ANÁLISIS DE PATRONES ESPACIALES EMERGENTES DE LLUVIA EN LA CIUDAD DE MEDELLIN

Analysis of rainfall emerging spatial patterns in the city of Medellin

DIEGO RANGEL VELÁSQUEZ

Trabajo de grado

DIRECTOR TOMAS OLARTE HERNANDEZ

CODIRECTOR JULIÁN SEPÚLVEDA BERRIO

UNIVERSIDAD EAFIT

ESCUELA DE CIENCIAS APLICADAS E INGENIERÍA

MAESTRÍA EN CIENCIAS DE LOS DATOS Y LA ANALÍTICA

MEDELLÍN

2025

CONTENIDO

LISTA DE FIGURAS....4
LISTA DE TABLAS....4
LISTA DE ECUACIONES....5
RESUMEN....6
ABSTRACT....7
INTRODUCCIÓN....8
PLANTEAMIENTO DEL PROBLEMA....9
JUSTIFICACIÓN....10
OBJETIVOS....11
GENERAL....11
ESPECÍFICOS....11
MARCO CONCEPTUAL....12
CONTEXTO DEL CASO DE USO....12
TECNICAS DE PREDICCIÓN....12
FUENTES DE INFORMACIÓN....13
IN SITU....13
REMOTAS....15
ARQUITECTURA ELEGIDA: UNETS....17
DISEÑO METODOLÓGICO....19
ENTENDIMIENTO DEL NEGOCIO Y DE LOS DATOS....19
PREPARACION DE LOS DATOS....20
MODELAMIENTO DE DATOS....21

EVALUACION 22
DESPLIEGUE 23
RESULTADOS 24
CONCLUSIONES 29
BIBLIOGRAFIA 31

LISTA DE FIGURAS

Figura 1: Simulación en Python de un evento de precipitación en el Valle de Aburrá mediante teselación por polígonos de Voronoi a partir de datos de estaciones pluviométricas del SIATA. Elaboración propia. 13

Figura 2: Ejemplos de interpolaciones de valores de mediciones pluviométricas. [12] 14

Figura 3: Visualización de precipitaciones en el aplicativo web del SIATA.[16]..... 15

Figura 4: Visualización de evento de precipitación cercano al Valle de Aburrá según RRQPE del GOES-16. Elaboración propia. 16

Figura 5: Esquema de la UNET Propuesta inicialmente por Ronneberger et al. en 2015. [20]..... 17

Figura 6: Grafica de elaboración propia de los eventos de precipitación del 29 de diciembre a las 00:50 GMT 0 centrados en el Valle de Aburrá pero visualizando gran parte del territorio nacional colombiano donde se observan diferentes eventos de precipitación en la región. 24

Figura 7: Histograma de tiempos entre muestras demostrando una alta consistencia temporal. 25

Figura 8: Histograma de valores medidos de precipitación en el dataset. Se observa una distribución bimodal de los datos alrededor de los 5mm/h y 50mm/h..... 25

Figura 9: Ejemplo de episodio de precipitación del 3 de octubre del 2024. Donde el cuadrado interior de 32x32 pixeles la figura con numero de pixeles base 2 más pequeña capaz de encerrar todo el valle de Aburrá. 26

Figura 10: Distribución de datos de precipitación en pixeles verdaderos de salida en dataset para la UNET de 4 niveles en escala natural y en escala logarítmica. Se observa que la distribución de datos presenta una cola pesada que llega hasta donde inicialmente se truncaron los datos del dataset.. 27

Figura 11: Distribución de datos originalmente predichos por la UNET de 4 niveles. Se observa que se truncan aproximadamente en 40mm/h, mostrando que la distribución de datos de salida predichos no se corresponde con la distribución de datos de salida reales. 27

Figura 12: Ejemplo de Predicciones de evolución de precipitaciones con los diferentes modelos usados..... 28

LISTA DE TABLAS

Tabla 1: Ejemplo de fila en la tabla construida de imágenes satelitales crudas. 20

Tabla 2: Arquitectura implementada de Unet de 3 Capas. 21

Tabla 3: Arquitectura implementada de Unet de 4 Capas. 22

Tabla 4 Matriz de Confusión..... 23

Tabla 5: Nombre de cada modelo desarrollado. 26

Tabla 6: Métricas finales para los diferentes modelos. Se observa que el modelo UNET4 logra predecir en ventanas de un frame(15 minutos a futuro) la evolución de los eventos de precipitación, mucho mejor que el baseline LAST12. El modelo UNET3 no logra superar el baseline. 28

LISTA DE ECUACIONES

(1) 22
(2) 22
(3) 22
(4) 26

RESUMEN

El aprovechamiento de datos meteorológicos es importante para la pronta atención de emergencias causadas por fenómenos climatológicos. Los sistemas de monitoreo climático brindan información valiosa para la gestión de riesgos, pero su aprovechamiento está estrechamente relacionado a los modelos predictivos que se puedan construir basándose en esta información. En este caso, mediante análisis de mediciones pluviométricas se buscó identificar patrones espaciales emergentes en picos de lluvia que pueden llevar a emergencias que requieran intervención de entidades de atención a desastres. Aunque existen estudios sobre distribución de precipitaciones, potencial de desarrollar modelos que se ajusten mejor a las condiciones ambientales de ciudades específicas. Esta investigación desarrolló un modelo que se adapta a las condiciones específicas del valle de Aburrá, anticipando la llegada de torrenciales a zonas de riesgo específicas. Se encontró que es posible anticipar la evolución de precipitaciones en escenarios específicos de precipitaciones convencionalmente elevadas.

Palabras clave: Pronóstico del tiempo, predicción, eventos espaciales, patrones espaciales, nowcasting, forecasting, aprendizaje automático, aprendizaje profundo

ABSTRACT

The utilization of meteorological data is crucial for quick responses to emergencies caused by climatological phenomena. Climate monitoring systems provide valuable information for risk management; however, their effectiveness is closely linked to the predictive models that can be built using this data. In this case, through analysis of pluviometry measurements, there have been found emerging spatial patterns in rainfall peaks that may lead to emergencies requiring intervention from disaster response entities. Although studies on precipitation distribution exist, there is potential to develop models better adapted to the environmental conditions of specific cities. This research aimed to develop a model tailored to the specific conditions of the Aburrá Valley, anticipating the arrival of torrential rains in specific risk zones. The study found that it is possible to forecast precipitation trends in specific scenarios characterized by conventionally high rainfall levels.

Keywords: Weather forecasting, forecasting, nowcasting, spatial events, spatial patterns, machine learning, deep learning.

INTRODUCCIÓN

Las lluvias torrenciales pueden tener un impacto negativo en la vida diaria y en los sistemas socioeconómicos de las poblaciones, causando inundaciones, afectaciones al flujo vehicular y deslizamientos de tierra. En el contexto colombiano, la ausencia de estaciones climáticas bien definidas y la influencia de factores geográficos, como la cercanía a los océanos Atlántico y Pacífico, hacen que el país sea propenso a precipitaciones frecuentes con patrones espaciales variados. La ciudad de Medellín, ubicada en el Valle de Aburrá entre las cordilleras Occidental y Central, enfrenta condiciones geográficas únicas que favorecen la formación de lluvias orográficas, es decir, precipitaciones causadas por el ascenso del aire húmedo al chocar con las montañas. Por lo tanto, la predicción oportuna de estos eventos climáticos es crucial para la gestión de riesgos, la planificación urbana y para salvaguardar vidas y bienes, garantizando el bienestar de los habitantes[1], [2].En este contexto, la predicción de picos de lluvia, basada en datos de radar, pluviómetros y otras observaciones es una tarea crucial para la gestión de riesgos y la planificación urbana.

Tradicionalmente, la predicción meteorológica se ha basado en modelos de Predicción Numérica del Clima (NWP), los cuales utilizan la resolución de ecuaciones matemáticas que describen el comportamiento de los fenómenos atmosféricos. Sin embargo, estos modelos presentan limitaciones significativas. Por un lado, su alta complejidad computacional restringe su utilidad para predicciones a muy corto plazo. Además, suelen tener un margen de error considerable en las predicciones de eventos climáticos inminentes. Debido a estas limitaciones, ha surgido el paradigma del Nowcasting, es decir, modelos más eficientes y especializados en pronósticos de corto plazo, basados en algoritmos de aprendizaje de maquina en lugar de modelos físicos asociados a la naturaleza subyacente, que permitan una planificación efectiva en situaciones de desastres.[3], [4]

En esta investigación, se evaluaron distintos algoritmos de aprendizaje de maquina como herramientas para predecir la evolución de los patrones de precipitación en el Valle de Aburrá. Para esto se exploraron distintas fuentes de información meteorológica, se planteó una estructura de almacenamiento datos que permite almacenar información de forma liviana y recuperarla de forma ágil, se entrenaron distintos algoritmos y se analizaron sus métricas de desempeño. Logrando resultados que solo superan a un modelo baseline de persistencia en situaciones muy específicas. A pesar de esto se logró construir herramientas que permitirán a futuras investigaciones de patrones espaciales en el Valle de Aburrá, tener un punto de partida más sólido y desarrollar modelos más robustos.

PLANTEAMIENTO DEL PROBLEMA

A pesar de que existen diversas fuentes de información meteorológica en los que se puede consultar datos relativos a Medellín y el Valle de Aburrá, como los servicios satelitales de agencias ambientales, el radar Doppler y la red de sensores in situ del SIATA, la predicción de la evolución de eventos de precipitación a muy corto plazo sigue siendo un desafío. El estado del arte de la analítica para estos eventos se enfoca en mejorar los niveles de granularidad en las predicciones y en reducir los costos computacionales y de tiempo, implementando modelos de aprendizaje de máquina que, a pesar de lograr predicciones en ventanas de tiempo más cortas, son más veloces y confiables que los modelos de Predicción Numérica del Clima (NWP). Esta necesidad de predicciones rápidas y granulares se sustenta en su capacidad para mejorar la respuesta de las autoridades y de la población. Con esto en mente, surge la siguiente pregunta de investigación:

¿Es posible desarrollar un modelo de aprendizaje automático que, utilizando datos meteorológicos disponibles, sea capaz de predecir el comportamiento espacial de los eventos de precipitación en el Valle de Aburrá?

JUSTIFICACIÓN

El Valle de Aburrá presenta características geográficas y meteorológicas únicas, lo que implica que la predicción climática en la región dependa de modelos adaptados específicamente a su contexto. Esta necesidad se justifica en las condiciones demográficas de los municipios que lo conforman: una alta densidad de población, una gran variedad de topografías donde se ubican las viviendas, y un constante y elevado flujo vehicular. Estas condiciones hacen que los eventos climáticos extremos necesiten la mayor atención posible para prevenir colapsos en la movilidad, inundaciones en infraestructuras críticas, y, lo más importante, la pérdida de vidas humanas.

Por estas razones, las herramientas tecnológicas para la prevención y atención de desastres resultan de vital importancia en este contexto. En la actualidad, ya existen diversas fuentes de información climática que monitorean constantemente el territorio, tanto mediante dispositivos in situ como a través de sensores remotos que registran, entre otras variables, mediciones de precipitación.

Gracias a la alta frecuencia con la que estos sensores generan datos, es posible modelar el comportamiento de los patrones de lluvia para predecir su evolución y planear de manera eficiente el despliegue de los distintos servicios de atención ante emergencias desencadenadas por eventos de lluvia.

OBJETIVOS

GENERAL

Evaluar la capacidad de modelos basados en Machine Learning para hacer predicciones eventos de precipitación en el Valle de Aburrá a partir de la información meteorológica disponible de enero del 2024 a enero del 2025 del SIATA y de la Oficina Nacional de Administración Oceánica y Atmosférica de los Estados Unidos.

ESPECÍFICOS

- Identificar fuentes de datos meteorológicos para el Valle de Aburrá.
- Seleccionar una de las fuentes para desarrollar la investigación.
- Diseñar una estructura de datos en la que se pueda almacenar y consumir la información de la fuente escogida.
- Entrenar y compara las métricas de desempeño de diferentes modelos en orden incremental de complejidad, tomando como línea base de desempeño un modelo de persistencia.
- Identificar cuántos timestamps a futuro pueden predecirse con cada modelo y a cuánto tiempo de predicción equivale este número de timestamps.

MARCO CONCEPTUAL

CONTEXTO DEL CASO DE USO

La complejidad de la geografía de Colombia, influenciada por su ubicación ecuatorial y los efectos de fenómenos oceánicos como El Niño y La Niña, hace que el clima del país sea altamente variable durante todo el año. El Valle de Aburrá, donde se encuentra la ciudad de Medellín, presenta condiciones topográficas que intensifican las precipitaciones. La cercanía a las cordilleras Occidental y Central, que presentan alturas desde los 500 msnm hasta los 3000 msnm, provoca el ascenso de aire húmedo, las lluvias provocadas por estos movimiento de aire húmedo, se conocen como lluvias orográficas. Este fenómeno climático, combinado con la alta densidad de población, aumenta la vulnerabilidad de la región a desastres naturales como inundaciones y movimientos de masas[2].

En este contexto, en Medellín se crea el SIATA (Sistema de Alertas Tempranas). Un proyecto estratégico para la gestión ambiental y de riesgos del Área Metropolitana del Valle de Aburrá que, entre otras actividades, realiza monitoreo en tiempo real de fenómenos ambientales y busca pronosticar la ocurrencia de fenómenos naturales o antrópicos que pueden tener afectaciones a las poblaciones del Valle de Aburrá.[5]

Para mitigar estos riesgos, el SIATA ha implementado una red de monitoreo meteorológico que, a través de sensores in situ, radares y datos satelitales de terceros, recopila información crítica sobre las condiciones atmosféricas. Sin embargo, este sistema enfrenta el desafío de predecir fenómenos meteorológicos a muy corto plazo (0 a 6 horas) para poder planear de forma efectiva la atención a desastres. Superar este reto requiere modelos capaces de procesar y analizar grandes volúmenes de datos en periodos muy breves para generar alertas oportunas.

TECNICAS DE PREDICCIÓN

Tradicionalmente, las predicciones meteorológicas se han realizado mediante métodos numéricos, en los que se resuelven ecuaciones que describen el comportamiento físico de múltiples variables medioambientales en la atmosfera. [3] Sin embargo, estos presentan limitaciones debidas a la alta complejidad de las ecuaciones que describen estos fenómenos, junto con la gran cantidad de recursos computacionales que requieren y las limitadas ventanas de tiempo en las que pueden desarrollar predicciones. [4]

Debido al alcance restringido de estos modelos para generar predicciones que faciliten una planificación efectiva en la atención a desastres, se han desarrollado nuevos modelos basados en Machine Learning que buscan mejorar las predicciones a corto plazo y presentan diversos niveles de complejidad, que van desde modelos autorregresivos espaciales hasta arquitecturas de redes

neuronales. A los modelos desarrollados dentro de estos nuevos paradigmas se les conoce como Nowcasting.[6], [7], [8], [9]

FUENTES DE INFORMACIÓN

IN SITU

Para el caso de la fuente de datos in situ, El SIATA ha venido desplegando en el Valle de Aburrá desde 2010. Esta red monitorea los niveles de lluvia en sitio, minuto a minuto, utilizando dos sensores por estación. Además, se emplean códigos de verificación para indicar la confiabilidad de cada medición por minuto en cada estación. Las mediciones realizadas por los sensores se expresan en milímetros (mm), lo que equivale a litros por metro cuadrado.[10].

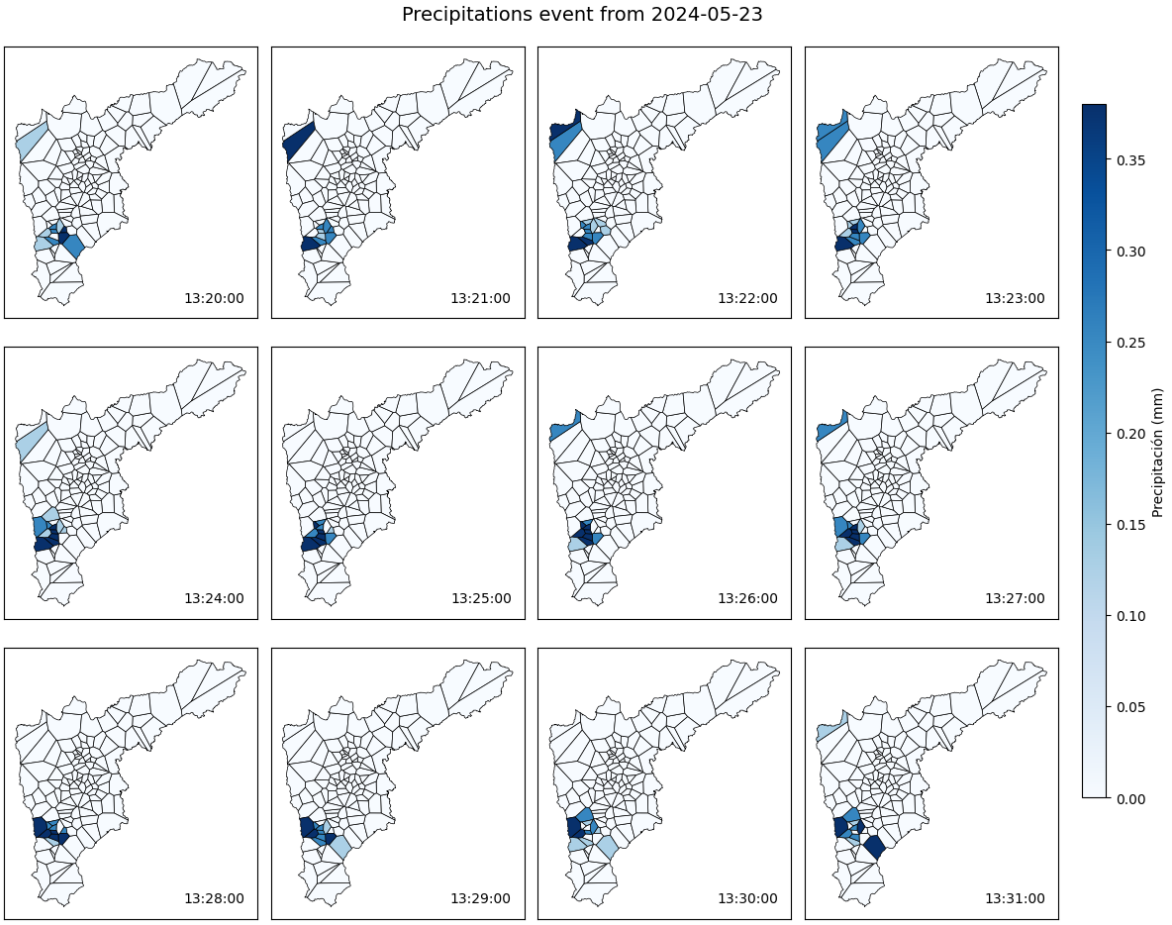

Figura 1: Simulación en Python de un evento de precipitación en el Valle de Aburrá mediante teselación por polígonos de Voronoi a partir de datos de estaciones pluviométricas del SIATA. Elaboración propia.

Este tipo de problemas se aborda generalmente desde el marco conceptual de los patrones espaciales. Este enfoque implica el análisis de la configuración de elementos en una distribución espacial. Los datos para este análisis suelen obtenerse mediante técnicas de censado remoto, que permiten recopilar información sobre un área sin contacto físico directo. Estas metodologías facilitan la observación y estudio de fenómenos ambientales a diversas escalas, proporcionando una base sólida para la predicción y modelado de eventos como las precipitaciones.[11]

Con este enfoque, a partir de las mediciones, sus ubicaciones y fechas de censado, se aplican diversas técnicas estadísticas para estimar los valores en puntos no muestreados. Un primer acercamiento a la estimación de valores en puntos no censados, es asumir en el modelo, que las áreas no censadas, presentan los mismos valores que el sensor más cercano a ellas, como se observa en la **Figura 1**, donde esta aproximación tesela el área de estudio con polígonos de Voronoi. Entre los métodos más utilizados para esta interpolación espacial, que buscan representaciones más precisas de las regiones no censadas, se encuentran el Kriging y el K-NN (K-Nearest Neighbors). Estas técnicas permiten generar una representación continua de las variables meteorológicas a partir de datos discretos, facilitando así el análisis y la predicción de patrones espaciales de fenómenos como la precipitación, como se observa en la **Figura 2**. Una vez inferido el comportamiento climatológico de toda la región de estudio para cada ventana temporal del intervalo sobre el cual se modelarán las predicciones, se procede a aplicar los modelos predictivos.[12]

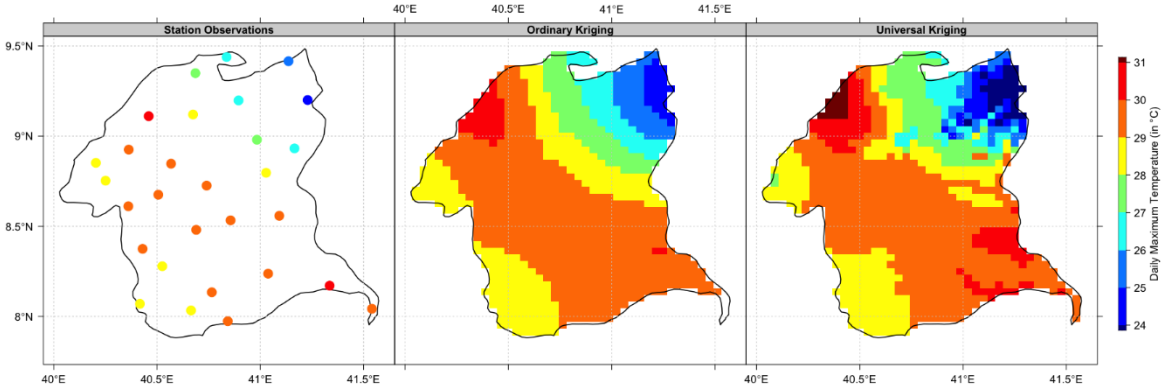

Figura 2: Ejemplos de interpolaciones de valores de mediciones pluviométricas. [12]

Cabe destacar que para esta clase de datos, se pueden realizar técnicas de Nowcasting sin necesidad de interpolar la información de toda la superficie a estudiar con la ayuda de redes neuronales alimentadas por grafos modelados a partir de las vecindades entre sensores in situ [13].

REMOTAS

RADAR DOPPLER

El SIATA (Sistema de Alerta Temprana de Medellín y el Valle de Aburrá) desplegó en el año 2011 un radar meteorológico Doppler, en el corregimiento de Santa Elena, Medellín. Este radar permite monitorear en tiempo real la evolución, movimiento y comportamiento de los sistemas de nubes asociados a precipitaciones en un radio de hasta 240 kilómetros alrededor de su ubicación, cubriendo gran parte del departamento de Antioquia y zonas aledañas, como se observa en la **Figura 3**. Esta fuente de información actualmente es muy valiosa para los municipios del Valle de Aburra debido a estar disponible apoyada con un aplicativo web que permite a cualquier usuario con acceso a internet conocer casi en tiempo real la evolución de las precipitaciones dentro del rango de cobertura del radar. [14], [15], [16]

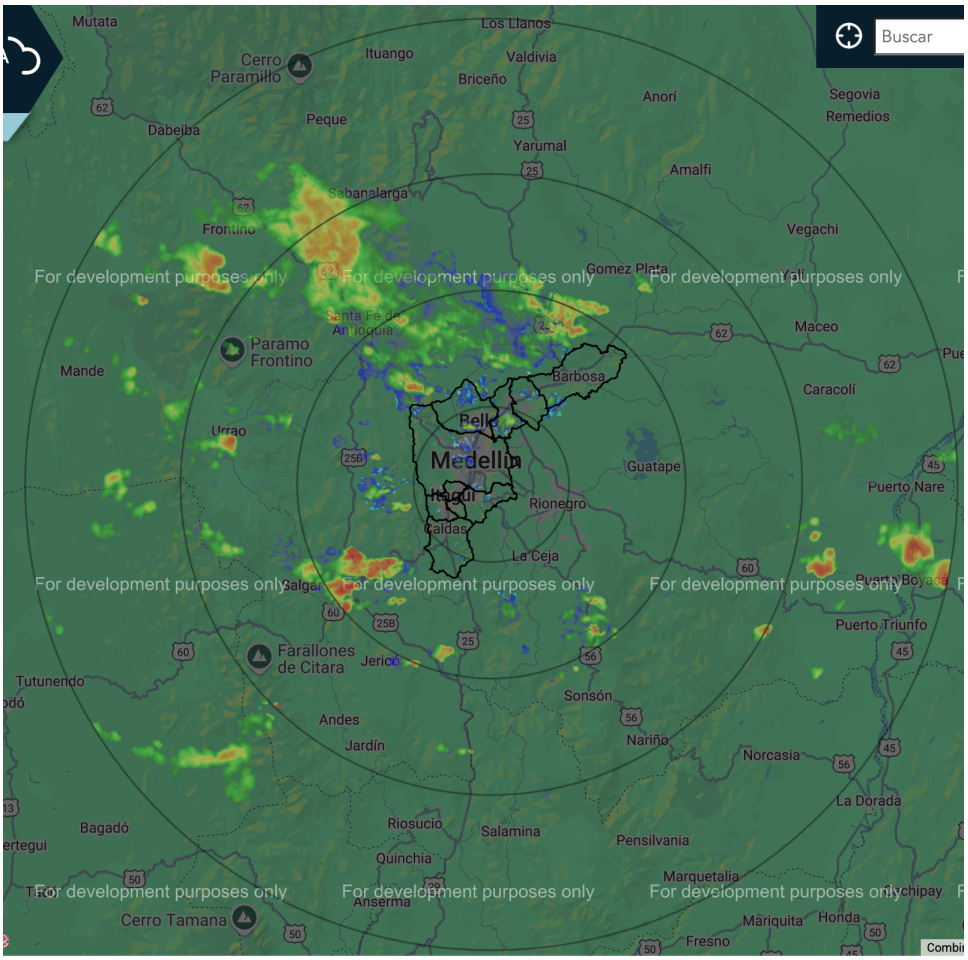

Figura 3: Visualización de precipitaciones en el aplicativo web del SIATA.[16]

INFORMACIÓN SATELITAL

Para el caso de la fuente satelital, la Oficina Nacional de Administración Oceánica y Atmosférica de los Estados Unidos (NOAA) opera el sistema de Satélites Geoestacionarios de Operación Ambiental (GOES). Dentro de este sistema de satélites, el GOES-16 mantiene una posición geoestacionaria fija sobre el ecuador, desde donde observa continuamente el hemisferio occidental, incluyendo el valle de Aburrá[17]. Este satélite mide la radiación reflejada y emitida por la superficie terrestre y la atmósfera en distintas bandas del espectro electromagnético, utilizando el sensor Advanced Baseline Imager (ABI). A partir de estas mediciones directas, se generan mediciones derivadas mediante algoritmos específicos que estiman variables atmosféricas, como la Estimación Cuantitativa de la Tasa de Precipitación (RRQPE, por sus siglas en inglés) Esta medición utiliza principalmente canales infrarrojos para asignar a cada píxel del área observada una estimación de la tasa de precipitación instantánea en unidades de mm/h [18], [19]. En este caso ya hay un grado de inferencia de los comportamientos meteorológicos para la región, donde se asume que toda la región esta teselada por los pixeles de la imagen satelital, como se muestra en la **Figura 4.** Lo que hace que esta fuente permita reducir los esfuerzos en inferencias y regresiones permitiendo enfocar más esfuerzos en la solución del problema de predicción.

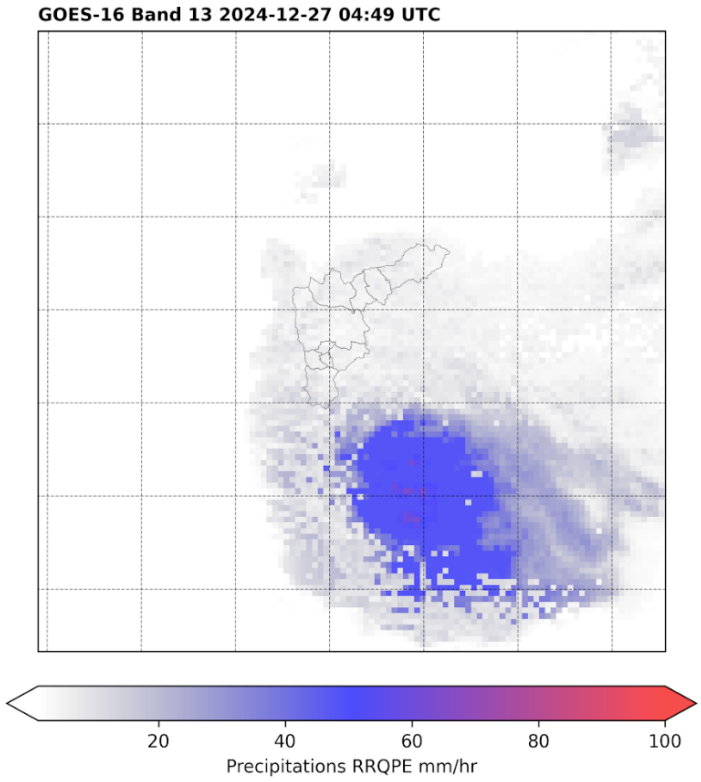

Figura 4: Visualización de evento de precipitación cercano al Valle de Aburrá según RRQPE del GOES-16. Elaboración propia.

ARQUITECTURA ELEGIDA: UNETS

La arquitectura U-Net, propuesta originalmente por Ronneberger et al. en 2015 para tareas de segmentación de imágenes biomédicas, se caracteriza por su diseño simétrico en forma de “U” como se observa en la **Figura 5.** Este combina un proceso de codificación, en el que se aplican operaciones de convolución y max-pooling para extraer características relevantes, con un proceso de decodificación que restaura la resolución espacial mediante upsampling y conexiones residuales (skip connections). Estas conexiones permiten transferir información de alta resolución desde las capas iniciales del codificador hacia las capas correspondientes del decodificador, lo cual es fundamental para lograr una localización espacial precisa.[20].

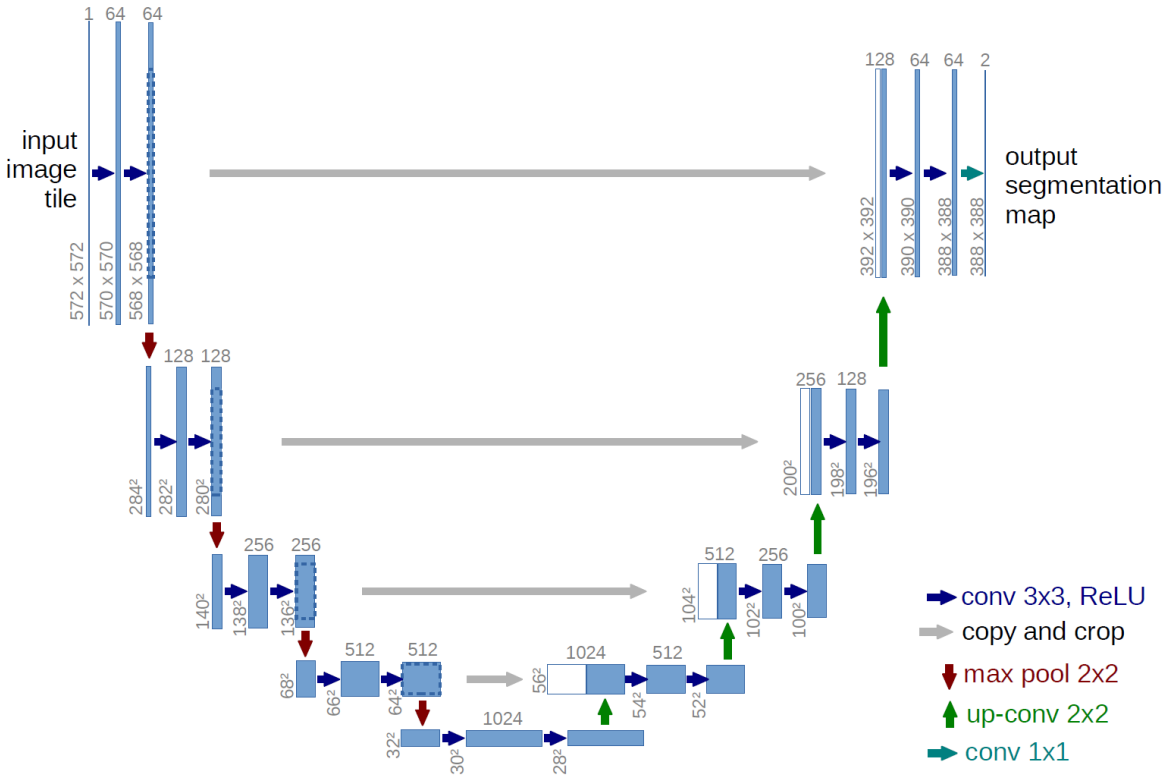

Figura 5: Esquema de la UNET Propuesta inicialmente por Ronneberger et al. en 2015. [20]

Para adaptar U-Net como regresor, se realizan ajustes fundamentales en su estructura y objetivos de aprendizaje. En la capa final, se elimina la función de activación softmax o sigmoide, que permite usarlo como clasificador, y se sustituye por una función lineal (sin activación), lo que permite al modelo generar salidas continuas, que permiten la implementación a modo de regresor. Esto implica reducir los canales de salida a uno, ya que la tarea de regresión requiere predecir un único valor numérico por píxel (para este caso, en unidades de mm/h) . Simultáneamente,

se redefine la función de pérdida para optimizar métricas de error continuo: el error cuadrático medio (MSE).[21]

Este enfoque ha abierto nuevas posibilidades en el análisis de fenómenos geoespaciales, donde la arquitectura no solo predice valores numéricos, sino que conserva la coherencia estructural.[22]

DISEÑO METODOLÓGICO

Bajo la metodología CRISP-DM, los pasos de la metodología fueron los siguientes:

ENTENDIMIENTO DEL NEGOCIO Y DE LOS DATOS

Debido a la alta necesidad de disponibilidad y a la baja latencia requerida en el caso de uso, se optó por emplear los datos de la capa RRQPE de las mediciones del satélite GOES-16, puesto que el tiempo de espera entre la toma de una imagen y su disponibilidad en AWS puede esperarse que no supere los 20 minutos.

Para obtener la data del entrenamiento, primero se probó la conexión al servicio de almacenamiento en AWS y se listaron los contenidos de la carpeta asociada a las mediciones RRQPE. Al validar la nomenclatura de los archivos y la estructura de carpetas, se observó que todos los archivos históricos se pueden acceder siguiendo la estructura:

$$\text{"noaa} - \text{goes16/ABI} - \text{L2} - \text{RRQPEF}/\mathbf{YEAR}/\mathbf{DAY}/\mathbf{MONTH}/\text{OR_ABI} - \text{L2} - \text{RRQPEF} - \text{M6_G16_s}\boldsymbol{DATE_{START}}\text{_e}\boldsymbol{DATE_{END}}\text{_c}\boldsymbol{DATE_{SAVED}}.\text{nc"}$$

Donde:

- $\boldsymbol{YEAR}$ es el año en formato de cuatro dígitos.
- $\boldsymbol{DAY}$ es el día del año en formato numérico entre 1 y 366, completado a tres dígitos.
- $\boldsymbol{MONTH}$ es el mes del año en formato de dos dígitos.
- $\boldsymbol{DATE_{START}}$, $\boldsymbol{DATE_{END}}$, $\boldsymbol{DATE_{SAVED}}$ son, respectivamente, las fechas de inicio del escaneo superficial, fin del escaneo superficial, y fecha de guardado del archivo. Cada una es una secuencia de 14 caracteres que sigue el patrón:
    - 4 dígitos para el año
    - 3 dígitos para el día del año (entre 1 y 366)
    - 2 dígitos para la hora
    - 2 dígitos para los minutos
    - 2 dígitos para los segundos
    - 1 dígito para las décimas de segundo

  Todo en horario GMT+0.

Un ejemplo de archivo, sería:

$$noaa - goes16/ABI - L2 - RRQPEF/\mathbf{2024}/\mathbf{362}/\mathbf{04}/OR_ABI - L2 - RRQPEF - M6_G16_s\mathbf{20243620420203}_e\mathbf{20243620429511}_c\mathbf{20243620429597}.nc$$

Lo cual se interpretaría como el archivo que consolida el día 362 del año 2024, a las 04:29:59.7 GMT+0, en el que se realizó un muestreo entre las 04:20:20.3 y las 04:29:51.1.

PREPARACION DE LOS DATOS

Una vez identificado el patrón, se listaron los elementos registrados desde enero del 2024 a enero del 2025 en una tabla, la cual se almacenó como un archivo PARQUET que facilitaría más adelante futuros exportes, como se observa en la **Tabla 1**.

Campo	Valor Ejemplo
files	noaa-goes16/ABI-L2-RRQPEF/2024/362/04/OR_ABI-L2-RRQPEF-M6_G16_s20243620420203_e20243620429511_c20243620429597.nc
sample_start	2024/12/27 04:20:20
sample_finish	2024/12/27 04:29:51
file_creation	2024/12/27 04:29:59

Tabla 1: Ejemplo de fila en la tabla construida de imágenes satelitales crudas.

Inicialmente, dentro de los insumos a utilizar, se había planteado definir un espacio para data cruda y otro para data procesada. Sin embargo, debido al gran volumen de la información cruda, que llegó a superar los 100 GB, se decidió procesar los archivos de forma individual recortando el área correspondiente a un cuadrado centrado en la ubicación del radar meteorológico del SIATA en Santa Elena, con un alcance similar al reportado por dicha institución.

El radar ubicado en Santa Elena tiene un alcance máximo de 120 km. No se eligió el alcance del radar porque se pretende aplicar redes neuronales U-Net, que funcionan mejor con imágenes cuyas dimensiones sean potencias de 2. Y, en zonas cercanas al ecuador, cada píxel de las imágenes satelitales equivale a aproximadamente 2 km × 2 km. Con esto en mente, se realizó un recorte de la capa RRQPE del GOES-16, centrado en Santa Elena, con un tamaño de 128 × 128 píxeles, ya que el diámetro del círculo que cubre el alcance máximo del radar (≈240 km) es mayor, y 128 pixeles cubren adecuadamente el diámetro de interés de ≈240 km, pero optimizado al tamaño base 2.[23]

Debido a la falta de librerías especializadas en las necesidades de ingesta de información expuestas. Fue necesario utilizar los Scripts de Procesamiento de datos satelitales geoestacionarios Desarrollados por Souza, Diego y Batista, Rogeiro en el Instituto Nacional de Investigación espacial del Brasil y disponibilizados en [24]. De esta manera se logra tener recortes del área de interés del sobrevuelo satelital de los satélites GOES-16, como se observa en la **Figura 4**.

La versión recortada de los archivos fue consolidada en archivos .H5 con una estructura simple de llave-valor, donde las llaves corresponden a las fechas de creación de los archivos y los valores son arreglos de tipo NumPy, en los que cada

entrada representa un píxel de un mapa de calor de precipitaciones. Los datos que no corresponden a valores flotantes positivos se imputan como 0, y se establece un valor máximo de 100, imputando con este valor cualquier entrada que lo supere. Esta decisión se tomó debido a que los valores superiores a 100 mm/h son extremadamente atípicos, y su truncamiento facilita el seguimiento del comportamiento de los datos durante los procesos de normalización en el entrenamiento del modelo.

Finalmente, siguiendo la estructura de datos propuesta por Trebing K, Stańczyk T, y Mehrkanoon S en [25], Para ello, se creó un nuevo archivo en formato .H5, separado en conjuntos de entrenamiento y validación. Cada entrada del conjunto de entrenamiento consiste en 12 frames consecutivos utilizados como predictores, bajo la condición de que al menos el primero de ellos tenga más del 20 % de sus valores con precipitaciones mayores a cero. Como variables objetivo se consideran los 6 frames siguientes en la secuencia. Sobre esta propuesta, debido a lo grande que es el área que se utilizó en proporción al área metropolitana, se aplicó esta regla del 20% específicamente en el cuadrado más central de 64 pixeles x64 pixeles, como se observa en la **Figura 9**.

MODELAMIENTO DE DATOS

A partir del dataset compuesto por 12 imágenes predictoras y 6 imágenes objetivo, se entrenaron dos configuraciones de redes neuronales tipo U-Net, utilizando el framework Lightning para PyTorch: una con 3 niveles de profundidad con la arquitectura descrita en la **Tabla 2** y otra con 4 niveles de profundidad con la arquitectura descrita en la **Tabla 3**. Estas dos arquitecturas fueron comparadas contra un modelo de persistencia, el cual consiste en predecir que las imágenes objetivo son idénticas al último frame del conjunto de imágenes predictoras.

Capa	Tipo	Entrada	Salida	Tamaño de Kenel	Stride	Padding
inc	DoubleConv	Canales = 12	Filtros Iniciales = 48	3	1	1
down1	Down	Filtros Iniciales	Filtros Iniciales*2	MaxPool2d(2) + Conv 3x3	2+1	0+1
down2	Down	Filtros Iniciales*2	Filtros Iniciales*4	MaxPool2d(2) + Conv 3x3	2+1	0+1
down3	Down	Filtros Iniciales*4	Filtros Iniciales*8//2	MaxPool2d(2) + Conv 3x3	2+1	0+1
up1	Up	Filtros Iniciales*8	Filtros Iniciales*4//2	Upsample + Conv 3x3	2+1	-1
up2	Up	Filtros Iniciales*4	Filtros Iniciales*2//2	Upsample + Conv 3x3	2+1	-1
up3	Up	Filtros Iniciales*2	Filtros Iniciales	Upsample + Conv 3x3	2+1	-1
outc	OutConv	Filtros Iniciales	Predicciones Deseadas	1	1	0

Tabla 2: Arquitectura implementada de Unet de 3 Capas.

Capa	Tipo	Entrada	Salida	Tamaño de Kenel	Stride	Padding
inc	DoubleConv	Canales = 12	Filtros Iniciales = 48	3	1	1
down1	Down	Filtros Iniciales	Filtros Iniciales*2	MaxPool2d(2) + Conv 3x3	2+1	0+1
down2	Down	Filtros Iniciales*2	Filtros Iniciales*4	MaxPool2d(2) + Conv 3x3	2+1	0+1
down3	Down	Filtros Iniciales*4	Filtros Iniciales*8	MaxPool2d(2) + Conv 3x3	2+1	0+1
down4	Down	Filtros Iniciales*8	Filtros Iniciales*16//2	MaxPool2d(2) + Conv 3x3	2+1	0+1
up1	Up	Filtros Iniciales*16	Filtros Iniciales*8//2	Upsample + Conv 3x3	2+1	-1
up2	Up	Filtros Iniciales*8	Filtros Iniciales*4//2	Upsample + Conv 3x3	2+1	-1
up3	Up	Filtros Iniciales*4	Filtros Iniciales*2//2	Upsample + Conv 3x3	2+1	-1
up4	Up	Filtros Iniciales*2	Filtros Iniciales	Upsample + Conv 3x3	2+1	-1
outc	OutConv	Filtros Iniciales	Predicciones Deseadas	1	1	0

Tabla 3: Arquitectura implementada de Unet de 4 Capas.

EVALUACIÓN

Se definió un umbral que permitiera binarizar los píxeles tanto de las imágenes reales como de las predichas, diferenciando entre presencia y ausencia de eventos de precipitación significativa. A continuación, debido al alto desbalance entre píxeles con y sin lluvia significativa, se empleó como métrica de evaluación el F1-Score, dada su capacidad para reflejar de manera precisa el desempeño en conjuntos de datos desbalanceados. Esta métrica de evaluación está dada por

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{1}$$

Dónde

$$precision = \frac{TP}{TP + FP} \tag{2}$$

$$recall = \frac{TP}{TP + FN} \tag{3}$$

A su vez siendo estas variables los **componentes** de la matriz de confusión de un proceso de clasificación como se describe en la **Tabla 4**.

		Valor Real	
		Positivo	Negativo
Valor predicho	Positivo	TP	FP
	Negativo	FN	TN

Tabla 4 Matriz de Confusión.

Finalmente, se analizó cuántos frames pueden predecirse manteniendo una relación adecuada entre la calidad de la evaluación y la duración de la ventana temporal predicha.

DESPLIEGUE

Se publicó un repositorio en GitHub en [26], que contiene todos los códigos desarrollados, junto con un instructivo para su despliegue en distintas infraestructuras. Este incluye la posibilidad de configurar el uso de recursos de cómputo especializados para redes neuronales, tanto en tarjetas gráficas NVIDIA, como en entornos con capacidades de cómputo de Apple y en sistemas basados únicamente en CPU.

RESULTADOS

Preliminarmente se realizaron visualizaciones de eventos en un área más grande de la que se estudiaría para validar la consistencia de los datos contra la realidad nacional. Como se evidencia en la **Figura 6,** los eventos de precipitación aparecen muy bien definidos. Además la presencia de eventos constantes en la costa del Pacífico concuerda con las estadísticas que afirman que esta es una de las regiones con más precipitaciones en el planeta [27].

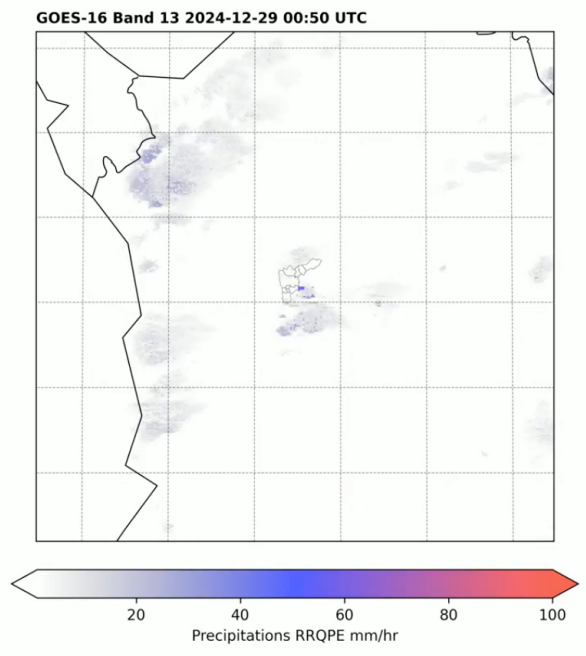

Figura 6: Grafica de elaboración propia de los eventos de precipitación del 29 de diciembre a las 00:50 GMT 0 centrados en el Valle de Aburrá pero visualizando gran parte del territorio nacional colombiano donde se observan diferentes eventos de precipitación en la región.

Al analizar los archivos de escaneos satelitales, se observó que en la gran mayoría de los casos las muestras están separadas por intervalos regulares de aproximadamente 10 minutos, como se observa en la **Figura 7**. Se identificó un evento atípico con una interrupción de los datos de alrededor de 700 minutos, pero

fuera de este caso, el 99.55 % de las muestras presenta una separación temporal entre 9 y 11 minutos.

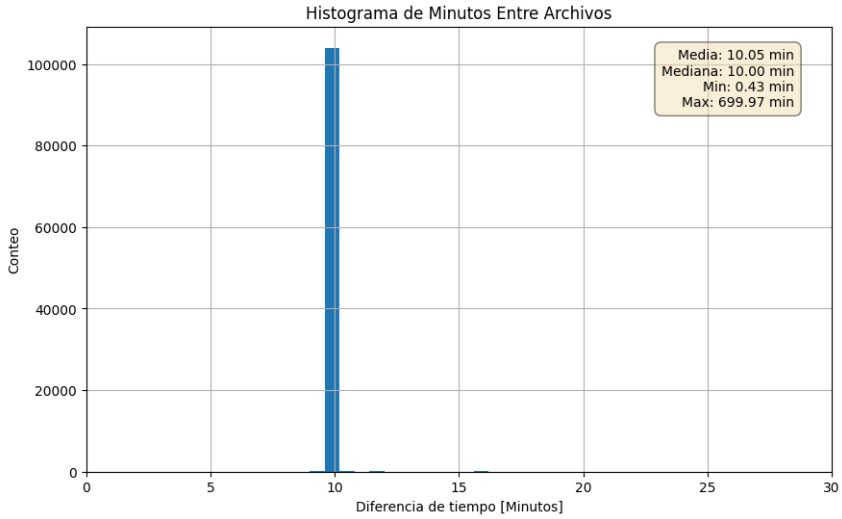

Figura 7: Histograma de tiempos entre muestras demostrando una alta consistencia temporal.

Como se observa en la **Figura 8**, se evidencia que las precipitaciones históricas en la región de estudio presentan una distribución bimodal, con picos de frecuencia en torno a los 5 mm/h y 50 mm/h.

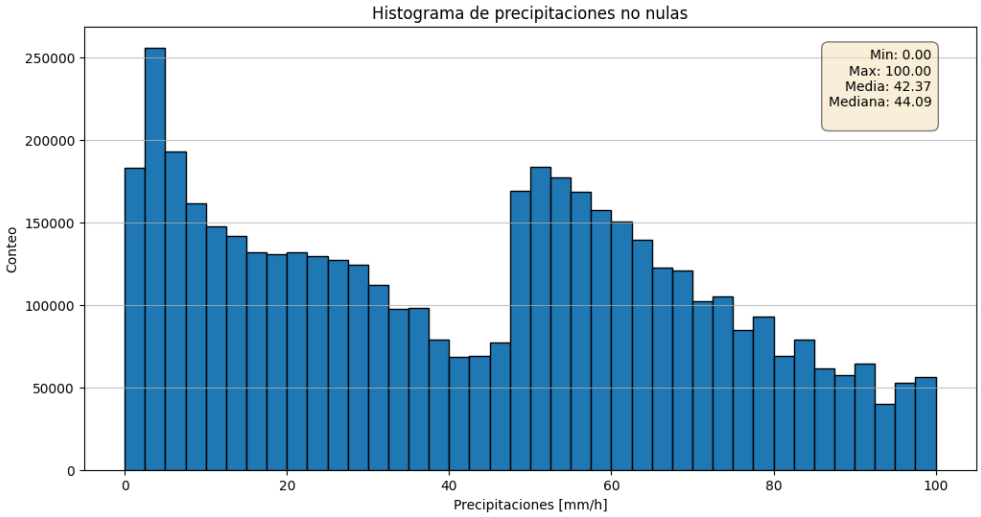

Figura 8: Histograma de valores medidos de precipitación en el dataset. Se observa una distribución bimodal de los datos alrededor de los 5mm/h y 50mm/h

Adicionalmente, como se observa en la **Figura 9**, al evaluar visualmente la construcción de los objetos del conjunto de entrenamiento, se verificó que se cumple

la condición establecida: el primer frame de cada secuencia predictora presenta un porcentaje significativo de píxeles centrales con presencia de precipitación.

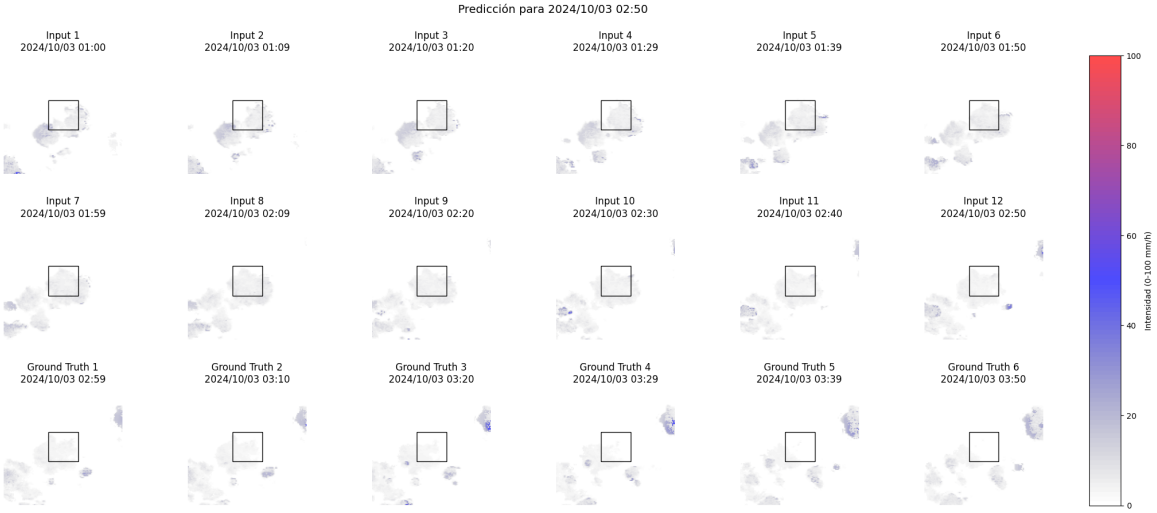

Figura 9: Ejemplo de episodio de precipitación del 3 de octubre del 2024. Donde el cuadrado interior de 32x32 pixeles la figura con numero de pixeles base 2 más pequeña capaz de encerrar todo el valle de Aburrá.

Para el desarrollo de los modelos, se usó la nomenclatura descrita en la **Tabla 5** para cada uno de ellos.

Nombre	**Modelo**
LAST12	Modelo de persistencia del doceavo frame de las imágenes predictoras.
UNET3	Unet de 3 niveles
UNET4	Unet de 4 Niveles

Tabla 5: Nombre de cada modelo desarrollado.

Durante el entrenamiento del modelo, se eligió como función de perdida para regresión el Error Cuadrático Medio.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i) \tag{4}$$

Lamentablemente, se observó que al intentar predecir más de un frame a futuro, el modelo perdía la capacidad de identificar correctamente que el valor de fondo correspondía a 0 mm/h de precipitación. Debido a esto, fue necesario limitar el modelo a la predicción de un solo frame futuro.

Posteriormente, se buscó identificar un umbral de precipitación que permitiera, mediante la binarización de las imágenes, predecir de forma más efectiva si en cada región de la imagen se presentaría un evento por encima de dicho umbral.

En los esfuerzos iniciales, no se lograba con ningún umbral, obtener un mejor desempeño que el modelo de persistencia. Se encontró, como se observa en la **Figura 10** y la **Figura 11**, que esto se debía a que la distribución de datos predichos por el modelo, no conservaba los órdenes de magnitud de los datos de entrada ni de los datos objetivo. Por lo que se procedió a escalarlos aprovechando el acotamiento inicial que se hizo de los datos a valores de 0 a 100mm/h.

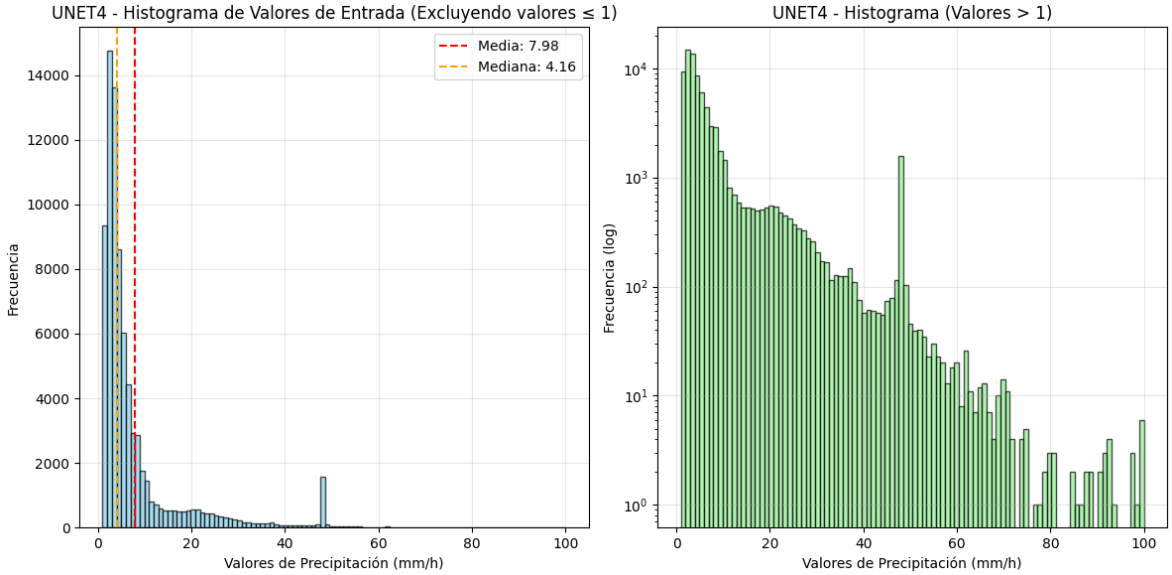

Figura 10: Distribución de datos de precipitación en pixeles verdaderos de salida en dataset para la UNET de 4 niveles en escala natural y en escala logarítmica. Se observa que la distribución de datos presenta una cola pesada que llega hasta donde inicialmente se truncaron los datos del dataset..

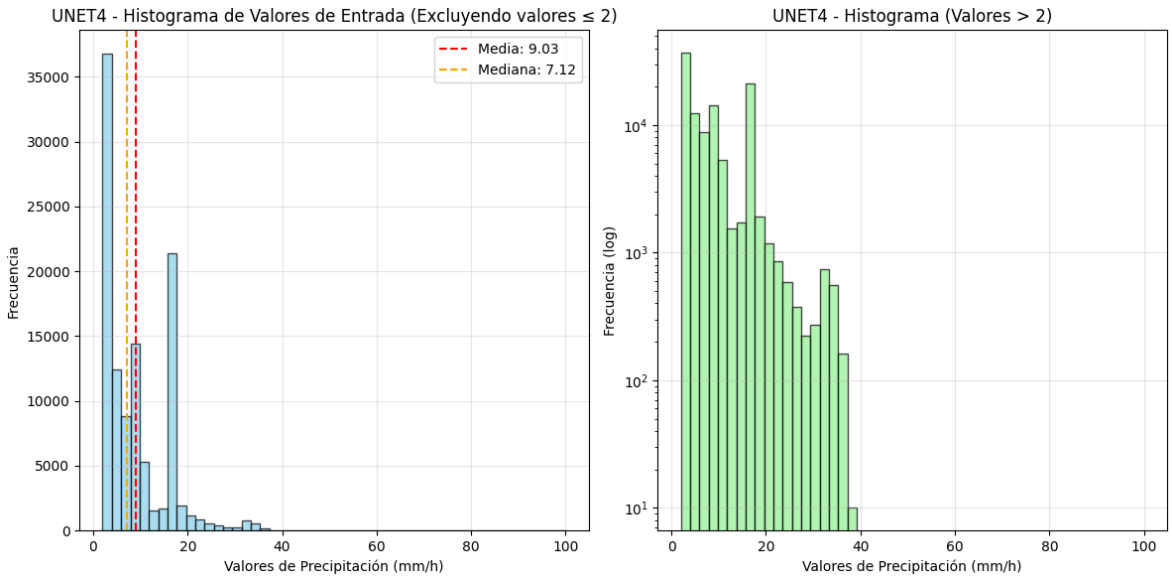

Figura 11: Distribución de datos originalmente predichos por la UNET de 4 niveles. Se observa que se truncan aproximadamente en 40mm/h, mostrando que la distribución de datos de salida predichos no se corresponde con la distribución de datos de salida reales.

Luego de escalar adecuadamente los datos predichos, se realizó un barrido sobre distintos niveles de precipitación, determinando que un valor de 20 mm/h ofrecía los mejores resultados. Según el F1-Score, este valor permite definir de forma más precisa un umbral para la presencia de precipitaciones significativas. Además, se alinea con lo que convencionalmente se considera una precipitación fuerte, por lo que se adoptó como referencia para las evaluaciones posteriores.

Finalmente, se evaluó cómo se comportan las predicciones al utilizar una ventana de un solo frame a futuro, binarizando los eventos de precipitación con el umbral de 20 mm/h. Obteniendo las métricas de desempeño presentadas en la **Tabla 6**. Según estos resultados, el mejor modelo evaluado es el UNET4. Observando en la **Figura 12** uno de los eventos de precipitación predichos por cada uno de los 3 modelos.

Modelo	**F1 Score -20mm/h**
LAST12	0.240
UNET3	0.233
UNET4	0.325

Tabla 6: Métricas finales para los diferentes modelos. Se observa que el modelo UNET4 logra predecir en ventanas de un frame(15 minutos a futuro) la evolución de los eventos de precipitación, mucho mejor que el baseline LAST12. El modelo UNET3 no logra superar el baseline.

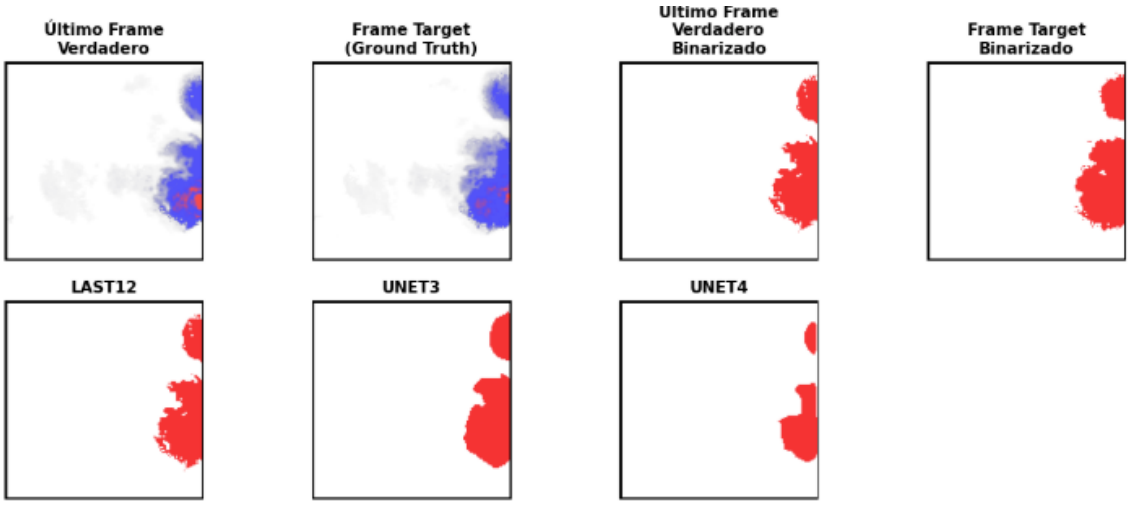

Figura 12: Ejemplo de Predicciones de evolución de precipitaciones con los diferentes modelos usados.

CONCLUSIONES

Se encontró que es posible realizar predicciones de eventos espaciales asociados a la precipitación sobre Medellín y el Valle de Aburrá mediante modelos basados en redes neuronales convolucionales, particularmente, arquitecturas tipo U-Net. Cumpliendo el objetivo de evaluar la capacidad de los modelos basados en Machine Learning para predecir eventos de precipitación. Aunque la predicción de los eventos espaciales fue posible, los beneficios obtenidos frente a un modelo de persistencia (que asume la invariabilidad de los patrones espaciales) fueron mínimos.

Estos resultados, aunque limitados, abren una discusión sobre la aplicabilidad y las limitaciones de este tipo de arquitecturas para la tarea de nowcasting en el contexto específico de Medellín y el Valle de Aburrá.

En particular, sólo el mejor modelo, UNET4, logró superar el desempeño del enfoque de persistencia, un hito importante, pero solo al predecir un único frame a futuro. Esta predicción equivale a una ventana promedio de aproximadamente 10 minutos.

Este hallazgo se convierte en una de las conclusiones más importantes del trabajo, al delimitar la ventana de predicción real que puede ser lograda bajo las condiciones actuales de procesamiento de datos.

Esta ventaja temporal resulta insuficiente para aplicaciones prácticas, dado que el proceso de almacenamiento en la nube del escaneo satelital puede tardar entre 10 y 15 minutos, lo que imposibilita un uso operativo en tiempo real.

La limitada ganancia de UNET4 sobre el modelo de persistencia, puede atribuirse a la elección de la función de pérdida MSE, la cual penaliza los errores grandes en mayor medida que los pequeños. Este comportamiento explicaría las dificultades del modelo para aprender que la mayoría de los píxeles corresponde a zonas sin precipitación, lo que es un desafío en la tarea de predicción.

No obstante, la contribución más significativa de esta tesis radica en el desarrollo de un proceso completo de ingesta, almacenamiento y preprocesamiento de datos satelitales on-premise, que maneja conexiones vía API con los repositorios oficiales del GOES. Este logro sienta las bases para futuras investigaciones al permitir tareas de descarga masiva e ingesta automatizada, lo que facilita el entrenamiento de modelos más robustos y el desarrollo de nuevas metodologías.

A partir de este trabajo, se abren diversas líneas de investigación para abordar las limitaciones encontradas y explorar enfoques más avanzados en la predicción de precipitación, como lo son las propuestas por Trebing et al. en [25], quienes emplean depthwise separable convolutions en lugar de convoluciones convencionales, reduciendo así el número de parámetros entrenables. A esto se suma la incorporación de mecanismos de atención en las conexiones internas de las redes U-Net, lo cual podría mejorar significativamente la capacidad del modelo para

aprender relaciones espaciales complejas. De igual forma, resulta relevante el enfoque de Kaparakis y Mehrkanoon en [9] donde se entrena un modelo multivariado utilizando imágenes satelitales en distintos espectros y otras variables obtenidas por sensores remotos. Adicionalmente se abre la posibilidad de realizar remuestreos espacial de los datos, integrando las coordenadas geográficas y las mediciones de sensores in situ, con el fin de aumentar la cantidad y calidad de la información disponible para el modelo, a la vez que se genera un mapeo más preciso entre las posiciones de los pixeles de los mapas de calor y los sitios de interés para las entidades meteorológicas y los servicios de atención a desastres.

Finalmente, se concluye que el presente estudio, al establecer las bases para la predicción de precipitación con modelos avanzados de aprendizaje automático, indica el potencial de estas tecnologías para fortalecer las capacidades de alerta temprana y toma de acciones ante los desafíos los fenómenos medioambientales.

BIBLIOGRAFÍA

[1] E. C. Paniagua Paniagua and D. Hernández García, “La perspectiva del cambio climático en el departamento de Antioquia,” *trilogía Ciencia Tecnología Sociedad*, vol. 5, no. 9, p. 115, Dec. 2013, doi: 10.22430/21457778.390.

[2] O. ’Poveda, G. ’Carvajal, L. F. ’Mesa, *Introducción al clima de Colombia*. Universidad Nacional de Colombia, 1997.

[3] W. Stolz, “Predicción del clima y modelos numéricos,” *Revista de Ciencias Ambientales*, vol. 35, no. 1, p. 34, Jun. 2008, doi: 10.15359/rca.35-1.7.

[4] J.-I. Yano *et al.*, “Scientific Challenges of Convective-Scale Numerical Weather Prediction,” *Bull Am Meteorol Soc*, vol. 99, no. 4, pp. 699–710, Apr. 2018, doi: 10.1175/BAMS-D-17-0125.1.

[5] Universidad Eafit, “Sistema de Alertas Tempranas SIATA.” Accessed: Nov. 18, 2024. [Online]. Available: https://www.eafit.edu.co/siata

[6] F. Sigrist, H. R. Künsch, and W. A. Stahel, “An autoregressive spatio-temporal precipitation model,” *Procedia Environ Sci*, vol. 3, pp. 2–7, 2011, doi: https://doi.org/10.1016/j.proenv.2011.02.002.

[7] Eco-Sustainable Informatics, “Precipitation Nowcasting with Explainable AI.” Accessed: Nov. 18, 2024. [Online]. Available: https://www.youtube.com/watch?v=S2a_YQ8XTPM

[8] B. Pradel, L. Perez, and L. Berthomier, “Cloud Cover Nowcasting with Deep Learning,” in *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, IEEE, Sep. 2020.

[9] C. Kaparakis and S. Mehrkanoon, “WF-UNet: Weather Fusion UNet for Precipitation Nowcasting,” Feb. 2023, [Online]. Available: http://arxiv.org/abs/2302.04102

[10] Sans Irene, “¿Cómo se mide la lluvia y por qué en mm?” Accessed: Nov. 18, 2024. [Online]. Available: https://www.clima.com/noticias/como-se-mide-la-lluvia-y-por-que-en-mm

[11] E. F. Lambin, “Remote Sensing and Geographic Information Systems Analysis,” in *International Encyclopedia of the Social & Behavioral Sciences*, N. J. Smelser and P. B. Baltes, Eds., Oxford: Pergamon, 2001, pp. 13150–13155. doi: https://doi.org/10.1016/B0-08-043076-7/04200-5.

[12] International Research Institute for Climate and Society, “Spatial Interpolation Methods.” Accessed: Nov. 18, 2024. [Online]. Available: https://iri.columbia.edu/~rijaf/CDTUserGuide/html/interpolation_methods.html

[13] T. Stańczyk and S. Mehrkanoon, “Deep Graph Convolutional Networks for Wind Speed Prediction,” Jan. 2021.

[14] Area Metropolitana Valle de Aburrá, “Así es el primer radar meteorológico para la gestión del riesgo.” Accessed: Jun. 08, 2025. [Online]. Available: https://www.metropol.gov.co/Paginas/Noticias/primer-radar-meteorologico-para-la-gestion-del-riesgo-que-existio-en-colombia.aspx

[15] A. ’Zapata, “Así se monitorean desde Santa Elena las nubes de Antioquia,” *https://www.elcolombiano.com/antioquia/funcionamiento-del-radar-metereologico-de-santa-elena-las-nubes-de-antioquia-JC14349990*, 2020. Accessed: Jun. 08, 2025. [Online]. Available: https://www.elcolombiano.com/antioquia/funcionamiento-del-radar-metereologico-de-santa-elena-las-nubes-de-antioquia-JC14349990

[16] SIATA, “Radar.” Accessed: Jun. 08, 2025. [Online]. Available: https://siata.gov.co/siata_nuevo/

[17] NOAA, “GOES East (GOES 16).” Accessed: Jun. 08, 2025. [Online]. Available: https://www.star.nesdis.noaa.gov/GOES/conus.php?sat=G16

[18] NOAA & NASA, “GOES-R Series Data Products.” Accessed: Jun. 08, 2025. [Online]. Available: https://www.goes-r.gov/products/overview.html

[19] NOAA & NASA, “GOES-R SERIES DOCUMENTS.” Accessed: Jun. 08, 2025. [Online]. Available: https://www.goes-r.gov/resources/docs.html

[20] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.

[21] R. ’Harish, B. ’Sandeep, C. ’Richard, M. ’John, T. ’P. Jyoteeshkumar, “PAUNet: Precipitation Attention-based U-Net for rain prediction from satellite radiance data,” 2023, Accessed: Jun. 10, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2311.18306

[22] “Roßberg Thomas”, “Schmitt Michael” “Baumann Anton,” “Probabilistic MIMO U-Net: Efficient and Accurate Uncertainty Estimation for Pixel-wise Regression,” 2023, Accessed: Jun. 10, 2025. [Online]. Available: https://arxiv.org/abs/2308.07477

[23] D. ’Lossos, “Beginner’s Guide to GOES-R Series Data.” Accessed: Jun. 09, 2025. [Online]. Available: https://noaa-goes16.s3.amazonaws.com/Version1.1_Beginners_Guide_to_GOES-R_Series_Data.pdf

[24] D. ’Batista, R. ’Souza, “geo-sat-python-mar-2021,” 2021. Accessed: Jun. 09, 2025. [Online]. Available: https://github.com/diegormsouza/Geo-Sat-Python-Mar-2021

[25] K. Trebing, T. Stańczyk, and S. Mehrkanoon, “SmaAt-UNet: Precipitation nowcasting using a small attention-UNet architecture,” *Pattern Recognit Lett*, vol. 145, pp. 178–186, May 2021, doi: 10.1016/j.patrec.2021.01.036.

[26] D. "Rangel, “unet_2,” 2025. Accessed: Jun. 10, 2025. [Online]. Available: https://github.com/drangelv/unet_2

[27] Ministerio de Ambiente y Desarrollo Sostenible and Instituto de Investigaciones Ambientales del Pacífico, “Plan integral de cambio climático Chocó - Colombia.”