

INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE OCCIDENTE

**DEPARTAMENTO DE ELECTRÓNICA, SISTEMAS E INFORMÁTICA
MAESTRÍA EN SISTEMAS COMPUTACIONALES**

**MSC2526A – PROGRAMACIÓN PARA ANÁLISIS DE DATOS
PRIMAVERA 2023**



ITESO, Universidad
Jesuita de Guadalajara

**ANÁLISIS Y GRAFICACIÓN DE LA CONCENTRACIÓN DE
MASA DE AEROSOL DE MEDICIONES DEL SENSOR
MODIS DEL SATÉLITE TERRA**

PROYECTO FINAL

Presenta:

*José Francisco Villanueva Cortés,
742015, francisco.villanueva@iteso.mx*

Tlaquepaque, Jalisco. Mayo de 2023.

RESUMEN

Se presenta una breve introducción al análisis de las bases de datos de las mediciones del sensor MODIS del satélite Terra, el cual tiene como objetivo principal comprender las variables de estudio de los aerosoles, enfocado de manera particular en el análisis y estudio de la variable que representa mediciones temporales de la *concentración de masa de los aerosoles* presente en la superficie terrestre, específicamente en la región de México. Para el desarrollo de este, se trabaja con distintos archivos (en formato .HDF) de la base de datos del sensor antes mencionado descargados desde un sitio web de la NASA (institución que mantiene en órbita el satélite antes mencionado) para un período de tiempo definido y fijo en cuatro meses, se analizan los datos contenidos en estos y se lleva a cabo una etapa de procesamiento para su futuro uso en la maestría dentro de la materia de TOG. Finalmente, se grafica tal variable de estudio promediada sobre un período temporal de algunos meses con el fin de entender el proceso de creación de mapas de zonas geográficas que representen tales datos.

TABLA DE CONTENIDOS

1. OBJETIVO DE INVESTIGACIÓN.....	4
1.1. INTRODUCCIÓN.....	5
1.2. ANTECEDENTES.....	5
1.3. JUSTIFICACIÓN.....	5
1.4. PROBLEMA	6
1.5. HIPÓTESIS.....	6
1.6. OBJETIVOS.....	7
1. Objetivo General:.....	7
2. Objetivos Específicos:.....	7
1.7. NOVEDAD CIENTÍFICA, TECNOLÓGICA O APORTACIÓN.....	7
2. OBTENCIÓN DE LOS DATOS.....	8
2.1. PROCESO DE OBTENCIÓN DE LOS DATOS	9
2.2. CONCLUSIONES	15
3. PREPARACIÓN DE LOS DATOS.....	16
3.1. PROCESO DE PREPARACIÓN DE LOS DATOS	17
3.2. CONCLUSIONES	18
4. EXPLORACIÓN DE LOS DATOS.....	19
4.1. PROCESO DE EXPLORACIÓN DE LOS DATOS	20
4.2. CONCLUSIONES	23
5. CONSTRUCCIÓN DE MODELOS	24
5.1. PROCESO DE CONSTRUCCIÓN DE MODELOS.....	25
5.2. CONCLUSIONES	25
6. PRESENTACIÓN DE RESULTADOS	26
6.1. PRESENTACIÓN DE RESULTADOS	27
6.2. CONCLUSIONES	29
7. CONCLUSIONES.....	30
7.1. CONCLUSIONES	31
7.2. TRABAJO A FUTURO	31

1. OBJETIVO DE INVESTIGACIÓN

Resumen: *En esta sección se presenta la introducción, antecedentes, justificación y problema a resolver haciendo uso de algunos de los datos obtenidos para mediciones de la concentración de masa de aerosoles del sensor MODIS del satélite Terra para su uso posterior dentro de la materia de TOG.*

1.1. Introducción

Este reporte describe a detalle la metodología de ciencia de datos utilizada para estudiar, comprender, limpiar y ordenar los datos o mediciones de la *concentración de masa de aerosoles* (*Aerosol Mass Concentration*) obtenidas por el satélite Terra en órbita el cual hace uso del sensor MODIS (*Moderate Resolution Imaging Spectroradiometer*) para llevar a cabo tales mediciones sobre la superficie terrestre.

Cabe destacar que el Trabajo de Obtención de Grado (TOG) hará uso de los datos o mediciones de la concentración de masa de aerosoles para más adelante hacer uso de ellos, junto con algunas otras variables de estudio o mediciones, tales como la profundidad óptica (*Aerosol Optical Depth* o *AOD*), como entrada de un modelo de Red Neuronal Recurrente (*Recurrent Neural Network* o *RNN*) para predicción, por lo que este trabajo tiene como finalidad tener una base de datos suficientemente madura y comprender lo que hay en ella para su correcta aplicación dentro del TOG.

1.2. Antecedentes

Dentro de la materia de TOG (IDI II), algunos trabajos de investigación han sido leídos para comprender el uso de los datos de las mediciones del sensor MODIS, particularmente de la profundidad óptica de los aerosoles (*AOD*), como podemos ver en [1], [2] [3], para investigar la influencia o los efectos que los aerosoles tienen sobre el clima a nivel global y regional, la calidad del aire y la salud humana, y los ecosistemas del planeta Tierra el cual habitamos.

En este caso, la medición de importancia para este proyecto es la *concentración de masa*, la cual también suele ser utilizada junto con la *profundidad óptica* dentro de algunos cálculos que se llevan a cabo para realizar análisis e incluso hacer correlaciones para tratar de comprender o encontrar la existencia de alguna relación entre los aerosoles con, por ejemplo, la calidad del aire o la salud de la población humana en alguna región del planeta, entre otras cuestiones.

Hablando del uso de este tipo de mediciones o datos en Redes Neuronales Recurrentes, hasta el momento de la redacción de este documento, no se ha encontrado algún trabajo de investigación relacionado con tal tópico, sin embargo, esto no significa que no exista alguno sino que, en todo caso, se requiere llevar a cabo un nuevo proceso de investigación y recopilación de material de investigación que pueda ayudar a conocer el uso de este tipo de redes neuronales artificiales para predicción utilizando datos espacio-temporales, es decir series de tiempo en el espacio, como es el caso con la *profundidad óptica* y la *concentración de masa* de los aerosoles medidos en la atmósfera terrestre.

1.3. Justificación

Dentro del trabajo de TOG, dos problemas u objetivos necesitan ser resueltos:

1. Construcción de una Red Neuronal Recurrente para predecir mediciones de características de aerosoles, tales como la concentración de masa o la profundidad óptica.
2. Creación de un mapa de riesgos sobre la salud humana del Área Metropolitana de Guadalajara generados por los aerosoles en dicha zona haciendo uso de variables de estudio como la concentración de masa o la profundidad óptica.

La justificación de llevar a cabo un análisis en el proyecto de esta materia de mediciones para la concentración de masa de aerosoles se justifica al ser un paso necesario para poder hacer uso de esta variable de estudio (y algunas otras) dentro de la materia de TOG, precisamente como entradas en un modelo de red neuronal artificial, para así poder resolver los dos problemas anteriores, así como comprender la estructura de los *datasets* necesarios que contienen esta información.

Antes de poder construir una red neuronal y alimentarla con datos (mediciones de la concentración de masa, de la profundidad óptica, tipo de aerosol, entre otras variables físicas de aerosoles), es necesario entender la manera en que tales datos están representados en las bases de datos o *datasets* que nos entrega el sensor MODIS, así como entender los valores que estas mediciones contiene, la calidad de los mismos y la conveniencia de realizar cambios o no en ellos.

1.4. Problema

El problema para resolver se basa en entender y comprender *datasets* de mediciones satelitales que puedan ser utilizados dentro de algún tipo de modelo de predicción. En este caso, el modelo que se desea utilizar es una Red Neuronal Recurrente. Más adelante, en el capítulo 5, se describe el porqué de la elección de este tipo de modelo, sin embargo, este proyecto se acota únicamente a la comprensión y análisis de los *datasets* que se pretende serán las entradas para este tipo de modelo de red neuronal.

1.5. Hipótesis

Hasta el momento de la redacción de este documento no se tiene alguna hipótesis definida para el uso de estos datos y la predicción de los mismos haciendo uso de Redes Neuronales Recurrentes (tema principal del TOG), sin embargo, al no conocer algún trabajo de investigación del uso de este tipo de redes neuronales, es posible definir ahora una hipótesis, a verificar más tarde dentro del TOG, la cual sería suponer que es posible llevar a cabo predicciones de características de aerosoles en la atmósfera terrestre, tales como las antes mencionadas (concentración de masa y profundidad óptica) a partir de los datos obtenidos del sensor MODIS.

Esta hipótesis se basa en una de las características de estos *datasets*: son series de tiempo, es decir, las mediciones cumplen con una relación de temporalidad ya que el satélite realiza mediciones diarias de la

variable de estudio de este proyecto, la concentración de masa, así como de algunas otras relacionadas con características de los aerosoles de la superficie terrestre.

1.6. Objetivos

1. Objetivo General:

El objetivo general de este proyecto es comprender las mediciones de *concentración de masa* obtenidas por el sensor MODIS y contenidas dentro de los *datasets* que pueden ser descargados de algunas plataformas de NASA.

2. Objetivos Específicos:

- 1) Comprender los valores que podemos encontrar dentro de las mediciones de la variable de estudio de interés de este proyecto, es decir, la concentración de masa de los aerosoles.
- 2) Comprender los valores que no son adecuados para la misma variable de estudio.
- 3) Entender el uso de las herramientas necesarias para su respectivo análisis.
- 4) Entender la manera en que se realiza la visualización de la misma.
- 5) Comprender el proceso general de análisis y preparación de datos satelitales.
- 6) Verificar el uso correcto de la plataforma para descargar los datos.

1.7. Novedad científica, tecnológica o aportación

La novedad de este proyecto radica en el estudio de datos satelitales de la *concentración de masa* de aerosoles obtenida mediante el sensor MODIS para el territorio de México, así como un uso futuro como entrada para un modelo de predicción.

2. OBTENCIÓN DE LOS DATOS

Resumen: *En esta sección se presenta el proceso de obtención de los datos de mediciones de la Concentración de Masa de Aerosoles del sensor MODIS del satélite Terra descargados a través del sitio web de la NASA para un posterior tratamiento de los mismos que permita hacer uso de ellos dentro de la materia del TOG.*

2.1. Proceso de obtención de los datos

Para obtener los datos del sensor MODIS del satélite Terra, es necesario acceder a la página de la NASA para *datasets* de la atmósfera **Level-1 and Atmosphere Archive & Distribution System Distributed Active Archive Center** o **LAADS** [4].

Una vez que se accede a la ruta antes referida, es necesario seleccionar el apartado **Atmosphere** para acceder a los datos de aerosoles (dentro de los cuales encontraremos la concentración de masa que necesitamos obtener):

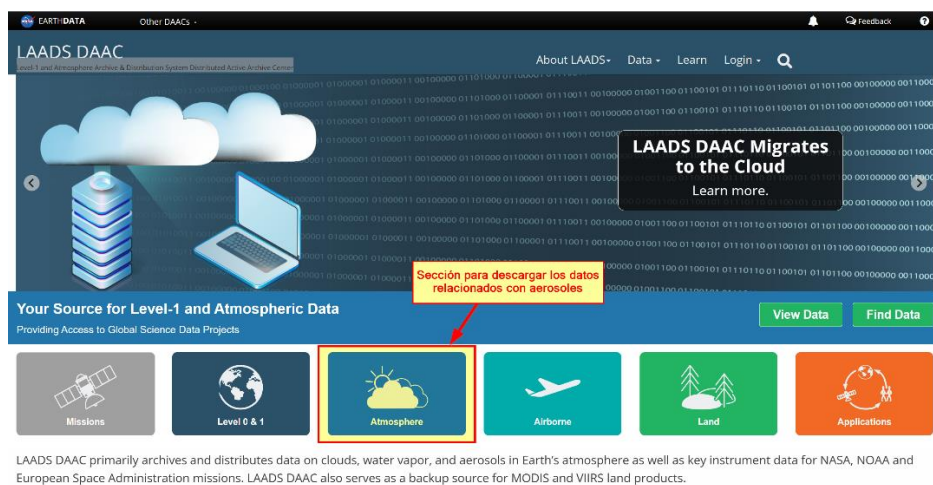


Figura 1 Página LAADS para descargar los datos relacionados con mediciones de aerosoles.

Cuando ya ha sido seleccionado el apartado **Atmosphere**, debajo aparecerá una nueva sección con el nombre **Aerosol**, lugar en el cual deberemos hacer clic para poder acceder a una nueva página donde podremos descargar los *datasets* de aerosoles generados por el sensor MODIS.

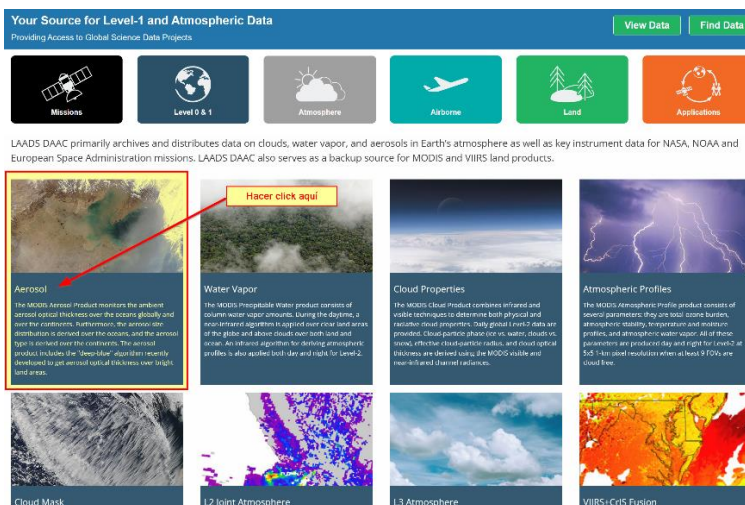


Figura 2 Seleccionar **Aerosol** para acceder al lugar donde obtendremos los datos del sensor MODIS.

Enseguida de haber seleccionado *Aerosol*, se abre una página en la cual seleccionaremos para el producto *Aerosol 5-Min L2 Swath 10km* del satélite Terra (**MOD04_L2**) la opción *Order Data*, lo cual nos permitirá acceder a una nueva página para poder escoger las características de los *datasets* que necesitemos para obtener la concentración de masa de los aerosoles:

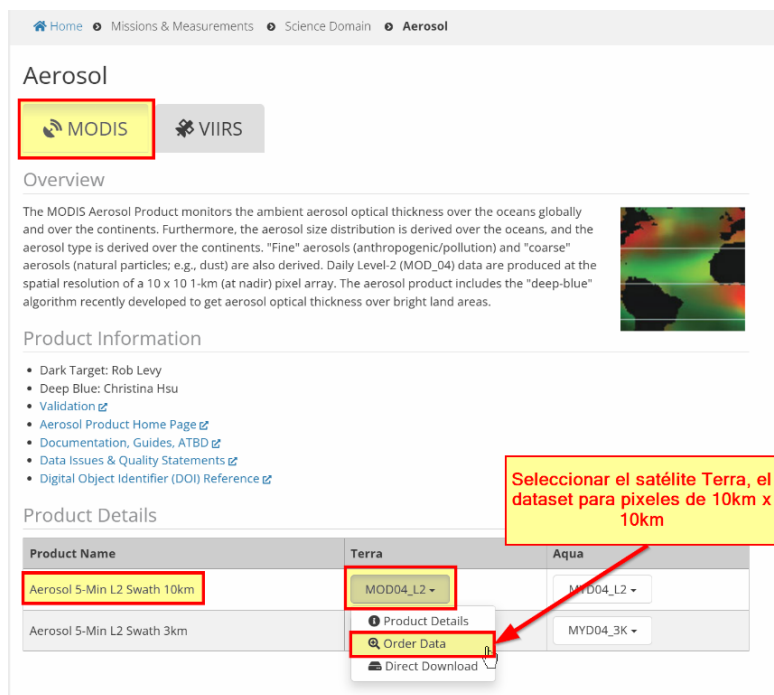


Figura 3 Selección del producto para acceder al sitio de descarga de *datasets*.

A continuación, se muestran una serie de capturas de pantalla con los siguientes pasos realizados para obtener *datasets* para un período de tiempo de 4 meses (enero-abril del 2023) para la variable de estudio, la *concentración de masa*, para datos obtenidos durante el día, del sensor MODIS del satélite Terra:

1. **Products:** Esta opción se definió automáticamente al seleccionar el producto *Aerosol 5-Min L2 Swath 10km* del satélite Terra en la página anterior, aunque es posible escoger un nuevo producto en caso de, por ejemplo, encontrarse con la necesidad de descargar un producto de con una menor resolución (3km x 3km) (en la imagen abajo se muestra cómo sería definir el producto que ya habíamos seleccionado antes en caso de no haber escogido el producto correcto, aclarando que en caso de necesitar otro producto la manera de seleccionarlo es prácticamente la misma).

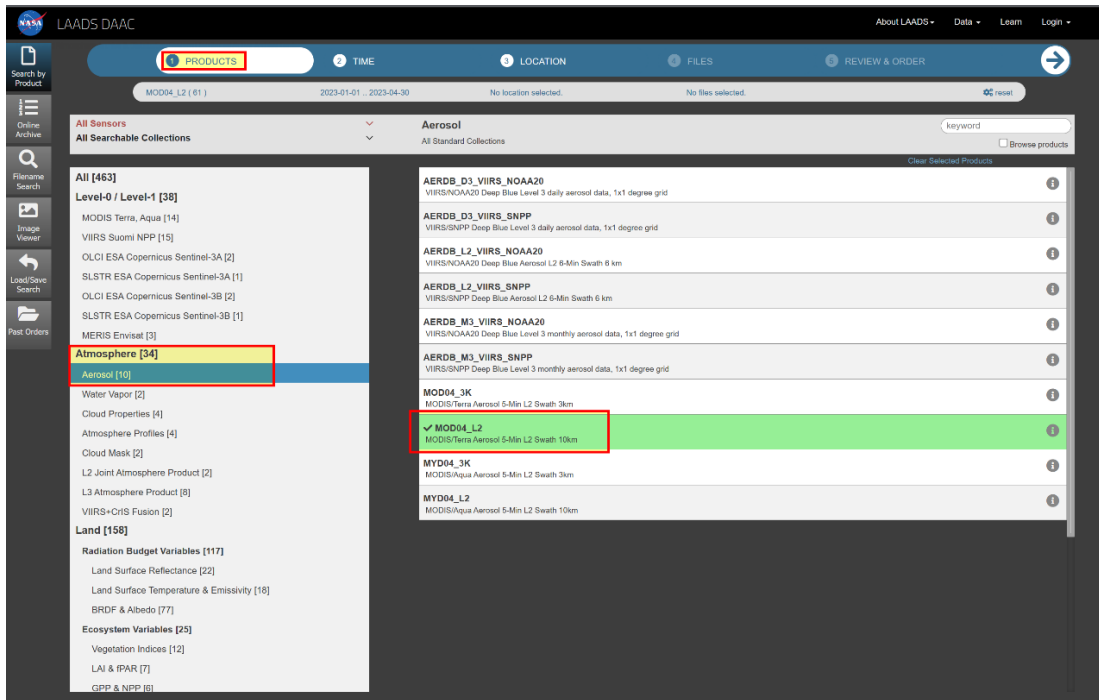


Figura 4 Opción *Products* para selección del *dataset* para MODIS Terra y otros sensores.

2. **Time:** En este apartado, se selecciona el período de tiempo para obtener datos entre los meses de inicio de enero hasta finales de abril del año 2023, lo cual da un total de datos obtenidos para el primer cuatrimestre del año 2023. Adicionalmente, se selecciona únicamente la casilla Day correspondiente a datos obtenidos durante el día.

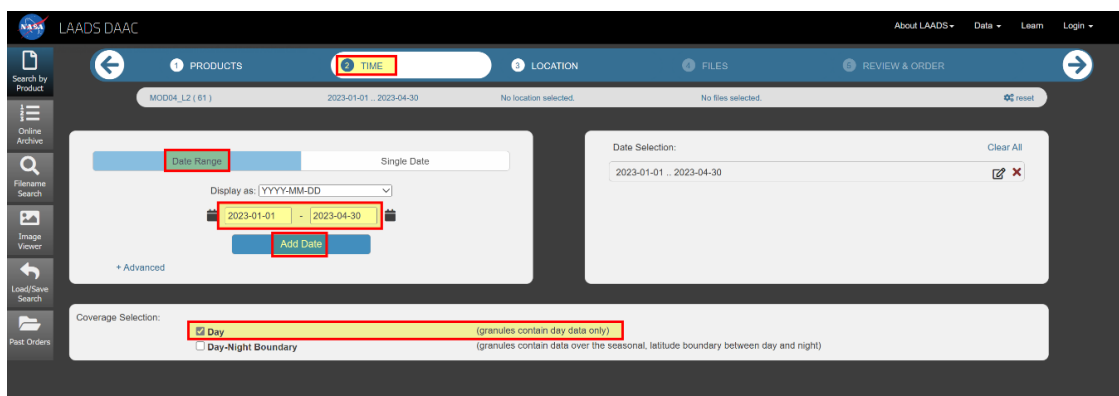


Figura 5 Opción *Time* para selección del período temporal de los datos.

3. **Location:** Esta sección se utiliza para definir la región o superficie para la cual se quieren obtener mediciones del sensor o producto seleccionado con anterioridad. Para ese caso, aunque para el TOG

la región de estudio es el Área Metropolitana de Guadalajara (AMG), se trabajará con todo México (esto debido a que el análisis de los datos resultó más sencillo ya que al seleccionar únicamente el AMG, la mayor parte de los datos en esa región contenían *NaN* lo cual hacía difícil confirmar que los datos obtenidos realmente correspondían al AMG y que realmente se estaban seleccionando los datos de la manera correcta, en cambio, para todo México, resultó más fácil hacer esta validación al observar que efectivamente se estaban obteniendo datos para la región correcta al momento de graficarlos).

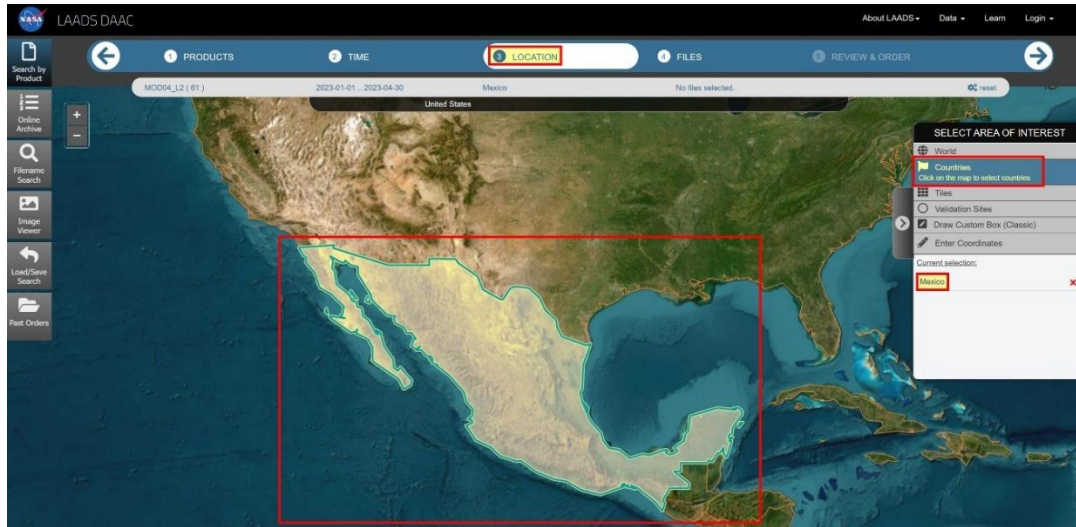


Figura 6 Opción *Location* para escoger la región para la cual se desea obtener datos.

4. **Files:** Esta sección permite seleccionar los archivos que se desea descargar. Para el análisis de los datos para este proyecto se seleccionarán todos los archivos que arrojen los resultados y a continuación, en el punto 5 (abajo), se hará un post-procesamiento para hacer más fácil el proceso de análisis y obtener realmente los datos deseados, en este caso, la concentración de masa únicamente.

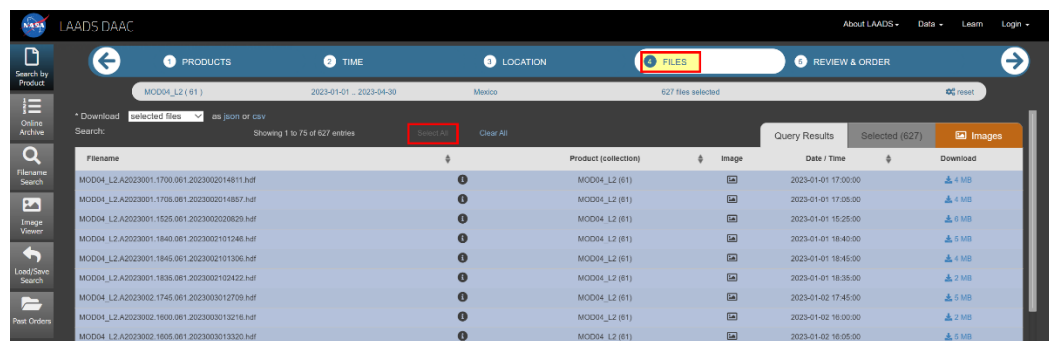


Figura 7 Opción *Files* para selección de *datasets* encontrados.

5. **Review & Order:** Esta última opción nos permite llevar a cabo una etapa de post-procesamiento para los archivos seleccionados antes. Es recomendable hacer este post-procesamiento para tener datasets con características específicas como, por ejemplo, el tipo de formato, las variables de estudio que se van a utilizar (como la concentración de masa y no la profundidad óptica), el tipo de proyección para su representación en mapas, entre otras opciones.

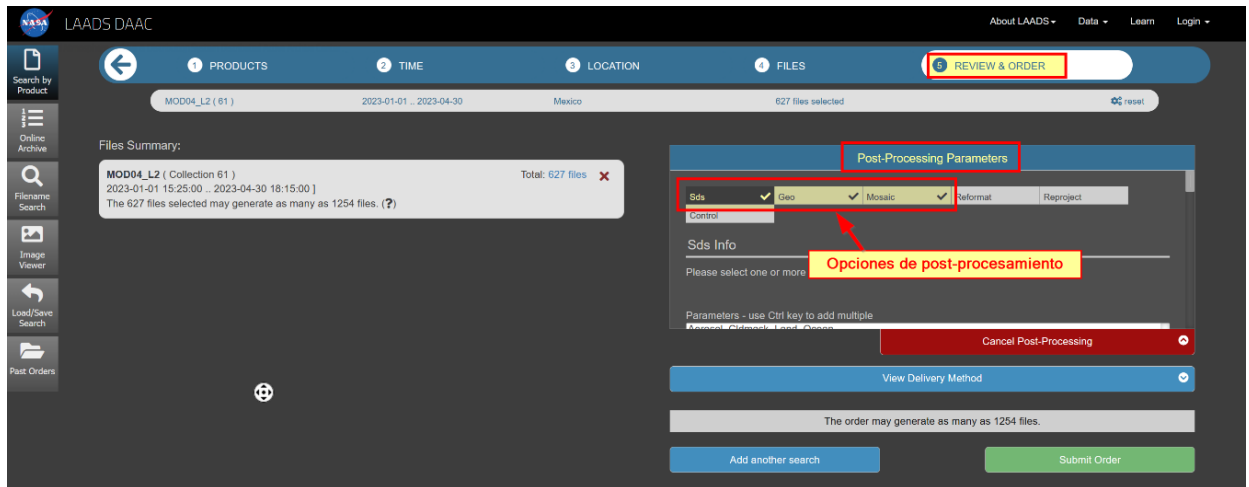


Figura 8 Opción *Review & Order* para post-procesamiento.

En el caso de este proyecto, se trabaja con las siguientes **opciones de post-procesamiento**:

- a. **Sds:** Permite seleccionar las variables de estudio (mediciones de los aerosoles) requeridas. En este caso se seleccionan dos, la variable de estudio del objeto de análisis de este proyecto, ***Mass_Concentration_Land*** (concentración de masa), y adicionalmente ***Aerosol_Type_Land*** (tipo de aerosol). Cabe mencionar que la segunda variable se seleccionó al observar, durante un primer análisis de los datos obtenidos en este sitio web, que puede ser importante relacionarla con la concentración de masa observada, así como en un futuro, con la profundidad óptica para no únicamente predecir sino tal vez llevar a cabo una clasificación de aerosoles haciendo uso de algunas otras variables que pudiesen ser importantes.

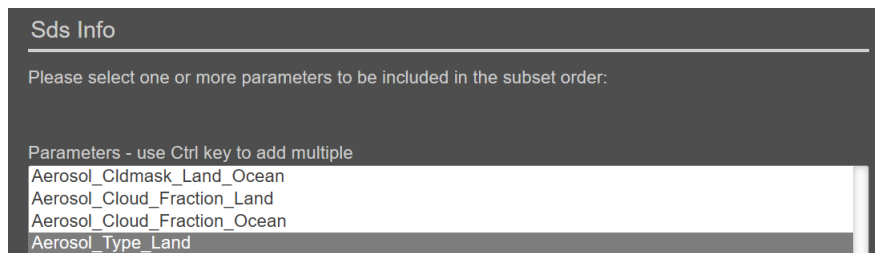


Figura 9 Opción *Sds* para selección de variables de estudio para el *dataset*.

- b. **Geo:** Permite seleccionar realmente la región de estudio. Durante un primer análisis de los datos obtenidos de este sitio, cuando no se llevaba a cabo esta etapa de post-procesamiento, se encontró que los datos obtenidos no siempre corresponden a toda la región de estudio, sino que son obtenidos más datos de los esperados. Para ello, se selecciona aquí la región de estudio definida para México (en forma de una caja, por lo cual también se obtienen datos de los océanos Pacífico y Atlántico en caso de haber seleccionado alguna variable de tipo *Ocean* en la opción *Sds*), la cual es dada por las coordenadas delimitantes al Norte, Sur, Este y Oeste en formato de longitudes y latitudes:

- i. **EastBound:** -83.5
- ii. **NorthBound:** 34.5
- iii. **SouthBound:** 14.5
- iv. **WestBound:** -121.3



Figura 10 Opción *Geo* para selección de límites espaciales para el *dataset*.

- c. **Mosaic:** Permite obtener *datasets* correspondientes a toda la región de estudio y no por partes. Cuando no se lleva a cabo el post-procesamiento, lo que se observa es que los *datasets* obtenidos en el punto 4 (*Files*) lo cual volvería más complicado el análisis de los mismos al tener que hacer la unión de los datos espaciales para una misma fecha a mano. Aquí, se selecciona la única opción disponible, , para lograr obtener *datasets* que correspondan a toda la región de estudio.

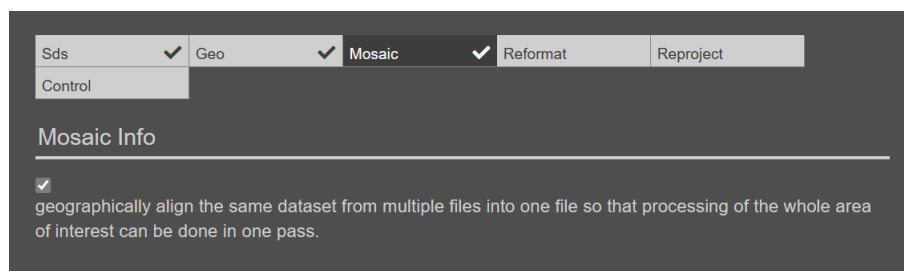


Figura 11 Opción *Mosaic* para aplicar los límites espaciales y obtener *datasets* dentro de esos límites para cada día seleccionado.

Una vez llevado a cabo este proceso, la misma plataforma LAADS generará una orden para poder descargar los datos, para así poder descargarlos tanto de manera manual como haciendo un request GET por medio de HTTP. Tal orden se ve de la siguiente manera dentro del histórico del usuario que requirió el post-procesamiento:

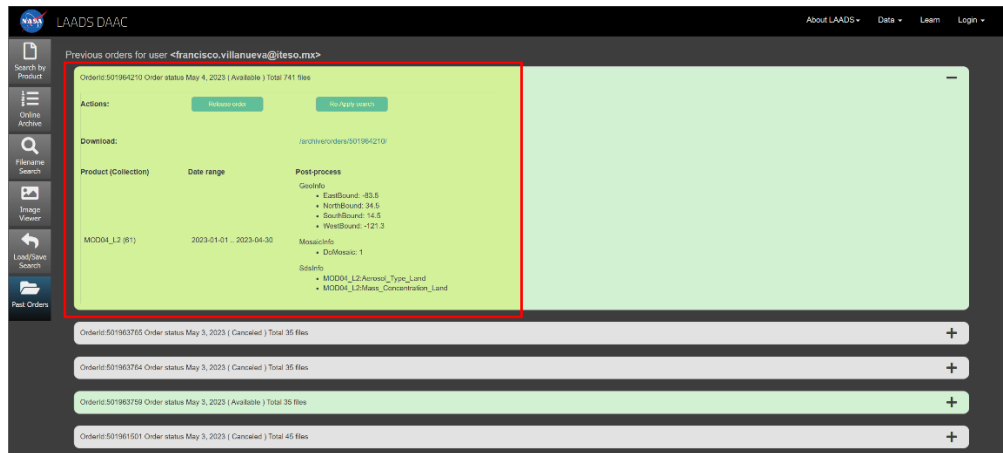


Figura 12 Orden en LAADS para descargar los datos post-procesados.

2.2. Conclusiones

Para esta etapa de post-procesamiento, se requirió llevar a cabo el proceso varias veces junto con el de los capítulos 3 y 4 siguientes hasta comprender la manera de obtener los *datasets* con los datos que realmente eran necesarios.

Cuando la etapa de post-procesamiento no se llevaba a cabo, los *datasets* obtenidos contenían demasiada información, la mayoría de la cual no era necesaria para este proyecto (e incluso para el TOG). Al mismo tiempo, esa información innecesaria se convertía en tener *datasets* demasiado pesados considerando el período de cuatro meses que se deseaba analizar en las siguientes etapas de los capítulos posteriores, así como en tiempo de descarga de los mismos.

Así mismo, la repetición del proceso sirvió para comprender mejor el uso de la plataforma LAADS, la manera en que se manejan los *datasets* dentro de ella, y las posibilidades y limitantes de la misma.

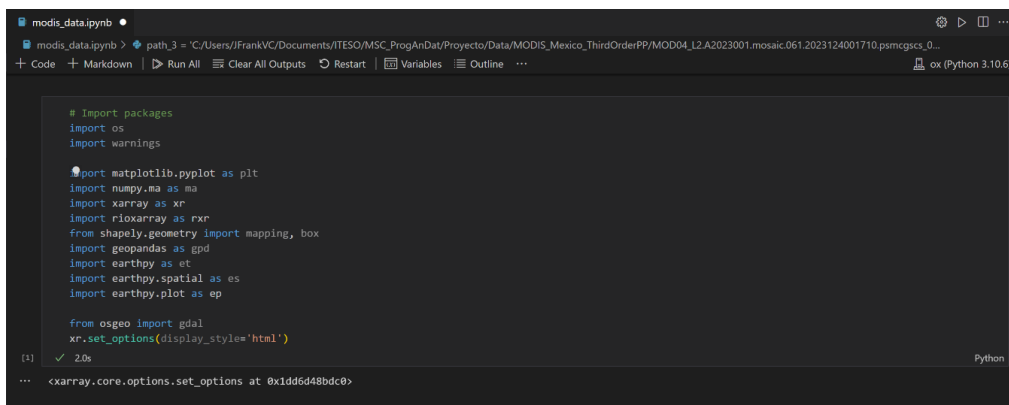
3. PREPARACIÓN DE LOS DATOS

Resumen: En esta sección se presenta el método utilizado para comprender los datos crudos de la Concentración de Masa de los Aerosoles obtenidos del sitio web de la NASA, así como un procesamiento posterior para así obtener un conjunto de archivos (o datasets) que puedan ser utilizados de manera segura y confiable para una posterior visualización y tratamiento de los mismos para así finalizar en un futuro en su uso dentro de la materia de TOG.

3.1. Proceso de preparación de los datos

Como parte del proceso de preparación de los datos, dos herramientas serán utilizadas tanto para analizarlos, comprenderlos y graficarlos, como para realizar cambios o ajustes en ellos en caso de ser requeridos:

1. **Python:** Se utilizará Python (3.10.6) para trabajar con los *datasets*, graficarlos, realizar la limpieza de los mismos, en resumen, llevar a cabo las acciones necesarias para preparar los datos para un uso en forma de input para, en este caso, un modelo de Red Neuronal Recurrente para predicción dentro de la materia de TOG. Así mismo, Python y sus librerías permitirán visualizar los mismos.



```
# Import packages
import os
import warnings

import matplotlib.pyplot as plt
import numpy.ma as ma
import xarray as xr
import rioxarray as rxr
from shapely.geometry import mapping, box
import geopandas as gpd
import earthpy as et
import earthpy.spatial as es
import earthpy.plot as ep

from osgeo import gdal
xr.set_options(display_style='html')
```

The screenshot shows a Jupyter Notebook interface with a dark theme. The code cell contains imports for various Python libraries used in geospatial data analysis. The output area shows a successful execution of the code, with a message indicating that the xarray options were set successfully.

Figura 13 Entorno de programación en Python.

- Dentro de Python, lo más importante es la comprensión de las librerías necesarias para trabajar con los *datasets* obtenidos en el capítulo 2 y basadas en las lecciones para manejo de datos geoespaciales de [5], las cuales son:
 - **Matplotlib** [6]: necesaria para la creación y visualización de imágenes.
 - **Numpy** [7]: utilizada para crear y realizar operaciones con arreglos multidimensionales, como los son las imágenes satelitales, es decir, los arreglos de píxeles.
 - **Xarray** [8]: permite hacer uso de etiquetas en la forma de dimensiones, coordenadas y atributos por encima de numpy, para dar la posibilidad de utilizar coordenadas para identificar píxeles.
 - **Rioxarray** [9]: para la lectura de *datasets* en formato HDF (Hierarchical Dataframe Format), el cual es el formato de los *datasets* obtenidos en LAADS y guardarlos haciendo uso de *xarray*.
 - **Shapely** [10]: no utilizada de momento, se trata de una librería que permite hacer uso de objetos geométricos para diseño y análisis de formas geométricas en objetos geoespaciales, como lo pueden ser mapas.
 - **Geopandas** [11]: no utilizada de momento, es una extensión de tipos de datos de *pandas* (por lo que *pandas* es implícitamente utilizada) para permitir

operaciones espaciales en objetos geométricos. Esta librería será de utilidad para análisis futuros dentro de TOG.

- **Earthpy** [12]: esta librería es utilizada para la visualización de los datos geoespaciales, es decir, los datasets obtenidos de LAADS, como se puede ver en el capítulo 6.

2. **Panoply**: Esta herramienta, en su versión 5.2.1, pertenece y es mantenida por la NASA. Con ella, el objetivo es obtener información de los datos de tal manera que sea posible comprender las variables que lo conforman, así como llevar a cabo una visualización previa. Esta herramienta se utilizará un paso antes de llevar tratamiento alguno con Python.

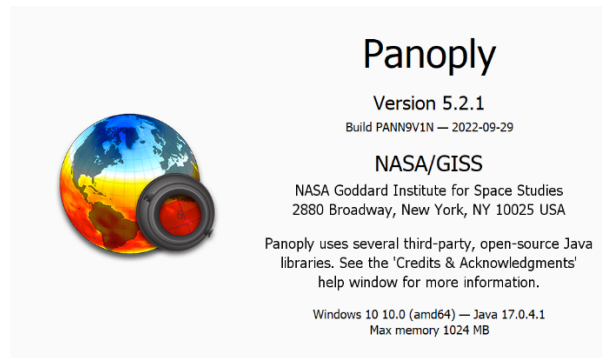


Figura 14 Aplicación Panoply.

A continuación, se describe el proceso utilizado para la preparación de datos antes de su análisis:

- Descarga de *datasets* de LAADS.
- Lectura de archivos HDF en Panoply.
 - Visualización de la variable de estudio
 - Visualización de los arreglos (matrices) de la variable de estudio
- Lectura mediante las librerías de Python.

3.2. Conclusiones

La preparación de los datos es un proceso necesario para hacer un uso correcto de ellos en algún modelo de predicción (como se tiene planteado para el trabajo de TOG mediante el uso de redes neuronales), así como para su análisis. En este caso, esta etapa es sencilla, ya que una gran parte de la preparación proviene de la misma plataforma LAADS (como se vio en el capítulo 2), sin embargo, también es importante entender el uso de las herramientas que permitirán llevar a cabo el análisis posterior, como sucede con las librerías de Python y la herramienta Panoply. En el capítulo siguiente, se podrá comprender más de esto a través del proceso de exploración el cual además permitirá comprender la naturaleza de los datos mismos.

4. EXPLORACIÓN DE LOS DATOS

Resumen: *En esta sección se presentan las herramientas utilizadas para comprender los datos ya procesados en una etapa anterior, así como para llevar a cabo la visualización de los mismos para definir si es necesaria una etapa más de post-procesamiento que permita hacer uso de los mismos, en un futuro, dentro de la materia de TOG.*

4.1. Proceso de exploración de los datos

Como proceso complementario del uso de las herramientas del capítulo 3 y el proceso brevemente descrito en el mismo, aquí se describe con precisión el proceso llevado a cabo para la exploración y comprensión de los datos de la variable de estudio (concentración de masa de aerosoles) contenida en los *datasets* (formato HDF) obtenidos de LAADS.

1. Exploración en Panoply:

- 1.1. El primer paso de exploración de los datos, después del proceso llevado a cabo en el capítulo 2, para la obtención de los datos, es abrir Panoply y abrir un archivo HDF para entender su contenido. En este caso se abre el primero de los archivos obtenidos para el período de tiempo antes descrito.

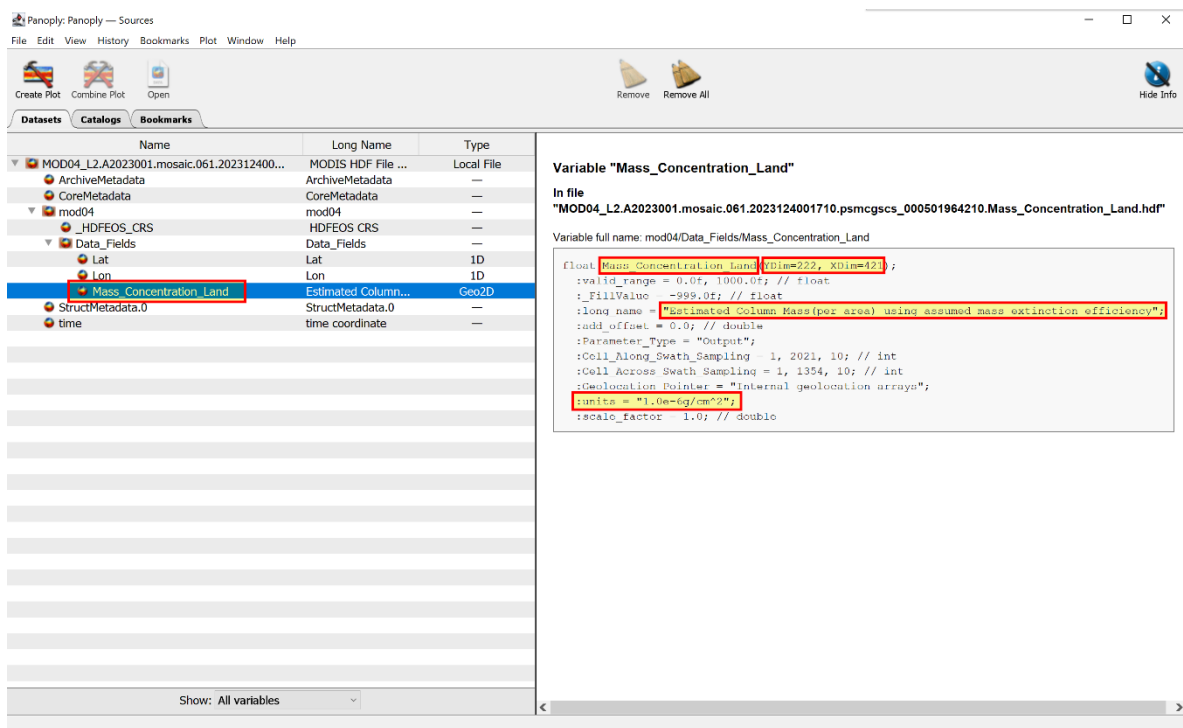


Figura 15 Archivo HDF número uno de los *datasets* obtenidos en LAADS abierto en Panoply.

Haciendo uso del primero de los 120 archivos (120 días de datos para un período de 4 meses) obtenidos en LAADS mediante Panoply, podemos observar una serie de características importantes del mismo:

- El archivo HDF contiene más información que las mediciones de concentración de masa. En sí, este archivo es una arreglo estructurado a manera de diccionario, en el cual podemos ver tanto la variable de estudio (concentración de masa en forma de arreglo bidimensional), como por ejemplo, las coordenadas a manera de arreglos unidimensionales que abarcan la región definida en LAADS al momento de descargar los *datasets*.

- También, para la variable de estudio, podemos ver las características necesarias para su análisis y posterior tratamiento, como lo son las dimensiones ($Y = 222$ por $X = 421$, dando un total de 93,041 píxeles), las unidades ($1.0e-6 \text{ g/cm}^2$), y la descripción de la variable de estudio (*Estimated Column Mass (per area) using assumed mass extinction efficiency*),

1.2. Como segundo paso, es importante comprender la forma del arreglo bidimensional de la variable de estudio, así como la visualización obtenida utilizando la misma herramienta, esto con el objetivo de validar un uso correcto de las librerías de Python.

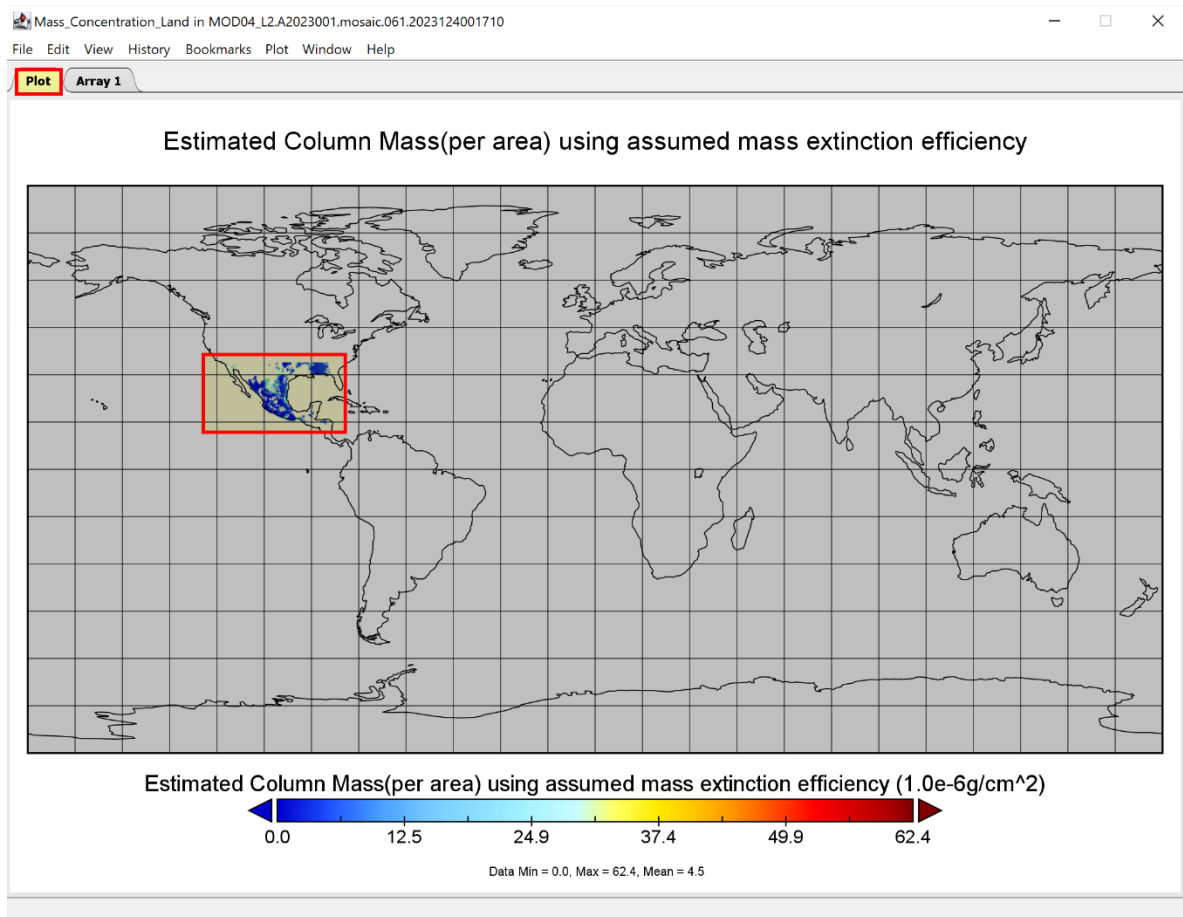


Figura 16 Mapa de generado para el *dataset* número uno de la concentración de masa.

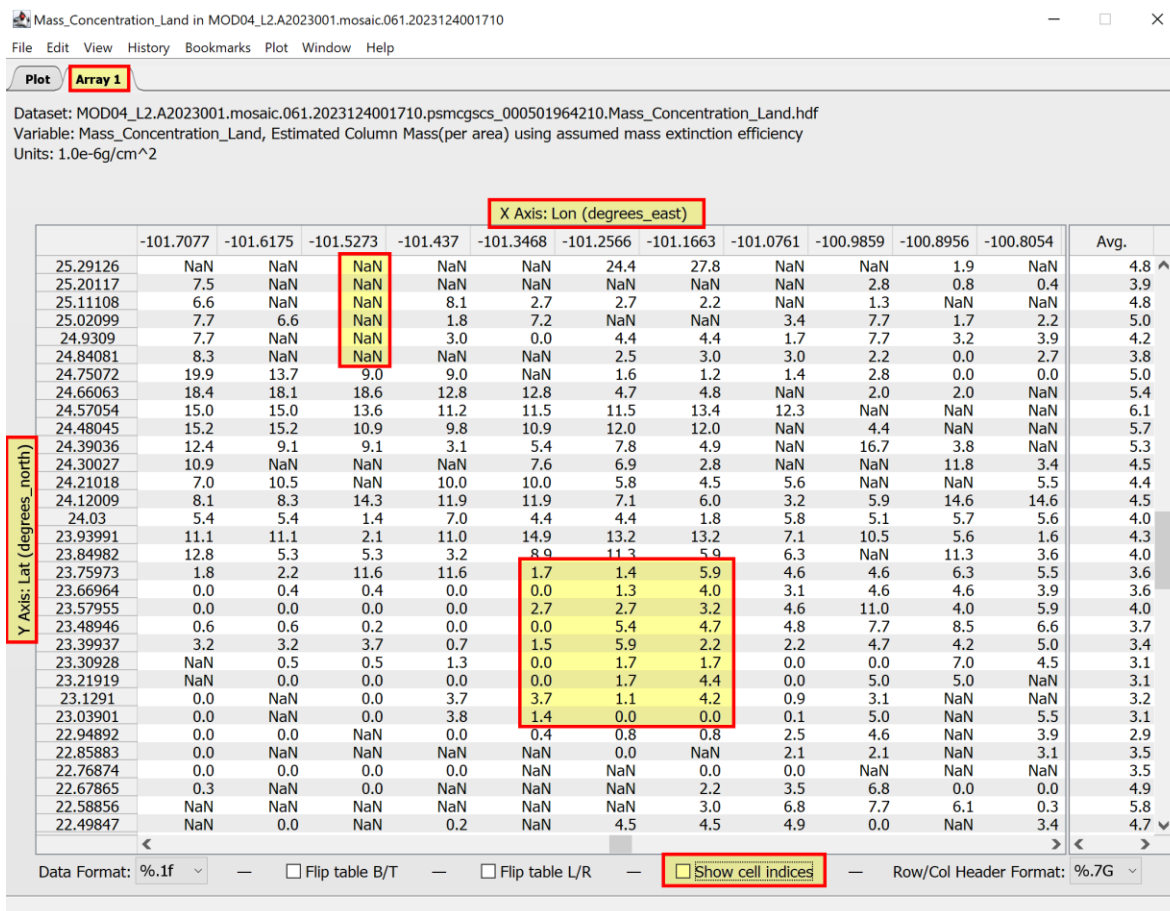


Figura 17 Arreglo del dataset número uno de la concentración de masa.

De la figura 15, es posible obtener información de gran interés para su posterior tratamiento con Python:

- Una gran cantidad de datos tiene el valor NaN, lo cual indica que las mediciones no se llevaron a cabo de la manera correcta (por cuestiones posiblemente relacionadas al clima de la superficie sensada por MODIS) o las características del post-procesamiento en LAADS.
- Los datos son de tipo flotante.
- Las dimensiones se muestran a manera de coordenadas (por lo cual la librería *xarray* de Python se vuelve de gran ayuda), aunque también es posible trabajar con índices (*Show cell indices*).
- Cada uno de los datos representa el valor de cada uno de los 93,041 píxeles.

2. Exploración en **Python**:

La lectura del *dataset* se lleva a cabo mediante *rioxarray*, y la imagen siguiente muestra la información contenida dentro de la variable *ds_copy* utilizada en este análisis para guardar el *dataset* número 1 (el mismo que se abrió con Panoply anteriormente):

```
ds_copy
✓ 0.1s

xarray.DataArray (y: 222, x: 421)

array([[ nan,    nan,    nan, ..., 2.228688, 1.262601,    nan],
       [ nan,    nan,    nan, ..., 1.800048, 1.725017,    nan],
       [ nan,    nan,    nan, ..., 1.720589, 1.539437,    nan],
       ...,
       [ nan,    nan,    nan, ..., 6.198546, 8.2982   ,    nan],
       [ nan,    nan,    nan, ..., 9.135132, 6.800264,    nan],
       [ nan,    nan,    nan, ...,      nan,      nan,    nan]],
      dtype=float32)

▼ Coordinates:
  band          ()      int32  1
  x             (x) float64 -121.3 -121.2 ... -83.63 -83.54
  y             (y) float64 34.45 34.36 34.27 ... 14.64 14.55
  spatial_ref   ()      int32  0

▼ Indexes:
  x              PandasIndex
  y              PandasIndex

▼ Attributes:
  add_offset :      0.0
  ArchiveMetadata : GROUP = ARCHIVEDMETADATA
                   GROUPTYPE = MASTERGROUP

                   OBJECT = LONGNAME
                   NUM_VAL = 1
                   VALUE = "MODIS/Terra Aerosol 5-Min L2 Swath 10km"
                   END_OBJECT = LONGNAME

                   GROUP = PROJECT

                   OBJECT = INSTRUMENTNAME
```

Figura 18 *Dataset* número uno en Python.

4.2. Conclusiones

Esta etapa permitió ver las características de los datos contenidos en los *datasets* con los que se estará trabajando (obtenidos de LAADS), así como corroborar la correcta lectura de los mismos en Python haciendo uso de las librerías descritas en el capítulo 3, lo cual permite a su vez poder comenzar la etapa de tratamiento y visualización de la concentración de masa.

5. CONSTRUCCIÓN DE MODELOS

Resumen: *En esta sección se presenta el modelo de predicción de datos que se desea construir que tome como entrada los datos trabajados en las etapas anteriores de este proyecto, sin por lo tanto construir el modelo en sí, sino únicamente entender el alcance y validar que la comprensión de los datos es la correcta.*

5.1. Proceso de construcción de modelos

La construcción de modelos se llevará a cabo dentro de la materia de TOG, pero a continuación se da una breve explicación del modelo de Red Neuronal que se desea utilizar para predicción de datos a partir de los *datasets* estudiados dentro de este proyecto.

El modelo que se espera construir, validar y utilizar para predicción de variables relacionadas con observaciones de aerosoles, es el modelo de Red Neuronal Recurrente. De acuerdo con lo investigado en [13], las Redes Neuronales Recurrentes están adaptadas para situaciones en las que los datos tienen forma de secuencia, como lo son las series de tiempo, y en donde lo que queremos detectar depende del contexto de los datos y de la evolución temporal de los mismos. En cuanto a su estructura, este tipo de redes neuronales tiene conexiones recurrentes en su capa oculta (*hidden layer*), lo que permite en cada momento que un cierto número de estados pasados sean considerados dentro de los cálculos que se llevan a cabo en dichas capas.

A continuación, se muestra la estructura de una red neuronal de este tipo de manera gráfica:

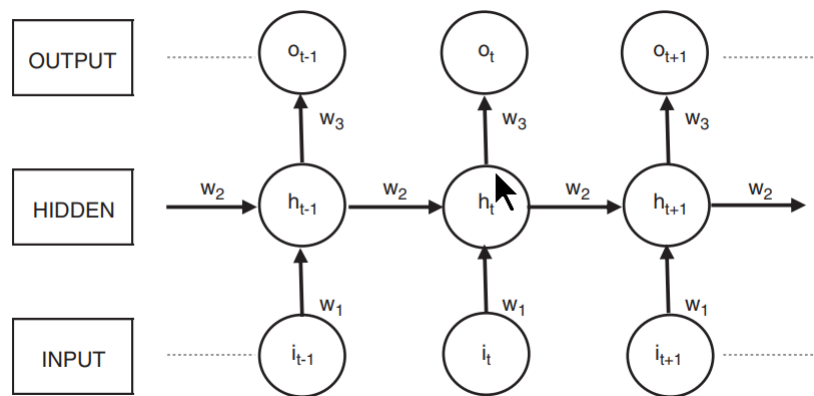


Figura 19 Model de Red Neuronal Recurrente.

5.2. Conclusiones

Para llevar a cabo la construcción de la Red Neuronal Recurrente, es necesario investigar más sobre el tema. A pesar de que la literatura indica que este tipo de Redes Neuronales Artificiales son las que deben ser utilizadas para predicción de series de tiempo, se requiere más información del tema, por lo cual dentro de la materia de IDI II y otras materias a venir dentro de la Maestría en Sistemas Computacionales el plan es concluir con dicha investigación y validar algunos modelos para verificar la hipótesis

6. PRESENTACIÓN DE RESULTADOS

Resumen: *En esta sección se presenta la visualización de los datos de la Concentración de Masa de Aerosoles obtenidos, procesados, post-procesados y promediados siguiendo la línea de trabajo definida en las etapas anteriores de este proyecto con el fin de seguir comprendiendo el alcance, validez y confiabilidad de los datos mismos para su uso dentro del modelo propuesto anteriormente que se pretende construir en un futuro dentro de la materia de TOG.*

6.1. Presentación de resultados

Para este capítulo, se llevará a cabo la visualización del *dataset* número 1 con el cual se ha trabajado en los capítulos anteriores. Así mismo, se llevará a cabo un promediado de cada pixel tomando en cuenta cada uno de los *datasets* obtenidos para el período de 120 días (120 *datasets* que comprenden 4 meses de mediciones de enero a abril del 2023) para llevar a cabo una visualización y ver las diferencias entre considerar *datasets* individuales y llevar a cabo operaciones con una mayor cantidad de datos (o *datasets*) para la variable de estudio, la concentración de masa de aerosoles.

- Durante la ejecución del promediado para los 120 *datasets*, es decir, considerando 120 días de datos o mediciones, el entorno de Visual Studio Code utilizado para llevar a cabo la programación (ejecutar el *kernel* de Python) se cierra de manera inesperada. La suposición hasta el momento es que el consumo de memoria RAM supera el alojado para este IDE, por lo cual es necesario investigar y encontrar una manera de evitar que se cierre, limpiar el buffer en caso de que alguno esté haciendo *overflow*, o utilizar menos *datasets*. **Para el cierre de este reporte, la decisión que se tomó fue la de disminuir el número de *datasets* para llevar a cabo el promediado, dejándolo en 20 días de mediciones, por lo cual los resultados contemplan los días del 1ro al 20 de enero del 2023.**

1.1. Visualización del *dataset* número 1 para la variable de estudio (1ro de enero del 2023)

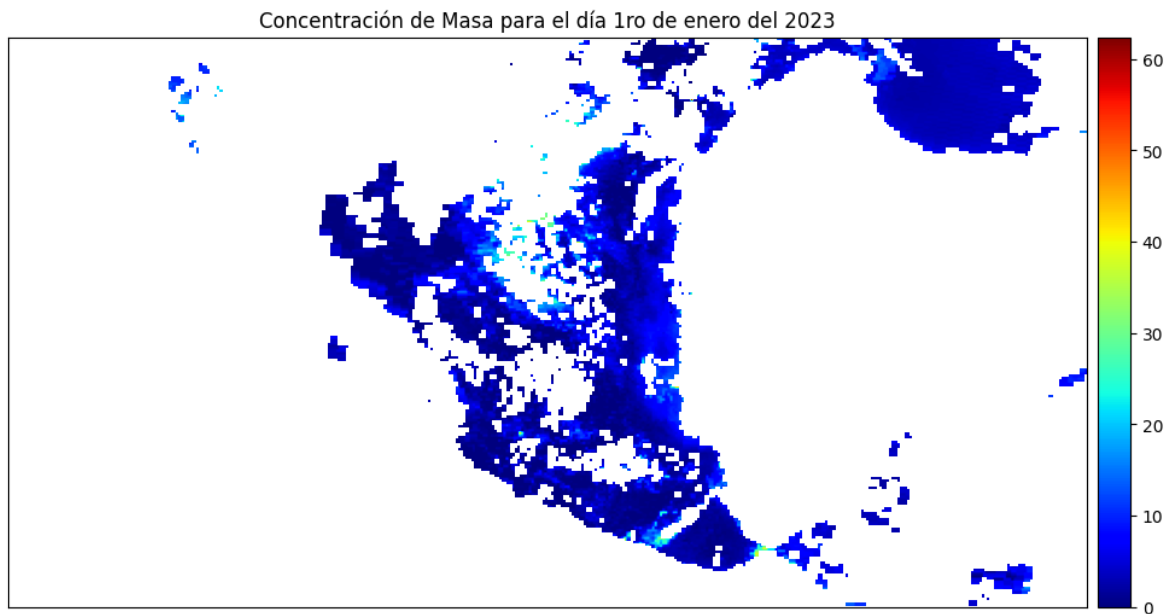


Figura 20 Concentración de masa del 1ro de enero del 2023.

- *De este primer gráfico de los datos para la concentración de masa, podemos ver que los valores NaN hacen que las mediciones correctas dejen espacios en el territorio mexicano sin información alguna, y que la mayor parte de las mediciones correctas se encuentran en valores que se encuentran entre 0 y $20 \times 10^{-6} \text{ g/cm}^2$.*

Adicionalmente, enseguida podemos ver los gráficos para los días 10 y 20 de enero, donde una situación similar ocurre (con los valores NaN), pero así mismo podemos ver que existe una variación entre aquellos valores que fueron NaN previamente y los que son NaN en los siguientes dos casos, por lo que podemos suponer que el promediado permitirá cubrir una mayor zona del mismo territorio.

1.2. Visualización del *dataset* número 10 para la variable de estudio (10 de enero del 2023)

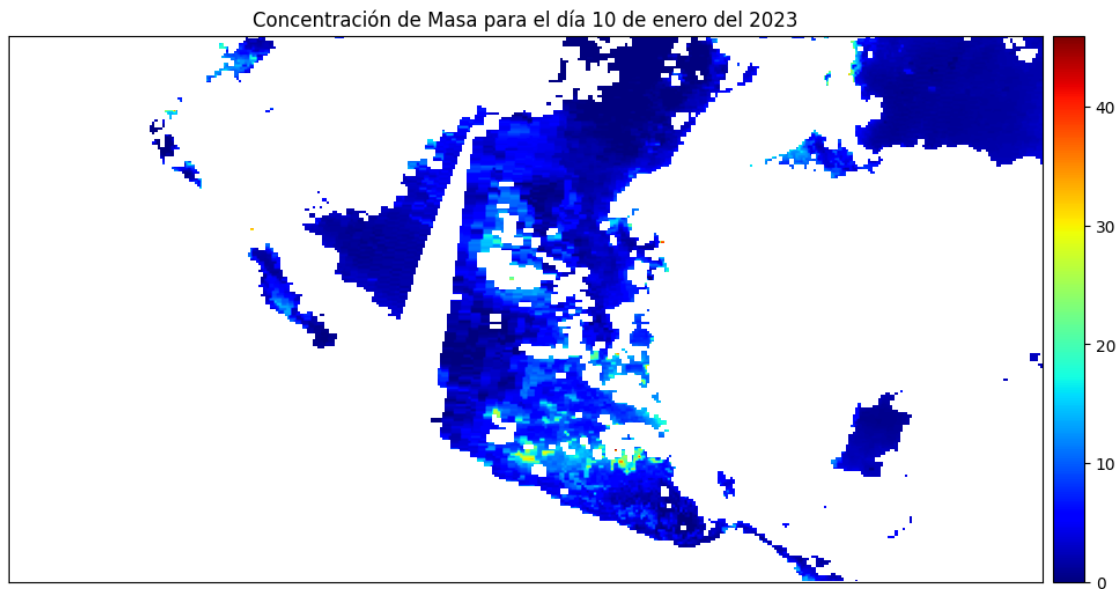


Figura 21 Concentración de masa del 10 de enero del 2023.

1.3. Visualización del *dataset* número 20 para la variable de estudio (20 de enero del 2023)

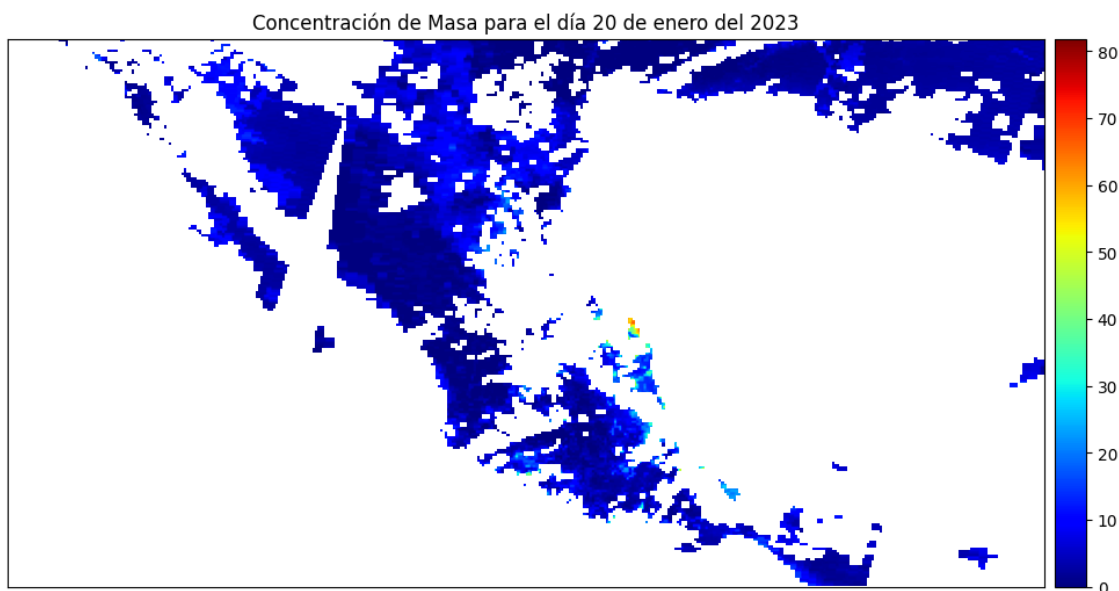


Figura 22 Concentración de masa del 20 de enero del 2023.

1.4. Visualización del promediado de la concentración de masa (del 1ro al 20 de enero del 2023)

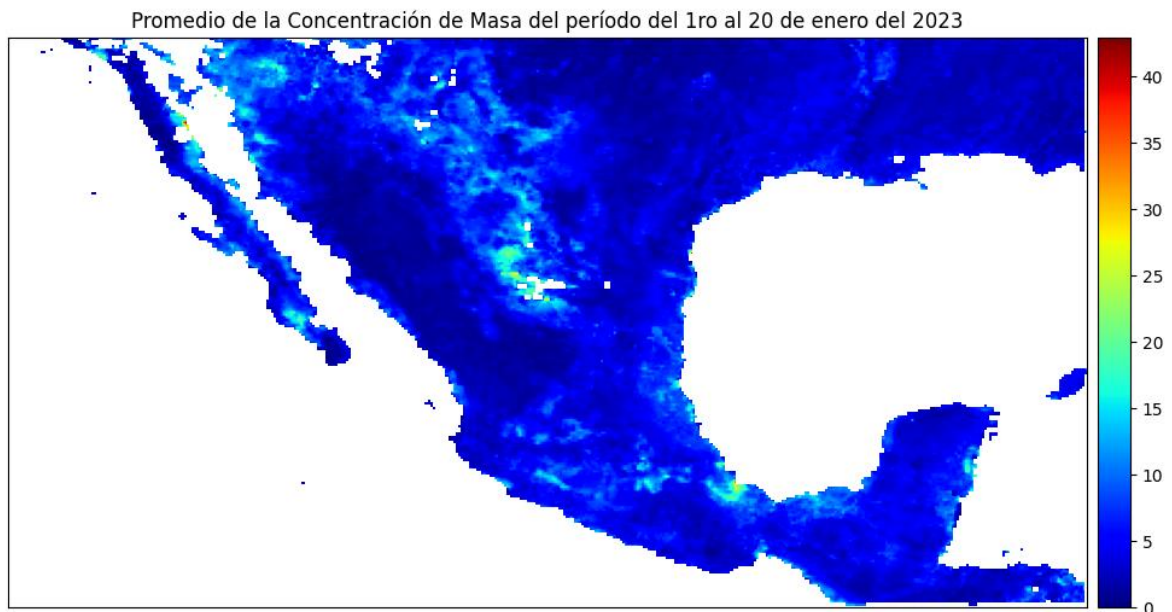


Figura 23 Promedio de concentración de masa de aerosoles para un período de 20 días.

- *Haciendo uso de una función de promediado (sin considerar valores NaN), se encuentra que el promedio de concentración de masa para un período de 20 días, comprendido entre el 1ro y el 20 de enero del 2023 (inclusivo), oscila entre 0 y $15 \times 10^{-6} \text{ g/cm}^2$.*
- *Además, la cobertura del territorio para observaciones de la concentración de masa mejora bastante, cubriendo zonas que para mediciones diarias no es posible.*

6.2. Conclusiones

De los resultados presentados en esta sección, es posible concluir que para hacer uso de las mediciones del sensor MODIS en un modelo de predicción, es probablemente más conveniente llevar a cabo antes una etapa de promediado, pixel por pixel, con el fin de obtener una mejor cobertura de la región de estudio.

Así mismo, tal procesamiento permite reducir el número de *datasets* iniciales, sin embargo, como se vio durante la ejecución de esta etapa, existen problemas que quedan por resolver, como lo es el uso de memoria dentro del IDE, por lo que conviene buscar alternativas para llevar a cabo este proceso y poder llevar a cabo un promediado para períodos mensuales, es decir, de 30 días.

7. CONCLUSIONES

Resumen: *En esta sección se presentan las conclusiones obtenidas del análisis de los datos de la Concentración de Masa de Aerosoles y el trabajo a futuro a realizar para que el modelo propuesto (en la etapa 5 de este proyecto) a construir dentro de la materia de TOG permita funcionar de manera exitosa.*

7.1. Conclusiones

Tomando como base los objetivos del proyecto, presentados en el capítulo 1, podemos concluir lo siguiente:

- 1) Se comprendió la estructura de los *datasets* y los tipos de datos contenidos en ellos.
- 2) Se entendió qué tipo de valores es posible utilizar, y cuáles requieren un tratamiento especial, como lo fue con los valores NaN.
- 3) Se logró hacer uso de las herramientas Panoply y Python tanto para el análisis como para un procesamiento de los datos para un uso futuro dentro de la materia TOG.
- 4) Un tipo de visualización, haciendo uso de una librería en Python, se logró llevar a cabo, sin embargo, aún queda trabajo por hacer, como lo es poder delimitar la región de estudio, en este caso, México.
- 5) El proceso de obtención, post-procesamiento, análisis, limpieza y visualización de los datos se comprendió, de tal manera que ahora resulta más sencillo trabajar con los *datasets* de la concentración de masa (visto en este proyecto), así como con otras variables importantes relacionadas con aerosoles, como lo son la profundidad óptica, el tipo de aerosol, etc.
- 6) Se verificó el uso correcto de la plataforma LAADS de la NASA para la obtención de los *datasets* en el formato correcto.

Además de los puntos arriba mencionados, se comprendió de manera general la estructura de algunos tipos de *datasets* geoespaciales y las herramientas que facilitan trabajar con ellos.

7.2. Trabajo a Futuro

El conocimiento obtenido con este proyecto es muy valioso, sin embargo, aún quedan algunos puntos por resolver:

- 1) Corregir el uso de la memoria RAM consumida en el IDE utilizado para poder llevar a cabo operaciones con los *datasets*, como el promediado pixel por pixel, para abarcar mayor cantidad de datos y poder reducir más el número de *datasets* finales para su uso en algún modelo de predicción, como lo sería con un modelo de Red Neuronal Recurrente.
- 2) Es necesario llevar a cabo un proceso con otras variables físicas (mediciones) relacionadas con los aerosoles, como podría ser la profundidad óptica, con el fin de obtener aún mayor

información de la región de estudio y los mismos aerosoles, con el fin de utilizarla para el diseño de la arquitectura del modelo de predicción.

- 3) Delimitar mediante objetos geométricos la región de estudio (México) y considerar realmente sólo esa superficie. Actualmente, como se puede apreciar en los resultados del capítulo 6, los mapas no tienen fronteras: es posible ver que la región de estudio es correcta, pero esta incluye porciones con datos que pertenecen a Estados Unidos de América. Por otro lado, en realidad el mapa es una representación de todos los píxeles con mediciones correctas, lo que a su vez le da la forma de la región de estudio, cuando, por el contrario, debería definirse primero la región, seleccionar los datos contenidos en ellas, y después, sobre un mapa de la región de interés, graficar los valores de los píxeles contenidos dentro de ella.
- 4) Para el TOG, queda trabajo por hacer con la comprensión del modelo que se espera diseñar y construir, así como llevar a cabo una validación haciendo uso del proceso trabajado en este proyecto final.

REFERENCIAS BIBLIOGRÁFICAS

- [1] M. Chin, P. Ginoux, S. Kinne, O. Torres, B. N. Holben, B. N. Duncan, R. V. Martin, J. A. Logan, A. Higurashi, and T. Nakajima, “Tropospheric aerosol optical thickness from the gocart model and comparisons with satellite and sun photometer measurements,” American Meteorological Society, 2002.
- [2] P. Ginoux, M. Chin, I. Tegen, J. M. Prospero, B. Holben, O. Dubovik, and S.-J. Lin, “Sources and distributions of dust aerosols simulated with the gocart model,” American Geophysical Union, 2001.
- [3] A. van Donkelaar, R. V. Martin, and R. J. Park, “Estimating ground-level pm 2.5 using aerosol optical depth determined from satellite remote sensing,” Journal of Geophysical Research, 2006.
- [4] Level-1 and Atmosphere Archive & Distribution System Distributed Active Archive Center - LAADS DAAC (nasa.gov). Disponible en: <https://ladsweb.modaps.eosdis.nasa.gov/>.
- [5] Open and Use MODIS Data in HDF4 format in Open Source Python | Earth Data Science - Earth Lab. Disponible en: <https://www.earthdatascience.org/courses/use-data-open-source-python/hierarchical-data-formats-hdf/open-MODIS-hdf4-files-python/>
- [6] Matplotlib. Disponible en: <https://matplotlib.org/>
- [7] Numpy. Disponible en: <https://numpy.org/>
- [8] Xarray. Disponible en: <https://docs.xarray.dev/en/stable/getting-started-guide/why-xarray.html>
- [9] Rioxarray. Disponible en: <https://github.com/corteva/rioxarray>
- [10] Shapely. Disponible en: <https://shapely.readthedocs.io/en/stable/index.html>
- [11] Geopandas. Disponible en: <https://geopandas.org/en/stable/>
- [12] Earthpy. Disponible en: <https://earthpy.readthedocs.io/en/latest/>
- [13] S. Tufféry, Deep Learning. From Big Data to Artificial Intelligence with R. John Wiley and Sons, Inc., 2023.