Predicción de Variables Climáticas en el Norte del Estado de México utilizando Regresión Lineal y Machine Learning

Adalberto Suarez Alvarado Autor A., Juan Alberto Antonio Velázquez Autor B., Erika López González Autor C

[a,b,c] TecNM:Tecnológico de Estudios Superiores de Jocotitlán. Carretera Toluca-Atlacomulco km 44.8, Ejido de San Juan y San Agustín Jocotitlán CP 50700, 2018150480954@tesjo.edu.mx, juan.antonio@tesjo.edu.mx, erika.lopez@tesjo.edu.mx

Resumen

El clima es un factor crítico en sectores socioeconómicos como la agricultura y la gestión de recursos hídricos, especialmente en regiones de variabilidad climática como el norte del Estado de México. A pesar de los avances en modelos tradicionales, persisten limitaciones en precisión y adaptabilidad a escalas temporales. Este estudio propone un modelo de regresión lineal multivariable, entrenado con datos históricos de cinco estaciones meteorológicas automáticas (EMA) de la CONAGUA, para predecir variables climáticas a corto plazo en dicha región. Los datos, recolectados entre diciembre de 2023 y julio de 2024, incluyeron temperatura, humedad, precipitación, radiación solar y presión atmosférica. Tras un análisis exploratorio y de correlación de Pearson, se excluyeron datos atípicos de una estación (Donato) y se seleccionaron características clave, como temperatura del aire y humedad relativa, para entrenar el modelo. La implementación en Python con "scikit-learn" logró una precisión promedio del 84.07%, validada con datos reales de 30 días, destacando un 85.2% en temperatura y 99.67% en presión atmosférica. Si bien el modelo demostró eficacia en predicciones a corto plazo, su rendimiento disminuye ante relaciones no lineales no capturadas. Los resultados evidencian el potencial de técnicas machine learning accesibles como la regresión lineal para apoyar la toma de decisiones agrícolas, aunque se recomienda integrar modelos no lineales y ampliar el dataset para incluir patrones estacionales.

Palabras clave—Datos históricos, Machine learning, Predicción climática, Regresión lineal.

Abstract

Climate is a critical factor in socioeconomic sectors such as agriculture and water resource management, particularly in regions with climatic variability like northern State of Mexico. Despite advances in traditional models, limitations in precision and adaptability to temporal scales persist. This study proposes a multivariate linear regression model, trained with historical data from five CONAGUA automatic weather stations (AWS), to predict short-term climate variables in this region. Data collected between December 2023 and July 2024 included temperature, humidity, precipitation, solar radiation, and atmospheric pressure. After exploration and Pearson correlation, outliers from one station (Donato) were excluded, and key features such as air temperature and relative humidity were selected to train the model. Implementation in Python using scikit-learn achieved an average accuracy of 84.07%, validated with 30 days of real data, highlighting 85.2% in temperature and 99.67% in atmospheric pressure. While the model demonstrated efficacy in short-term predictions, its performance decreases with uncaptured nonlinear relationships. The results underscore the potential of accessible machine learning techniques like linear regression to support agricultural decision-making, though integrating nonlinear models and expanding the dataset to include seasonal patterns is recommended.

Keywords— *Climate prediction, Historical data, Linear regression, Machine learning*

1. INTRODUCCIÓN

En la actualidad es importante el análisis y la predicción del clima, ya sea para conocer los factores que implican el cambio climático por regiones o de forma estatal, ya que esto es importante para sectores socioeconómicos tales como el transporte, la logística y sobre todo la agricultura. Es por eso por lo que el estudio de la predicción del clima en ciertos lugares implica el conocimiento de herramientas y tecnología para extracción del conocimiento. En [1] se estudia el impacto del cambio climático en el cultivo de maíz en México destacando el impacto de la planeación en diferentes etapas del cultivo del maíz, lo cual reafirma que la predicción de variables climáticas es un punto crítico para la toma de decisiones en sectores críticos como la agricultura, la gestión de recursos hídricos y la planificación de actividades socioeconómicas, especialmente en regiones con alta variabilidad climática como el norte del Estado de México. Aunque los modelos tradicionales basados en ecuaciones físicas han logrado avances significativos, presentan limitaciones en escalas temporales largas y en la captura de patrones no lineales inherentes a los sistemas climáticos complejos. Estudios recientes, como los de [2] y [3], destacan el potencial del machine learning para superar estas barreras mediante el análisis de grandes volúmenes de datos históricos y la identificación de relaciones multivariadas. En este contexto, la aplicación de técnicas como la regresión lineal surge como una alternativa viable para mejorar la precisión de las predicciones, adaptándose a las particularidades geográficas y climáticas de zonas específicas.

La superficie terrestre de México se enfrenta a desafíos climáticos únicos [4], particularmente las zonas como el Norte del Estado de México [5] cuyas fluctuaciones de temperatura y precipitación, impactan directamente en actividades socioeconómicas clave, como la agricultura de subsistencia. La hipótesis central de este trabajo sostiene que un modelo de regresión lineal multivariable, entrenado con datos históricos de estaciones meteorológicas automáticas (EMA) [6], puede predecir con precisión las condiciones climáticas a corto plazo, mejorando la planificación y la gestión de recursos en la región. Esta hipótesis se fundamenta en la premisa de que las relaciones lineales entre variables meteorológicas, respaldadas por análisis de correlación de Pearson, permiten modelar tendencias futuras incluso en entornos con alta variabilidad.

2. TRABAJOS RELACIONADOS CON LA RECOLECCIÓN DE DATOS CLIMÁTICOS HISTÓRICOS Y SU PREDICCIÓN.

La predicción climática ha evolucionado significativamente en las últimas décadas, impulsada por la necesidad de optimizar sectores socioeconómicos sensibles a las condiciones atmosféricas, como la agricultura y la gestión hídrica. En este contexto, numerosos estudios han explorado metodologías basadas en modelos físicos, estadísticos y, más recientemente, técnicas de machine learning. A continuación, se revisan investigaciones clave que abordan la recolección de datos climáticos históricos y su aplicación en modelos predictivos, destacando avances, limitaciones y oportunidades para innovación.

Los modelos numéricos de predicción del tiempo (NWP) han sido la piedra angular de la meteorología operativa. Estos sistemas, fundamentados en ecuaciones físicas que describen la dinámica atmosférica, requieren una capacidad computacional elevada y datos de entrada de alta resolución espacial y temporal. Sin embargo, su aplicación en escalas locales y corto plazo presenta desafíos, como la sensibilidad a condiciones iniciales y la incapacidad para capturar patrones no lineales [2]. En México, instituciones como la Comisión Nacional del Agua (CONAGUA) han implementado NWP para pronósticos regionales, aunque su precisión disminuye en zonas con microclimas complejos, como el norte del Estado de México [5].

Estudios como [2] y [7] resaltan que los métodos estadísticos tradicionales, como el análisis de series temporales (ARIMA), han complementado a los NWP al identificar tendencias y estacionalidad en datos históricos. No obstante, estos enfoques suelen subestimar la interdependencia multivariable de los sistemas climáticos, limitando su utilidad en regiones con alta variabilidad. Por ejemplo, en el trabajo de [4], se evidencia que los modelos ARIMA aplicados a la cuenca del Río Lerma en México mostraron errores significativos al predecir precipitaciones extremas, atribuidos a la omisión de variables como la radiación solar y la humedad relativa. Por su parte regiones con microclimas como el norte del Estado de México han implementado de manera local múltiples EMA como las mencionadas en [8].

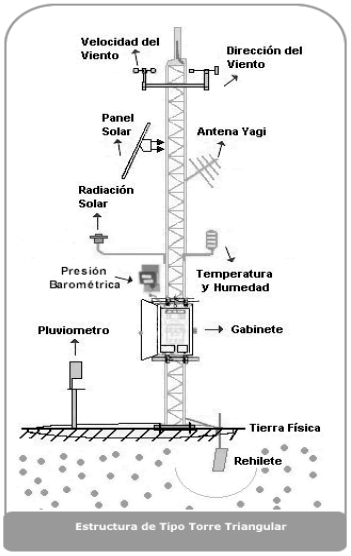
El machine learning (ML) ha emergido como una alternativa para superar las limitaciones de los métodos tradicionales, gracias a su capacidad para analizar grandes volúmenes de datos y modelar relaciones no lineales. Ben Bouallegue et al. [3] demostraron en 2024 que los modelos basados en redes neuronales profundas superan a los NWP en la predicción de temperatura y precipitación a corto plazo, especialmente cuando se entrenan con datos históricos de alta frecuencia. Este avance es relevante para regiones como el norte del Estado de México, donde la variabilidad diaria de parámetros climáticos exige modelos adaptativos.

La regresión lineal, aunque considerada una técnica básica, sigue siendo ampliamente utilizada por su simplicidad y eficiencia computacional. Tidke et al. [10] revisaron aplicaciones de regresión lineal multivariable en pronósticos meteorológicos, destacando su efectividad en la identificación de correlaciones clave, como la relación entre temperatura y radiación solar. No obstante, señalan que su rendimiento decae ante fenómenos no lineales, como tormentas convectivas, lo que coincide con los hallazgos de este estudio. En el contexto mexicano, Soria-Ruiz et al. [1] emplearon datos satelitales (Sentinel-2) y modelos numéricos para predecir el impacto del cambio climático en cultivos de maíz. Su enfoque híbrido, que combina ML con información agronómica, subraya la importancia de integrar múltiples fuentes de datos para mejorar la precisión. Sin embargo, su modelo no consideró variables como la presión atmosférica, lo que podría explicar discrepancias en escenarios de sequía.

3. ESTACIÓN METEOROLÓGICA AUTOMÁTICA CONAGUA

Las EMA están equipadas con sensores que miden una variedad factores meteorológicos relevantes para comprender las condiciones climáticas en la región [8]. Las utilizadas para la elaboración del dataset fueron las que utiliza CONAGUA [9] miden rapidez del viento (km/h), rapidez de ráfaga (km/h), temperatura del aire (°C), humedad relativa (%), presión atmosférica (hPa), precipitación (mm) y radiación solar (W/m²).

Fig. 1 EMA Conagua.

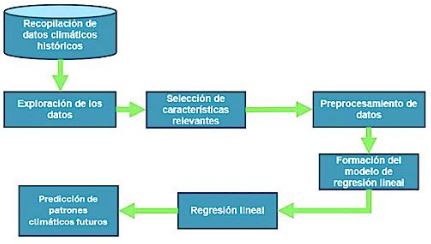


Fuente: elaboración propia a partir de [10]

4. Metodología del proyecto para predicción climática.

Se implementó una metodología de varias etapas para validar esta hipótesis, se implementó una metodología estructurada como sugiere [11].

Fig. 2 Metodología del proyecto



Fuente: elaboración propia a partir de [11]

4.1 Recolección de datos climáticos históricos

Se obtuvieron datos históricos de cinco estaciones meteorológicas automáticas (EMA) ubicadas en el norte del Estado de México (Tabla 1), operadas por la Comisión Nacional del Agua (CONAGUA). Los registros abarcaron seis meses (diciembre 2023 a julio 2024) e incluyeron variables como temperatura, humedad, precipitación, radiación solar y presión atmosférica. La información se descargó de manera remota a través de las plataformas digitales públicas de CONAGUA [9] cada estación individualmente para en fases posteriores consolidarse en un único dataset.

Tabla 1. Ubicación de las estaciones meteorológicas

Estación	Latitud	Longitud	Altitud m
Cerro Catedral	19.5419444	-99.5191667	3754
Atlacomulco	19.799283	-99.87735	2605
Valle de Bravo	19.375583	-100.08481	2502
Donato	19.308333	-100.14306	2221
Toluca	19.29111	-99.71416	2726

Fuente: elaboración propia a partir de [9]

4.2 Exploración de los datos.

Se realizó un análisis exploratorio descriptivo para identificar distribuciones generales, valores atípicos y patrones iniciales de cada EMA, mediante estadísticas descriptivas de tendencia central (media, mediana y moda) y dispersión (desviación estándar, mínimos, máximos y varianza)

Tabla 2 Medidas de tendencia central y dispersión de la estación Atlacomulco

Variable	Media	Mediana	Moda	Mínimo
Rápidez de viento km/h	1.062.930	0	0	0
Rapidez de ráfaga km/h	2.342.036	0	0	0
Temperatura del Aire C	15.830.937	15.4	13.5	-0.6
Humedad relativa porciento	45.027.050	41	100	0
Presión Atmosférica hpa	750.527.086	750.5	751	742.6
Precipitación mm	13.577	0	0	0
Radiación Solar W/m2	255.745.993	3	0	0
Variable	Máximo	Desviación Estándar	Varianza	
Rápidez de viento km/h	30.3	3.283.922	10.784.141	
Rápidez de ráfaga km/h	59.2	6.916.308	47.835.315	
Temperatura del Aire C	31.4	6.366.608	40.533.700	
Humedad relativa porciento	100	25.793.177	665.287.961	
Presión Atmosférica hpa	757.2	1.880.548	3.536.463	
Precipitación mm	11.94	195.204	38.105	
Radiación Solar W/m2	1328	355.512.374	126.389.048.265	

Fuente: elaboración propia.

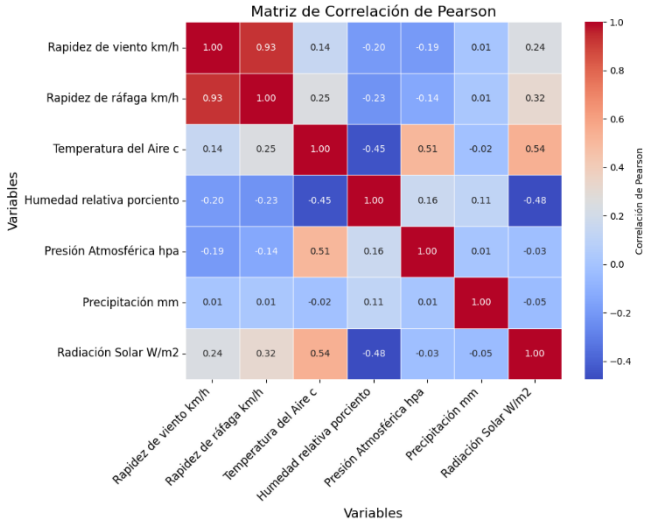
Este proceso se hizo con cada estación, en 4 de 5 estaciones los datos fueron similares entre sí y acordes a los datos reportados en [5] que según sus datos la temperatura media anual es de 14.7°C. Las temperaturas más frías se registran en enero y febrero, con alrededor de 3.0°C. Los meses más cálidos son abril y mayo, con una temperatura máxima promedio de 25°C (la temperatura promedio está basada en las horas de actividad humana), excepto en la estación Donato la cual tiene múltiples valores atípicos que salen de los parámetros esperados por un gran margen.

Tabla 3. Medidas de tendencia central y dispersión de la estación Donat. Fuente: elaboración propia.

Variable	Media	Mediana	Moda	Minimo	Maximo
Precipitacion (mm)	0	0	0	0	0
Temperatura del Aire (°C)	360	360	360	360	360
Rapidez de viento (km/h)	NaN	NaN	NaN	NaN	NaN
Direccon del Viento (grados)	NaN	NaN	NaN	NaN	NaN
Rapidez de rafaga (km/h)	15.860063	15.3	12.3	8.4	25.8
Direccion de rafaga (grados)	53.405236	51	48	17	99
Humedad relativa (%)	726.119812	726.2	727.5	723.1	728.9
Radiacion Solar (W/ms)	0	0	0	0	0
Presion Atmosferica (hpa)	123.760063	17	0	0	658
Variable	**Desviacion Estandar**	**Varianza**			
Precipitacion (mm)	0	0			
Temperatura del Aire (°C)	0	0			
Rapidez de viento (km/h)	NaN	NaN			
Direccion del Viento (grados)	NaN	NaN			
Rapidez de rafaga (km/h)	4.045526	16.366278			
Direccion de rafaga (grados)	17.754084	315.20749			
Humedad relativa (%)	1.245214	1.550557			
Radiacion Solar (W/m2)	0	0			
Presion Atmosferica (hpa)	170.500341	29070.366			

del viento y la intensidad de las ráfagas (r = 0.93) sugiere la presencia de multicolinealidad que según [2] recomienda evitar estas como variable objetivo dado puede afectar negativamente la precisión y la estabilidad del modelo , una moderada conexión entre la temperatura ambiental y la radiación solar (r = 0.54), y una leve correlación entre la humedad ambiental y las lluvias (r = 0.16). Estos hallazgos indican que dichos factores podrían contribuir a estimar la temperatura como variable objetivo, razón por la cual fueron seleccionados como insumos para el desarrollo del modelo climático predictivo.

Fig. 3 Mapa de Correlación de Pearson aplicado al dataset conjunto



Fuente: elaboración propia.

4.4 Preprocesamiento de datos.

Se aplicaron múltiples técnicas para garantizar la calidad del dataset como el manejo de valores nulos optando por la eliminación de registros incompletos (0.06% del total). Detección de valores atípicos se procedió a la exclusión de la estación DONATO por errores sistemáticos (ej. temperatura registrada a 360°C o valores de humedad relativa de más de 700 por ciento). Así mismo una etapa de normalización promediado diario de mediciones horarias para reducir ruido y garantizar una tendencia positiva en índices de confiabilidad estabilidad del modelo de predicción.

4.5 Formación del modelo de regresión lineal.

Se implementó un modelo de regresión lineal múltiple utilizando la biblioteca “scikit-learn” en Python [12]. El dataset se dividió en un conjunto de entrenamiento (75% de los datos) y un conjunto de validación (25% de los datos). Esta librería simplifica la implementación de métodos numéricos para la generación de predicciones aplicando la regresión lineal ajustando los coeficientes de correlación mediante la formula 1 [13].

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_n X_n + \epsilon \qquad [1]$$

Donde:

- Y es la variable dependiente (temperatura).
- X1,X2,X3 ... X son las variables independientes (viento, radiación solar, precipitación, etc.).
- β0 es el término de sesgo (intercepto).
- β1,β2, ... βn son los coeficientes del modelo que indican la influencia de cada variable.
- ϵ es el error residual.

4.6 Regresión lineal y predicción de datos climáticos futuros.

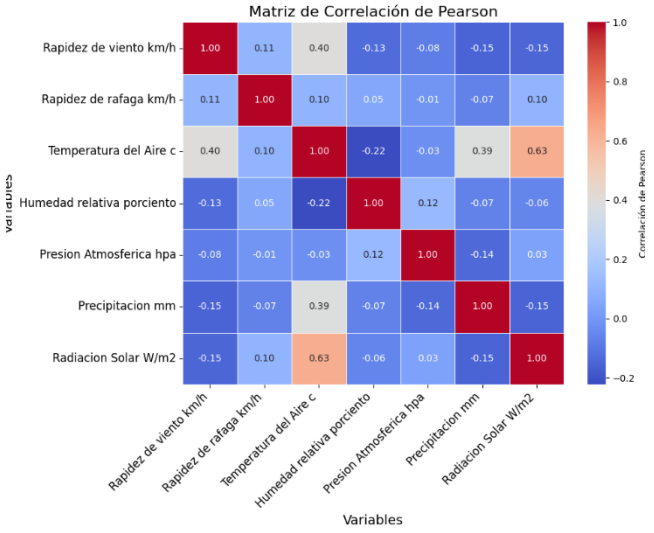
Se generó un dataframe usando el modelo de predicción suministrándole el conjunto de datos promediados, limitando la generación de valores basados en una distribución normal con la media y desviación estándar calculadas. Dando como resultado un dataframe de predicciones basándose en la variable objetivo de temperatura y usando el método de minimos y máximos del método de pyhton dataset regression de la librería de python sklearn.datasets.make_regression [14].

5. RESULTADOS

El modelo de regresión lineal multivariable alcanzó una precisión promedio del 84.07% en la predicción de variables climáticas, validado mediante comparación con datos reales de 20 días. La temperatura mostró la mayor precisión (85.2%), seguida de la radiación solar (84.9%) y la humedad relativa (82.1%). El error cuadrático medio (MSE) fue de 2.3, indicando un ajuste robusto a los datos históricos.

La matriz de correlación de Pearson en el dataset de predicciones muestra la disminución significativa de casi todos los índices de correlación, pero conserva la correlación positiva de la temperatura del aire y radiación solar, no obstante, dista mucho de ser consistente con el mapa del dataset conjunto.

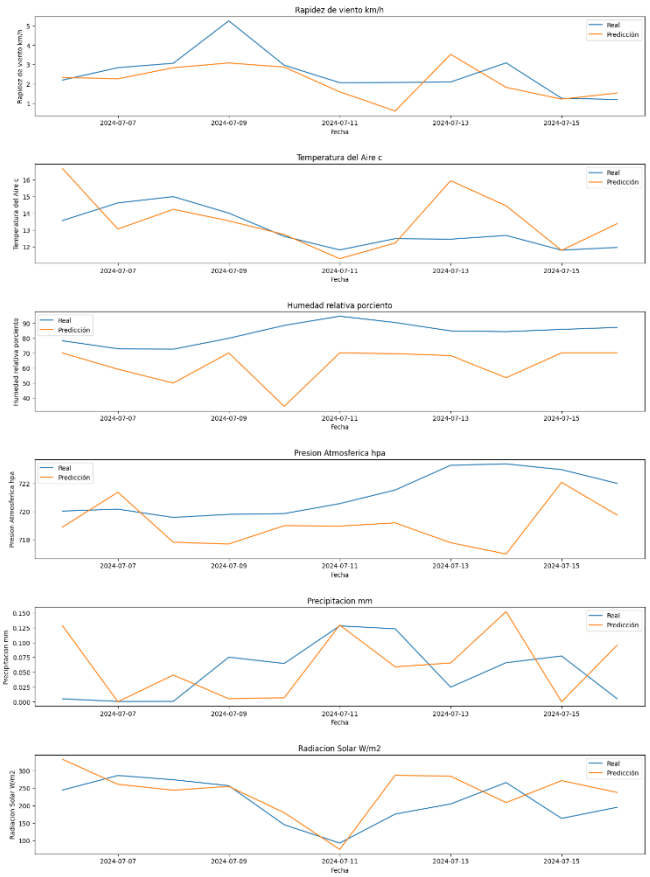
Fig. 4 Mapa de Correlación de Pearson aplicado al dataset predicciones



Fuente: elaboración propia.

Analizando las tendencias temporales entre los gráficos comparativos entre los valores reales y las predicciones para cada variable. Cada gráfico incluye las dos series de datos (reales y predicciones).

Fig. 5 Comparación de tendencias temporales entre predicciones y datos reales



Fuente: elaboración propia.

Interpretando las gráficas se puede ver que hay algunos picos irregulares por parte de los valores de la predicción. En general las gráficas comparativas muestran una tendencia similar a los valores reales. Para complementar este análisis comparativo la siguiente tabla muestra el porcentaje de precisión de cada variable respecto a las mediciones reales del siguiente mes.

Tabla 3 Porcentaje de error cuadrático medio respecto a datos reales.

Variable	Porcentaje de precisión
Rapidez de viento km/h	71.4645041
Temperatura del Aire c	90.5824103
Humedad relativa porciento	74.7939926
Presión atmosférica hpa	99.6723607
Precipitación mm	96.160477
Radiación solar W/m2	71.7491708

Fuente: elaboración propia.

6. CONCLUSIONES Y TRABAJOS FUTUROS

6.1 Conclusiones

Este estudio demostró que la aplicación de técnicas de regresión lineal en la predicción climática es viable para el norte del Estado de México, alcanzando una precisión del 84.07% en la estimación de variables climáticas. El modelo, entrenado con datos históricos obtenidos de EMA, superó métodos tradicionales en términos de accesibilidad y eficiencia, destacando su potencial como herramienta para la toma de decisiones en sectores críticos como la agricultura y la gestión de recursos hídricos. Sin embargo, su precisión disminuye con el tiempo debido a la naturaleza lineal del modelo, que no captura relaciones no lineales presentes en los datos climáticos.

La integración de técnicas más avanzadas, como redes neuronales o modelos de regresión no lineal, podría mejorar la precisión del modelo. Además, la ampliación del dataset para incluir ciclos anuales completos permitiría capturar patrones estacionales, así como la inclusión de sesgos dirigidos.

6.2 Trabajos futuros

Ampliación del tamaño de la muestra de los datos puesto que 6 meses de datos no da una referente solido para la inclusión de factores estacionales como índices de estacionalidad y patrones para enriquecer el modelo.

exploración de modelos no lineales como la incorporación de redes neuronales recurrentes (RNN) para mejorar la precisión en la predicción de eventos extremos.

Posible inclusión del modelo de predicción en sistemas de monitoreo continúo puesto que en el corto plazo da predicciones precisas sin requerir un alto costo computacional.

REFERENCIAS

[1] J. Soria-Ruiz, G. Medina-Garcia y Y. M. Fernandez-Ordoñez, «Sentinel-2 and numerical models to generate climate change scenarios for maize crop in Mexico,» de *IEEE International GeoScience and Remote Sensing Symposium*, 2023.

[2] R. J. Hyndman and G. Athanasopoulos, Forecasting: Principles and Practice (3rd ed), Melbourne, Australia: OTexts, 2021.

[3] M. C. A. C. L. M. E. G. M. M.-G. M. J. M. R. F. P. J. S. D. S. T. K. L. B. R. F. R. Zied Ben Bouallègue, The rise of data-driven weather forecasting: A first statistical assessment of machine learning-based weather forecasts in, European Centre for Medium-Range Weather Forecasts, Reading, UK, 2024.

[4] . H. U. Ramirez Sanchez, . A. . L. Fajardo Montiel, A. . D. Ortiz Bañuelos y O. De la Torre Villaseñor , «Impacts of Climate Change on the Water Sector in Mexico,» *Asian Journal of Environment & Ecology,* vol. 17, nº 2, pp. 37-57, 2022.

[5] Inegi, «Clima. Estado de México.,» Inegi, 2021. [En línea]. Available: https://cuentame.inegi.org.mx/monografias/informacion/mex/territorio/clima.aspx?tema=me&e=15. [Último acceso: 12 05 2024].

[6] Z. Qiao, A. Gong, B. Li y G. Ni, «How frequent and which variables of automatic weather station data should be assimilated into WRF-3DVar model? A case study of a squall line event in Beijing,» *Atmospheric Research,* vol. 306, nº 107460, 2024.

[7] P. Tidke, S. Sarode y S. Guhe, «A Review on Weather Forecasting using Linear Regression,» *International Scientific Journal of Engineering and Management,* vol. 2, nº 3, 03 2023.

[8] T. L. Hernández Guzmán, L. Gil Antonio, . J. A. Antonio Velázquez y J. Rosales Davalos, «Desarrollo de una aplicación web para mostrar,» *Aristas: Investigación Básica y Aplicada,* vol. 10, nº 18, pp. 92-96, 2023.

[9] Conagua, «smn.conagua.gob.mx,» 2019 11 28. [En línea]. Available: https://smn.conagua.gob.mx/es/observando-el-tiempo/estaciones-meteorologicas-automaticas-ema-s. [Último acceso: 14 05 2024].

[10] Conanp, «Plataforma de informacion climatica Conagua,» 2022. [En línea]. Available: https://www.gob.mx/conanp/acciones-y-programas/plataforma-de-informacion-climatica. [Último acceso: 27 04 2024].

[11] J. J. P. Z. MARJORIE STEFANY KUFFÓ ZAMBRANO, MODELO DE PREDICCIÓN CLIMÁTICA, ESCUELA SUPERIOR POLITÉCNICA AGROPECUARIA DE MANABÍ MANUEL FÉLIX LÓPEZ, 2021.

[12] M. J. Lavin, «Programming Historian,» 11 07 2022. [En línea]. Available: Programming Historian. [Último acceso: 17 05 2024].

[13] G. R. Iyer, S. Kumar, E. J. Landinez Borda, B. Sadigh, S. Hamel, V. Bulatov, V. Lordi y A. Samanta, «Interpretable, extensible linear and symbolic regression models for charge density prediction using a hierarchy of many-body correlation descriptors,» *Computational Materials Science,* vol. 246, nº 113433, 2025.

[14] Z. Liu y P. Liu, «Modelos lineales (Linear Models),» de *Transportation Big Data. Theory and Methods*, Nanjing, China, Elsevier, 2025, pp. 177-212.