeficientes empíricos, donde sus valores originales son 16 y 0.36, respectivamente, y deben ser calibrados acorde con el lugar de empleo; I = índice de calor anual; a = exponente empírico en función de I; N = tiempo máximo de insolación en horas; (K_{CA2} * ($3T_{máx}$ - $T_{mín}$) = temperatura efectiva, reemplazando a la temperatura media en la ecuación de Thornthwaite. El valor de I se define como la suma de 12 valores de índices de calor mensual, como se muestra en la siguiente ecuación:

$$I = \sum_{n=1}^{12} \left(T_{medj/5}\right)^{1.514} \quad (4)$$

donde T_{medj} = temperatura media mensual (°C).

y:

$$a = 6.751 * 10^{-7} * I^3 - 7.711 * 10^{-5} * I^2 + 1.792 * 10^{-2} * I + 0.492 \quad (5)$$

El valor de *a* varía de 0 a 4.25; mientras que el índice de calor anual *I* varía de 0 a 160.

Ajuste de parámetros en los métodos convencionales

Los métodos convencionales basados en temperatura deben ajustarse a las condiciones locales antes de ser utilizados (Almorox *et al*., 2018) para obtener buenas estimaciones de *ETo*. Por tal motivo, los coeficientes originales de las ecuaciones se calibraron utilizando técnicas de regresión no lineal mediante al algoritmo Levenberg-Marquardt.

Máquinas de soporte vectorial (SVM)

La técnica máquinas de soporte vectorial (SVM) fue desarrollada por Vapnik (2000) y es uno de los enfoques basado 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; selecciona un hiperplano único de separación de cada clase y 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, llamados vectores de soporte. 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, 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 $\varphi(x)$ para después encontrar una función de regresión (Fan *et al.*, 2018a; Mehdizadeh, Behmanesh, & Khalili, 2017; Quej, Almorox, Arnaldo, & Saito, 2017; Topi & Vanita, 2017; Wen *et al*., 2015):

$$f\ (x)\ =\ \omega\varphi(x)\ +\ b \tag{6}$$

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

$$R\ (C) = C\ \sum_i^N L_\varepsilon\ (f\ (\ \chi_i),y_i)\ \frac{1}{2}\ \|\omega\|^2 \tag{7}$$

Tecnología y ciencias del agua, ISSN 2007-2422, 12(3), 32-81. DOI: 10.24850/j-tyca-2021-03-02

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; (ε) es la función de pérdida de Vapnik (tamaño del tubo del modelo de SVM) y está definida como:

$$L_\varepsilon\,(f\,(\chi_i),y_i) = \begin{cases} 0 \text{ for} |f\,(x_i) - y_i| \leq \varepsilon \\ |f\,(x_i) - y_i| - \varepsilon \text{ otra manera} \end{cases} \quad (8)$$

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 la detección de valores atípicos, las variables de holgura ξ y ξ^* miden de arriba y abajo en el tubo de ε.

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

$$R\,(\xi,\xi^*,\omega,b) = \frac{1}{2}\|\omega\|^2 + C\,\sum_{i=1}^{n}(\xi_i + \xi_i^*) \quad (9)$$

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

Tecnología y ciencias del agua, ISSN 2007-2422, 12(3), 32-81. DOI: 10.24850/j-tyca-2021-03-02

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

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. 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):

$$K\left(x_i, x_j\right) = \exp\left(-\gamma \left\|x_i - x_j\right\|^2\right), \gamma > 0 \qquad (10)$$

Implementación del modelo SVM en la estimación de la *ETo*

En el presente estudio, el modelo SVM para estimar la *ETo* se construyó usando el *software R* (RDevelopment, 2009). Como variables de entrada se usaron los datos meteorológicos de $T_{máx}$, $T_{mín}$ y R_a, y como variable objetivo se utilizaron los valores de ETo-FAO56PM (Ecuación (1)). Para el entrenamiento y validación del modelo SVM, se utilizó el *software R* en

Tecnología y ciencias del agua, ISSN 2007-2422, 12(3), 32-81. DOI: 10.24850/j-tyca-2021-03-02

conjunto con el paquete *LIBSVM 3.1* (Chang, Lin, & Tieleman, 2013). Para la redimensión de los datos se utilizó la función Kernel de Base Radial (RBF) (Ecuación (10)). Con la finalidad de evitar el sobre ajuste y aumentar el desempeño del modelo SVM para estimar la *ETo*, los parámetros ε, *C* y ɣ de la SVM, y de la función Kernel de base radial se optimaron mediante el algoritmo genético (GA), utilizando validación cruzada (*VC* = 5 fólders) (Quej *et al*., 2017; Shrestha & Shukla, 2015), y variando los parámetros ε = 0.002 a ε = 2, *C* = 0.0001 a *C* = 10, y ɣ = 0.0001 a ɣ = 2. El GA se implementó utilizando el *software R* en conjunto con la librería *e1079* y *Caret*. Se utilizó el 60 % de los datos durante la etapa de entrenamiento y el 40 % en la etapa de validación.

Los parámetros optimizados por GA utilizados en el entrenamiento de la SVM se muestran en la Tabla 2.

Tabla 2. Parámetros SVM óptimos obtenidos mediante GA.

Estación/modelo ID	**Valores óptimos**		
	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

Programación de expresión genética (GEP)

La programación de la expresión génica (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 en la programación por computadora para resolver un problema definido por el usuario.

Los GEP se consideran un híbrido 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 con las variables o constantes.
- Funciones. Tienen descendientes, se asocian con operadores del algoritmo que se desea desarrollar.

En GEP, los individuos se codifican primero como cadenas lineales de longitud fija como en 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

Tecnología y ciencias del agua, ISSN 2007-2422, 12(3), 32-81. DOI: 10.24850/j-tyca-2021-03-02

dependiente (Mehdizadeh *et al*., 2017), creando funciones de manera aleatoria y seleccionando aquellas que presenten un mejor ajuste a los resultados experimentales, permitiendo la generación de algoritmos y expresiones matemáticas de manera automática para la solución de problemas (Mattar, 2018; Shiri, 2017).

Implementación del modelo GEP en la estimación de la *ETo*

En este estudio, la implementación de la técnica GEP se llevó a cabo utilizando el *software GenexprooTols* v. 5.0. Las variables de entrada son los valores de los datos meteorológicos de $T_{máx}$, $T_{mín}$, R_a y valores de *ETo*-FAO56PM como variable objetivo. Los operadores aritméticos y funciones matemáticas implementadas dentro del programa fueron $\{+,-,\times,\div, \sqrt{x}, \sqrt[3]{x}, x^2, x^3, In(x), e^x, \sin(x), \cos(x), Arctan(x)\}$, recomendados para estudios hidrológicos (Mattar, 2018; Mehdizadeh *et al*., 2017; Shiri, 2017). El 70 % de los datos se usaron para la etapa de entrenamiento y el 30 % para la validación, utilizando validación cruzada (*VC* = 5 fólders) para evitar el sobreajuste. Los parámetros GEP utilizados en el presente estudio se muestran en la Tabla 3 (Shiri *et al*., 2014).

Tabla 3. Parámetros del modelo GEP.

Parámetro	**Valor**
Número 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*).

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 modelo se basa en la teoría del aumento de gradiente, 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, evitando un sobreajuste y optimizando los recursos de cómputo. Esto se obtiene simplificando todas las funciones que permitan combinar términos predictivos y de regularización, pero que, a su vez, mantenga una velocidad computacional óptima durante todo el procesamiento. 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 (minimización de la raíz cuadrada media del error). 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" (Fan *et al*., 2018a). Bajo este contexto el modelo XGBoost se basa en la siguiente función objetivo: *pérdida* + *regularizador*:

Tecnología y ciencias del agua, ISSN 2007-2422, 12(3), 32-81. DOI: 10.24850/j-tyca-2021-03-02

$$Obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{i=1}^{t} \Omega\ (f_i) \qquad (11)$$

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:

$$Obj^{(t)} = -\frac{1}{2} \sum_{i=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma \mathrm{T} \qquad (12)$$

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, y *T* es el número de hojas en un árbol. 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, Antonanzas, Antonanzas-Torres, & Martinez-De-Pison, 2017).

Implementación del modelo XGBoost en la estimación de la *ETo*

Para la implementación del modelo XGBoost, como primer paso se optimizaron los hiperparámetros nrounds, max_depth, eta, gamma, colsample bytree, min_child_weight y subsample (Tabla 4) utilizando la librería *Caret* del *Software R* (Fan *et al.*, 2018a); segundo, se ajustó el modelo XGBoost utilizando la librería "Xgboost" del *Software R*, utilizando validación cruzada (*VC* = 5 fólders) para evitar el sobreajuste. El 70 % de los datos se usó para la etapa de entrenamiento y el 30 % para la validación. Las variables de entrada al modelo son los valores de los datos meteorológicos de $T_{máx}$, $T_{mín}$ y R_a, los valores de *ETo*-FAO56PM como variable objetivo.

Tabla 4. Hiperparámetros XGBoost optimizados.

Parámetros ajustados	**Calak-mul**	**Campe-che**	**Cd. del Car-men**	**Escár-cega**	**Moncl-va**	**Los Pete-nes**
Nrounds	50	50	50	50	150	50
Max_depth	2	3	3	3	2	3

Tecnología y ciencias del agua, ISSN 2007-2422, 12(3), 32-81. DOI: 10.24850/j-tyca-2021-03-02

Eta	0.3	0.3	0.3	0.3	0.3	0.3
Gamma	0	0	0	0	0	0
Colsample bytree	0.8	0.8	0.8	0.8	0.8	0.8
Min_child_weight	1	1	1	1	1	1
subsample	1	1	1	1	0.75	1

Análisis estadístico

En el presente estudio se utilizaron cuatro indicadores estadísticos para evaluar el desempeño de los modelos convencionales y de inteligencia artificial, estos indicadores son coeficiente de determinación (R^2; Ecuación (13)); raíz cuadrada media del error (RMSE; Ecuación (14)); error absoluto medio (MAE; Ecuación (15)); error medio de sesgo (MBE; Ecuación (16)):

$$R^2 = \frac{\left[\sum_{i=1}^{n}(P_i - P_{prom})(O_i - O_{prom})\right]^2}{\sum_{i=1}^{n}(P_i - P_{prom})^2 \sum_{i=1}^{n}(O_i - O_{prom})^2} \quad (13)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2} \quad (14)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(|P_i - O_i|) \quad (15)$$

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

donde n es el número de comparaciones; P_i y O_i, valores estimados y observados de *ETo*-FAO56PM, respectivamente; P_{prom}, el promedio de los valores estimados de *ETo*; O_{prom}, el promedio de los valores observados de *ETo*; las unidades de *ETo* se encuentran en mm d^{-1}.

R^2 se usa por lo común para estimar el desempeño de modelos hidrológicos. Representa la fracción de los valores estimados que son los más cercanos a la línea de datos de medición. Valores del coeficiente de determinación cercanos a 1 indican modelos más eficientes y la línea de regresión se ajusta mejor a los datos. EL RMSE es una medida utilizada con frecuencia para comparar errores de predicción en diferentes modelos; cuanto menor sea su valor, mejor será la capacidad predictiva de un modelo en términos de su desviación absoluta. El MAE es la suma de los valores absolutos de los errores divididos por el número de observaciones; se usa con frecuencia para medir qué tan cerca están los valores estimados de los valores observados. El MBE proporciona información sobre la tendencia del modelo a sobreestimar o subestimar la variable; cuantifica el error sistemático del modelo.

Tecnología y ciencias del agua, ISSN 2007-2422, 12(3), 32-81. DOI: 10.24850/j-tyca-2021-03-02

La evaluación de la incertidumbre de los indicadores estadísticos R^2, RMSE, MAE y MBE se realizó mediante la construcción de intervalos de confianza *bootstrap* (ICB) al 95 % de nivel de confianza. Para tal propósito se utilizó el método no paramétrico *bootstrap* percentil como técnica de remuestreo, utilizando B = 10 000 réplicas con remplazamiento, con el propósito de inducir mayor precisión en las estimaciones (Efron, 1992).

Los ICB ofrecen una manera de estimar con alta probabilidad un rango de valores en el que se encuentra el valor del parámetro (indicador estadístico).

También se computó el error estándar de la distribución (*ee*) y el valor de cada indicador estadístico *bootstrap*, calculando la desviación estándar y la media de las B réplicas.

Resultados

En el presente estudio se evaluaron dos ecuaciones convencionales y tres técnicas de inteligencia artificial para estimar la *ETo* utilizando la variable de temperatura del aire.

En la Tabla 5 se observan los índices estadísticos obtenidos mediante *bootstrap* (media de las B réplicas *bootstrap*). Entre las ecuaciones convencionales evaluadas, el modelo de Camargo mostró

mejores resultados globales (R^2 = 0.734; MAE = 0.564; RMSE = 0.721; MBE = -0.008) respecto a la ecuación de HS (R^2 = 0.727; MAE = 0.588; RMSE = 0.750; MBE = -0.032), siendo la estación de Monclova donde presentó mejor desempeño (R^2 = 0.815; MAE = 0.488; RMSE = 0.608; MBE = -0.011). De manera global, la ecuación de Camargo tuvo una tendencia a subestimar ligeramente valores de *ETo* según el indicador MBE = -0.008. Respecto a la ecuación de Camargo, el coeficiente 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 de 0.0015 a 0.0027.

Tabla 5. Índices estadísticos *bootstrap* (R^2, MAE, RMSE y MBE) de los modelos convencionales y de inteligencia artificial usados para la estimación de la *ETo* de cada estación meteorológica.

Estación/ modelo	R^2	MAE (mm d^{-1})	RMSE (mm d^{-1})	MBE (mm d^{-1})	K_{HS}	K_{CA1}	K_{CA2}
Calakmul							
HS	0.700	0.569	0.729	-0.055	0.0015		
Camargo	0.712	0.534	0.688	0.028		36.071	0.200
SVM	0.740	0.486	0.646	0.046			
GEP	0.696	0.544	0.719	-0.003			
XGBoost	0.771	0.467	0.607	-0.0004			
Campeche							
HS	0.703	0.550	0.709	-0.010	0.0020		
Camargo	0.635	0.623	0.797	0.037		40.256	0.218

Tecnología y ciencias del agua, ISSN 2007-2422, 12(3), 32-81. DOI: 10.24850/j-tyca-2021-03-02

SVM	0.731	0.519	0.680	-0.003			
GEP	0.695	0.561	0.726	0.036			
XGBoost	0.695	0.543	0.721	-0.012			
Cd. del Carmen							
HS	0.694	0.633	0.820	-0.002	0.0027		
Camargo	0.702	0.627	0.811	0.005		44.476	0.240
SVM	0.742	0.585	0.778	-0.056			
GEP	0.721	0.638	0.809	-0.034			
XGBoost	0.703	0.611	0.802	0.032			
Escárcega							
HS	0.711	0.654	0.825	-0.050	0.0018		
Camargo	0.783	0.554	0.705	-0.036		38.581	0.211
SVM	0.838	0.471	0.608	0.034			
GEP	0.773	0.561	0.713	0.030			
XGBoost	0.815	0.500	0.642	-0.043			
Monclova							
HS	0.796	0.523	0.651	0.014	0.0020		
Camargo	0.815	0.488	0.608	-0.011		39.391	0.214
SVM	0.852	0.426	0.531	-0.046			
GEP	0.816	0.500	0.618	0.009			
XGBoost	0.842	0.442	0.569	0.025			
Los Petenes							
HS	0.755	0.599	0.767	-0.089	0.0014		
Camargo	0.754	0.555	0.716	-0.071		34.922	0.195

SVM	0.801	0.392	0.574	-0.042			
GEP	0.812	0.406	0.574	-0.022			
XGBoost	0.804	0.414	0.584	0.041			
Todas las estaciones							
HS	0.727	0.588	0.750	-0.032			
Camargo	0.734	0.564	0.721	-0.008			
SVM	0.784	0.480	0.636	-0.011			
GEP	0.752	0.535	0.693	0.003			
XGBoost	0.772	0.496	0.654	0.007			

* En los modelos de inteligencia artificial, los índices estadísticos *bootstrap* corresponden a los obtenidos en el proceso de validación.

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

Respecto a los modelos de inteligencia artificial para estimar la *ETo*, la Tabla 5 muestra que el modelo SVM obtuvo el mejor desempeño con relación a los otros modelos evaluados, obteniendo valores de R^2 = 0.784, MAE = 0.480, RMSE = 0.636 y MBE = -0.011, siendo la estación de Monclova donde se presentó mejor desempeño (R^2 = 0.852; MAE = 0.426; RMSE = 0.531; MBE = -0.046). Le sigue el modelo XGBoost con resultados de R^2 = 0.772, MAE = 0.496, RMSE = 0.654, y MBE = 0.007, donde el mejor desempeño fue en la estación de Monclova (R^2 = 0.842; MAE = 0.442; RMSE = 0.569; MBE = 0.025). El modelo GEP fue el de menor rendimiento comparado con los otros modelos de inteligencia artificial, obteniendo resultados de R^2 = 0.752, MAE = 0.535, RMSE =

Tecnología y ciencias del agua, ISSN 2007-2422, 12(3), 32-81. DOI: 10.24850/j-tyca-2021-03-02

0.693 y MBE = 0.003; sin embargo, su rendimiento fue superior a lo obtenido por las ecuaciones convencionales.

El modelo SVM de manera general presentó una tendencia a subestimar valores de *ETo* de acuerdo con un valor global de MBE = -0.011, mientras que los modelos GEP y XGBoost presentaron una leve tendencia a sobreestimar valores de *ETo*.

El modelo SVM tiene buen rendimiento cuando se realiza el ajuste de los parámetros de Costo, Gamma y Épsilon utilizando el algoritmo genético. Asimismo, al usar validación cruzada se evita el sobreajuste del modelo.

Una de las principales ventajas del método SVM sobre los demás métodos radica en que el problema no lineal siempre convergerá en un mínimo global.

Por otra parte, una característica útil de la técnica GEP es que proporciona una expresión algebraica para estimar la *ETo*, que puede programarse en una hoja de cálculo, *software R*, *Matlab* o *Python*. En la Tabla 6 se presentan las expresiones algebraicas obtenidas por el modelo GEP para las seis estaciones meteorológicas.

Tabla 6. Expresiones algebraicas obtenidas por el modelo GEP para cada estación meteorológica.

Estación	Expresión matemática

Tecnología y ciencias del agua, ISSN 2007-2422, 12(3), 32-81. DOI: 10.24850/j-tyca-2021-03-02

Calakmul	$ETo = \frac{T_{máx}}{\sqrt[3]{(0.815 * T_{máx}}} + \frac{\text{Arctan}\,(T_{máx}) * (-9.223)}{\text{Arctan}\,(Ra - 5.126)} + \frac{Ra}{\text{Arctan}\,(T_{máx} - T_{mín} - \log(Ra + T_{mín})}$
Campeche	$ETo = exp\left[\cos(\sqrt{T_{máx}})^9\right] + exp\left[\sqrt{\cos\left(\sqrt{\frac{1.707}{T_{mín}} + Ra}\right)^3}\right] + \cos\left[\frac{\left(\frac{1.707}{T_{mín}}\right) + T_{máx}}{\sqrt{(T_{máx})^3}}\right]$
Cd. del Carmen	$ETo = \sqrt{\left(\frac{Ra - (T_{máx} + T_{mín}) * sin(T_{máx})}{Ra}\right)} + Arctan\left[(-3.589 * 8.418) + T_{máx} - \sqrt{T_{mín}} + log(T_{máx} - 7.291)\right] + sin\left[\frac{(Ra * -3.589) - T_{min}}{T_{máx} - \sqrt[3]{7.291 + T_{mín}}}\right]$
Escárcega	$ETo = \log\left(\frac{7.347}{\sqrt[27]{T_{mín}}}\right) + \left(\frac{Ra * \sqrt[3]{6.801}}{7.347 - 9.023 + T_{máx}}\right) + \left(\frac{Ra * \sqrt{T_{máx} - T_{mín}}}{7.347^2 - T_{máx}}\right)$
Monclova	$ETo = -13.981 + \sqrt[3]{2Ra} + \frac{T_{máx}}{\sqrt[3]{\frac{\frac{T_{máx}}{T_{mín}}}{0.611 + Ra}}} + \sqrt[3]{\sqrt[6]{Ra} + \sqrt[3]{-3.443 + Ra}}$
Los Petenes	$ETo = \log(\log(\log(4.244 + T_{máx}) + T_{máx}^{27} + \left(\frac{\sqrt[3]{2Ra}}{\log\left(\frac{T_{máx}}{T_{mín}}\right)}\right) + \log\left(\frac{\frac{Ra}{4.244^3} * Ra}{T_{mín}}\right)$

Como un ejemplo práctico del modelo GEP, se proporciona la fórmula en formato de Microsoft Excel®, cuyas entradas corresponden a las variables de $T_{mín}$, $T_{máx}$, R_a:

Tecnología y ciencias del agua, ISSN 2007-2422, 12(3), 32-81. DOI: 10.24850/j-tyca-2021-03-02

$B2 / ((M2*T_{máx}) \text{ ^ } (1/3)) + (\text{ATAN}(T_{máx})*M3) / (\text{ATAN}(R_a\text{-}M4)) + R_a / (\text{ATAN}((T_{máx}\text{-}T_{mín}) - \text{LOG}(R_a)) + T_{mín})$

donde $M2 = 0.0815$; $M3 = 9.223$; $M4 = 5.126$ son constantes en el modelo.

Así se verifica que al ingresar los valores de $T_{máx} = 31.7$, $T_{mín} = 18.40$ y $R_a = 38.87$, obtendremos el valor de la $ETo = 3.407$ mm d^{-1}.

La Tabla 7 muestra los intervalos de confianza *bootstrap* (ICB) al 95 % de nivel de confianza y el error estándar de la distribución (EE) de los índices estadísticos R^2, MAE, RMSE y MBE de los modelos convencionales y de inteligencia artificial. De manera general, los intervalos de confianza muestran una amplitud reducida, relacionada con error estándar, lo cual indica que si utilizamos muestras aleatorias y se determinan sus indicadores estadísticos de evaluación, éstos variarán en un rango siempre aceptable.

Tabla 7. ICB (límite inferior [LI] = 2.5 % y límite superior [LS] = 97.5 %) al 95 % de nivel de confianza, y el error estándar de la distribución (*ee*) de los índices estadísticos R^2, MAE, RMSE y MBE de los modelos convencionales y de inteligencia artificial.

Estación / modelo	R^2			MAE (mm d^{-1})			RMSE (mm d^{-1})			MBE (mm d^{-1})		
	LI	LS	*ee*	LI	LS	*ee*	LI	LS	*ee*	LI	LS	*ee*
Calakmul												

Tecnología y ciencias del agua, ISSN 2007-2422, 12(3), 32-81. DOI: 10.24850/j-tyca-2021-03-02

HS	0.683	0.717	0.008	0.554	0.584	0.007	0.709	0.748	0.009	-0.079	-0.031	0.012
Camargo	0.696	0.727	0.007	0.519	0.548	0.007	0.669	0.706	0.009	0.006	0.049	0.011
SVM	0.715	0.765	0.013	0.462	0.511	0.012	0.615	0.677	0.016	0.009	0.083	0.019
GEP	0.667	0.724	0.014	0.516	0.572	0.014	0.682	0.757	0.019	-0.046	0.039	0.022
XGBoost	0.757	0.786	0.007	0.453	0.482	0.007	0.589	0.626	0.009	-0.023	0.023	0.012
Campeche												
HS	0.688	0.717	0.007	0.537	0.562	0.006	0.692	0.726	0.008	-0.029	0.009	0.009
Camargo	0.617	0.652	0.008	0.609	0.636	0.007	0.778	0.815	0.009	0.014	0.059	0.011
SVM	0.707	0.754	0.011	0.497	0.54	0.011	0.652	0.708	0.014	-0.036	0.029	0.017
GEP	0.669	0.72	0.013	0.537	0.584	0.012	0.694	0.758	0.016	-0.001	0.073	0.018
XGBoost	0.662	0.727	0.016	0.517	0.569	0.013	0.682	0.759	0.019	-0.052	0.028	0.021

Tecnología y ciencias del agua, ISSN 2007-2422, 12(3), 32-81. DOI: 10.24850/j-tyca-2021-03-02

Cd. del Carmen												
HS	0.662	0.726	0.016	0.603	0.662	0.015	0.400	1.241	0.214	-0.049	0.045	0.024
Camargo	0.671	0.732	0.016	0.597	0.656	0.015	0.770	0.852	0.021	-0.041	0.052	0.023
SVM	0.692	0.792	0.025	0.534	0.637	0.026	0.700	0.856	0.039	-0.134	0.021	0.039
GEP	0.675	0.766	0.023	0.585	0.691	0.027	0.742	0.875	0.034	-0.118	0.050	0.043
XGBoost	0.639	0.767	0.032	0.551	0.672	0.031	0.718	0.886	0.043	-0.061	0.125	0.047
Escárcega												
HS	0.695	0.727	0.008	0.635	0.673	0.009	0.802	0.847	0.012	-0.082	-0.018	0.016
Camargo	0.769	0.796	0.007	0.537	0.570	0.008	0.685	0.724	0.010	-0.063	-0.009	0.014
SVM	0.817	0.859	0.011	0.445	0.496	0.013	0.574	0.642	0.017	-0.006	0.074	0.021
GEP	0.745	0.800	0.014	0.530	0.592	0.015	0.677	0.749	0.018	-0.018	0.079	0.025

XGBoost	0.791	0.838	0.012	0.470	0.531	0.015	0.606	0.678	0.018	-0.093	0.006	0.025
Monclova												
HS	0.781	0.811	0.007	0.507	0.539	0.008	0.632	0.670	0.009	-0.058	-0.004	0.014
Camargo	0.802	0.829	0.006	0.473	0.503	0.007	0.590	0.626	0.009	-0.036	0.014	0.013
SVM	0.832	0.871	0.009	0.403	0.449	0.012	0.503	0.558	0.014	-0.084	-0.007	0.019
GEP	0.792	0.839	0.012	0.471	0.528	0.014	0.585	0.651	0.016	-0.037	0.056	0.024
XGBoost	0.818	0.866	0.012	0.412	0.472	0.015	0.532	0.605	0.018	-0.022	0.072	0.024
Los Petenes												
HS	0.710	0.800	0.023	0.564	0.633	0.017	0.715	0.818	0.026	-0.144	-0.035	0.027
Camargo	0.710	0.798	0.022	0.523	0.588	0.016	0.665	0.767	0.026	-0.122	-0.019	0.026
SVM	0.724	0.878	0.039	0.338	0.446	0.027	0.463	0.686	0.057	-0.114	0.030	0.037

GEP	0.738	0.886	0.037	0.352	0.460	0.027	0.463	0.684	0.056	-0.097	0.054	0.038
XGBoost	0.742	0.866	0.032	0.353	0.473	0.030	0.494	0.674	0.046	-0.042	0.123	0.042

Discusión

Se evaluaron dos modelos convencionales y tres técnicas de inteligencia artificial para estimar valores de *ETo*, siendo las variables de entrada para todos los modelos los datos de $T_{máx}$, $T_{mín}$ y R_a, excepto para el modelo de Camargo, que utiliza las horas máximas de insolación para un determinado sitio; de esta manera, los modelos tienen un ámbito de uso espacial y temporal delimitado por la amplitud térmica.

En un caso de clima árido y súper húmedo donde existe una amplitud térmica mayor, Camargo *et al*. (1999) presentaron una modificación al método de Thornthwaite usando el término "temperatura efectiva" $Tef = 0.36\ (3\ T_{máx} - T_{mín})$, 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

Tecnología y ciencias del agua, ISSN 2007-2422, 12(3), 32-81. DOI: 10.24850/j-tyca-2021-03-02

la península de Yucatán, México, para estimar la *ETo* cuando solo existen datos de temperatura, lo anterior debido a que en las estaciones estudiadas existe mayor amplitud térmica. Asimismo, la calibración del coeficiente K_{HS} de la ecuación de HS coincide con el obtenido por Bautista, Bautista y Delgado-Carranza (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 Cd. del 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). Similar resultado obtuvieron Quej, Almorox, Arnaldo y Moratiel (2019) al evaluar la *ETo* diaria utilizando la ecuación de HS y Camargo, obteniendo valores de RMSE de 0.70 y 0.80 mm d^{-1}, respectivamente. También en un estudio regional realizado por Kelso-Bucio *et al*. (2013) se calibró el coeficiente empírico de la ecuación de Hargreaves, y se consiguieron valores de RMSE en el rango de 0.68 a 0.87 mm d^{-1} para la región norte y centro del estado de Campeche, México, valores similares a este estudio, donde el valor de RMSE en la ecuación de HS tuvo una variación de 0.651 a 0.820 mm d^{-1}.

Entre las técnicas de inteligencia artificial evaluadas para estimar la *ETo*, el modelo SVM utilizando el kernel de base radial presentó 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 semiáridas 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 convencionales 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 inteligencia artificial y que las ecuaciones convencionales. Por otra parte, los resultados de las técnicas SMV y XGBoost coinciden con un estudio realizado en China por Fan *et al*. (2018b), donde evaluaron algunos métodos de inteligencia artificial, SVM y XGBoost entre ellos. Los resultados obtenidos en el clima subhúmedo utilizando solo datos de temperatura del aire y con la técnica XGBoost obtuvo un valor de RMSE = 0.723 mm d^{-1} y con la técnica SVM un valor de RMSE = 0.717 mm d^{-1}.

Finalmente, la técnica XGBoost obtuvo un desempeño muy cercano a SVM. Sin embargo, se recomienda el uso de SVM de base radial al ser una técnica más robusta por sus fuertes bases matemáticas, con menos gasto computacional al no usar todos los datos para el cálculo sino solo algunos llamados vectores de soporte, además es una técnica con menos tendencia al sobreajuste una vez que sus parámetros han sido ajustados vía algoritmo GA.

Conclusiones

Tecnología y ciencias del agua, ISSN 2007-2422, 12(3), 32-81. DOI: 10.24850/j-tyca-2021-03-02

De las ecuaciones convencionales 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 de la región del estudio. 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 inteligencia artificial, el modelo SVM de base radial se recomienda para realizar estimaciones de *ETo*.

El ajuste previo de los parámetros de los modelos de inteligencia artificial mediante algoritmos es fundamental para evitar un sobreajuste 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; como se comprobó en el presente estudio, el modelo GEP superó ligeramente a los modelos convencionales.

Los modelos de inteligencia artificial son una excelente opción para estimar valores de *ETo* al superar a las ecuaciones convencionales; sin embargo, para su implementación se requiere de un conocimiento especializado en el uso de *software* y ejecución de códigos de programación.

Los modelos evaluados en este estudio se pueden utilizar en regiones de clima cálidos subhúmedos y en los rangos de temperatura señalados en la Figura 2.

En futuros trabajos se podría evaluar el efecto de la humedad relativa, velocidad del viento en la estimación de la *ETo* para diferentes meses del año y bajo condiciones extremas de lluvia.

Para la implementación de los modelos de inteligencia artificial descritos en este estudio, se recomienda el uso del *software* libre *R*. En caso de requerir los códigos para su implementación, se pueden solicitar vía correo electrónico al autor para correspondencia de este artículo.

Agradecimientos

Al Consejo Nacional de Ciencia y Tecnología (Conacyt) y al Colegio de Postgraduados Campus Campeche por el apoyo financiero para realizar los estudios de maestría, de la que se deriva esta investigación.

Referencias

Allen, R. G., Pereira, L. S., Raes, D., & Smith, M. (1998). Crop evapotranspiration. Guidelines for computing crop water requirements. FAO Irrigation and drainage paper 56. *Irrigation and Drainage*, 300(9), D05109.

Almorox, J., Senatore, A., Quej, V. H., & Mendicino, G. (2018). Worldwide assessment of the Penman-Monteith temperature approach for the estimation of monthly reference evapotranspiration. *Theoretical and Applied Climatology*, 131(1-2), 693-703. Recuperado de https://doi.org/10.1007/s00704-016-1996-2

Antonopoulos, V. Z., & Antonopoulos, A. V. (2017). Daily reference

evapotranspiration estimates by artificial neural networks technique and empirical equations using limited input climate variables. *Computers and Electronics in Agriculture*, 132, 86-96. Recuperado de https://doi.org/10.1016/j.compag.2016.11.011

Bautista, F., Bautista, D., & Delgado-Carranza, C. (2009). Calibration of the equations of Hargreaves and Thornthwaite to estimate the potential evapotranspiration in semi-arid and subhumid tropical climates for regional applications. *Atmósfera*, 22(4), 331-348.

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. Recuperado de https://doi.org/10.1016/j.iswcr.2017.07.002

Camargo, A. P., Marin, F. R., Sentelhas, P. C., & Picini, A. G. (1999). Adjust of the Thornthwaite's method to estimate the potential evapotranspiration for arid and superhumid climates, based on daily temperature amplitude [JOUR]. *Revista Brasileira de Agrometeorologia*, 7(2), 251-257.

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. Recuperado de https://doi.org/10.1145/1961189.1961199

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tre boosting system. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *19*(6).

Recuperado de https://doi.org/10.1145/2939672.2939785

Efron, B. (1992). *Bootstrap methods: Another look at the jackknife*. Recuperado de https://doi.org/10.1007/978-1-4612-4380-9_41

Fan, J., Wang, X., Wu, L., Zhou, H., Zhang, F., Yu, X.,… & Xiang, Y. (2018a). 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. Recuperado de https://doi.org/10.1016/j.enconman.2018.02.087

Fan, J., Yue, W., Wu, L., Zhang, F., Cai, H., Wang, X.,… & Xiang, Y. (2018b). 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, 225-241. Recuperado de https://doi.org/10.1016/j.agrformet.2018.08.019

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. Recuperado de https://doi.org/10.1016/j.jhydrol.2016.02.053

Ferreira, C. (2001). Gene Expression Programming: a New Adaptive Algorithm for Solving Problems. *Arxiv.org*. Recuperado de https://arxiv.org/abs/cs/0102027

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. Recuperado de https://doi.org/10.1016/j.agwat.2016.11.010

INEGI, Instituto Nacional de Estadística y Geografía. (2017). *Anuario estadístico y geográfico de Campeche 2017*. Recuperado de 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. Recuperado de https://doi.org/10.1016/j.compag.2018.04.003

Kelso-Bucio, H., Ba, K. M., Magaña, H. F., Sánchez, M. S., Reyes, L. D., & Pascual, R. F. (2013). Recalibración regional de los coeficientes de Hargreaves (HE y KRS) en México. *XXII Congreso Nacional de Hidráulica y 1er Congreso Internacional de Ingeniería Agrícola*, México.

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

Mattar, M. A., & Alazba, A. A. (2019). GEP and MLR approaches for the prediction of reference evapotranspiration. *Neural Computing and Applications*, 31(10), 5843-5855. Recuperado de https://doi.org/10.1007/s00521-018-3410-8

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. Recuperado de 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. Recuperado de https://doi.org/10.1016/j.compag.2017.05.002

Quej, V. H., Almorox, J., Arnaldo, J. A., & Moratiel, R. (2019). Evaluation of temperature-based methods for the estimation of reference evapotranspiration in the Yucatán Peninsula, Mexico. *Journal of Hydrologic Engineering*, 24(2). Recuperado de https://doi.org/10.1061/(ASCE) HE.1943-5584.0001747

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, 62-70. Recuperado de https://doi.org/10.1016/j.jastp.2017.02.002

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

Salazar, E. A. Q., Ureña, W. A., & Gallego, H. A. B. (2010). Interfaz gráfica para la interpolación de datos a través de splines. *Scientia et Technica*, 1(44), 195-200.

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. Recuperado de https://doi.org/10.1016/j.agwat.2017.04.009

Shiri, J., Sadraddini, A. S., 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. Recuperado de 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. Recuperado de https://doi.org/10.1016/j.agrformet.2014.09.025

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

Torrente-Cantó, L. (2018). *Reconstrucción basada en interpolación de Hermite aplicada a funciones discontinuas*. Recuperado de http://repositorio.upct.es/handle/10317/7584

Urraca, R., Antonanzas, J., Antonanzas-Torres, F., & Martinez-De-Pison, F. J. (2017). Estimation of daily global horizontal irradiation using extreme gradient boosting machines. *Advances in Intelligent Systems and Computing*, 527, 105-113. Recuperado de

https://doi.org/10.1007/978-3-319-47364-211

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory* (2nd ed.). New York, USA: Springer-Verlag.

Webb, C. P. (2010). *Boreau of meteorology reference evapotranspiration calculations*. (February), 20.

Wen, X., Si, J., He, Z., Wu, J., Shao, H., & Yu, H. (2015). Support-vector-machine-based models for modeling daily reference evapotranspiration with limited climatic data in extreme arid regions. *Water Resources Management*, 29(9), 3195-3209. Recuperado de https://doi.org/10.1007/s11269-015-0990-2

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