Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 221 (2023) 292–299

Procedia Computer Science

www.elsevier.com/locate/procedia

10th International Conference on Information Technology and Quantitative Management

System design for a near real-time Caribbean Sargassum monitoring platform in the National Laboratory for Earth Observation (LANOT), Mexico. Implementing Spatial Data Science Tools.

López, M.[a], Couturier, S.[a]*

[a] Laboratorio de Análisis Geoespacial, Instituto de Geografía, UNAM, Circuito Exterior, Cd. Universitaria, C.P.04510, México, D. F.

Abstract

A sudden change in global and regional environmental conditions has triggered the invasion of Sargassum algae in parts of the Caribbean coasts since 2014. To date, it has not been possible to revert the trend of seasonal Sargassum invasion, but some public institutions of subtropical countries are in the process of building monitoring systems based on satellite earth observation. Algorithms applied on high spatial resolution (but low revisiting frequency) data have been reported successful for Sargassum detection. GOES-16, MODIS (Aqua & Terra) and VIIRS imagery are acquired daily at the reception station of the National Laboratory for Earth Observation (LANOT), hosted in the Geography Institute, National Autonomous University of Mexico (UNAM). In this research, we implement a near real time Sargassum monitoring platform off the shore of Honduras, Belize and Mexico, based on the above-mentioned satellite imagery. The system design of this platform is first described, including the Big Data infrastructure for image acquisition and storage. Then, Sargassum potential presence is mapped from each of the three sensors, using Python-based processing tools and Sargassum detection algorithms. The coarse spatial resolution products obtained could complement higher spatial resolution studies by providing inputs for temporal modelling of propagation and onshore accumulation of Sargassum.

Keywords: Spatial Data Science; GOES-16; MODIS-AQUA; MongoDB; Python3; Sargassum.

Tel.: + 521 55-5622-4240 ext. 45514; fax: +521 55-5616-2145
E-mail address:marco@geografia.unam.mx.

10.1016/j.procs.2023.07.040

1. Introduction

As a result of the unsustainable economic growth and climate change, entire regions of the biosphere and hydrosphere are moving away from their previous Holocene equilibrium [1]; some biological species are emerging as a regional threat to entire ecosystems. This is the case of Sargassum in the coastal ecosystems of the Caribbean region [2].

Governmental institutions in the region have acknowledged great limitations in their capacity to respond to the associated crisis [3]. International cooperation networks, such as the Global Earth Observation System of Systems (GEOSS) make some remote sensing based processed data available [4], but the voluntary basis of these networks makes the response to a novel threat largely dependent on the efforts and technical capacity of the countries affected. Automatic detection tools have been developed in Mexico and applied on Sentinel-2 and MODIS imagery, but comprehensive platforms using large satellite time series with Big Data-type handling are still lacking.

Although with coarse (more than decametric) spatial resolution, the Geostationary Operational Environmental Satellite (GOES) satellite products are characterized by a revisiting time of 5, 15 and 30 minutes, with the potential of almost real-time monitoring. The National Observatory for Earth Observation (LANOT) of the United States of Mexico, launched as a multi-institutional cooperation effort in March 2017, has permitted the installation of a renewed satellite reception station in the Geography Institute at the National Autonomous University of Mexico (UNAM). The reception station of LANOT ensures the continuity of freely available, historical satellite datasets acquired by the ERISA reception station since the 1990s. Imagery from a range of satellite constellations has been acquired on a daily basis since 2017, including GOES 16, the Moderate Resolution Imaging Spectroradiometer (MODIS Aqua and Terra) and the Visible Infrared Imaging Radiometer Suite (VIIRS). Minghelli et al. [5] reported a potential synergy for Sargassum monitoring between very high frequency GOES data and MODIS data.

In this research, we build up a prototype geoinformatics system that potentially can provide high temporal frequency information for the modeling and monitoring of the Sargassum in parts of the Caribbean. Geoinformatics tools were selected among open source and free software solutions from data science, including the disciplines of applied informatics, big data management and geo-computation. The focus of this paper is on the design and operational implementation of the geoinformatics platform for the storage, processing, and distribution of massive raster data in the National Laboratory for Earth Observation (LANOT) in Mexico.

In the following section, we review the overall set of techniques from data science that are implemented in the infrastructure of the LANOT laboratory for our purpose; the characteristics of the acquired satellite imagery are then specified; we finally describe the geoinformatics tools and the processing flow selected in our system. In section 3, the algorithms for the detection of Sargassum potential presence are shown. In section 4, sample images of Sargassum presence detection by the three above mentioned satellite imagery types are shown as an output of the geoinformatics system.

2. Data and methods

In the past 30 years, the analysis and processing of satellite data within the Geographic Information Sciences have integrated tools such as ENVI, SEADAS, GDAL, ENVI, QGIS and ArcGIS. More recently, the design of systems that swiftly store spatial data and allow their analysis, has led to the implementation of complex spatial data infrastructures, with data science tools [6,7]. With this new paradigm, ad hoc solutions to specific needs have emerged, using a range of options from open source and free software more suitable for low budget academic and public institutions.

In the last decade, several features of the computational architecture and software have been restructured in the Laboratory for Geospatial Analysis (LAGE) in the Geography Institute, National Autonomous University of Mexico to meet the new paradigm [8-10]. In this section, we present the set of Data Science tools, supported by the renewed infrastructure in LAGE, to come up with the design of the geoinformatics system for Sargassum monitoring. These

Data Science tools include the Big Data – type database MongoDB [11], NoSQL query language [12] and Python language programs using spatial data libraries [7,13] (Figure 1).

2.1 Satellite imagery

The challenge of a near real-time storage and processing system includes the high volume of daily satellite image data (raster format) from our reception station. The imagery comes from geostationary (GOES-16) and polar orbit (e.g. VIIRS, MODIS) satellite constellations (Table 1). All images used in this study were acquired in April 2023.

GOES-16 is intended for the study of land cover and oceanic conditions; the ABI sensor mounted on it spans 16 large spectral bands ranging from 0.47 µm to 13.3 µm, including two channels in the visible spectrum, four channels in the near infrared range, and 10 channels in the infrared range. The spatial resolution of ABI is 1 km for the blue band (470 µm) and the NIR band (860 µm) and 500 m for the red band (640 µm). The signal to noise ratio (SNR) is around 1200, 500 and 800 for the blue, red, and NIR bands, respectively [5].

Likewise, the VIIRS sensors collect visible and infrared images within 22 spectral bands (412 nm to 12.1 µm), but with a spatial resolution of 375 and 750m. Its temporal resolution can be daily, every 8 days or every 16 days. MODIS sensors acquire data in 36 spectral bands ranging from 0.4 µm to 14.4 µm. Observations are made at spatial resolutions depending on the spectral band (2 bands at 250 m, 5 bands at 500 m, and 29 bands at 1 km).

Table 1. Imagery acquired by the LANOT reception station.

Satellite	Imager	Periodicity	Spatial Resolution
GOES 16	Conus	1x15 Minutes	500 m
GOES 16	Mesoscale	1x5 Minutes	500 m
VIIRS	NOAA-14 (Historical)	1x24 Day	Bands I1-5: 375 m Bands M1-17: 750 m
MODIS-Aqua	NOAA-14 (Historical)	1x24 Day	500m

2.2 Data storage and processing

The tool used for the administration and deployment of the acquired imagery is Studio 3T, version 2023.3 (Figure 2). Added collections of imagery (such as the Sargassum detection products) may be monitored on the interface. The data was stored in Big Data-type documents of a MongoDB database [14], version 4.14.19. The use of the GridFS tool was essential for the storage of high periodicity satellite images storage and management [15]. The implementation of the storage and processing was done in the LANOT infrastructure, Geography Institute, UNAM, where mass storage systems and high-performance computing equipment were acquired for the distribution of a range of products (www.lanot.unam.mx).

The image processing in this application was done with scripts written in Python 3.0 (see Section 3: Algorithms) through a processing flowchart shown in Figure 3. Python libraries, with their fast and flexible data structures, have become the de facto standard for data-centric Python applications, offering a rich set of built-in functions for analyzing the details of structured data. Built on top of other core Python libraries such as NumPy and Matplotlib, pymongo takes advantage of these background libraries to manipulate data quickly and easily, allowing to take advantage of their functionality with less coding (Table 2).

Table 2. Python libraries used in the geo-informatics system for calculations, connection to the database and graphic display.

Library	Function
xarray	Work with labeled multidimensional arrays in Python.
numpy	Perform mathematical operations on multidimensional vectors and matrices.
cv2	Process images.
matplotlib	Build graphs from matrices and vectors.
pymongo	Work with the MongoDB database.
gridfs	Store large objects in MongoDB.
os	Provide the functions of the operating system.
request	Call requests in the http language.
glob	Search for files that match a specific pattern or file name.
re	Check if a given string matches a given pattern.

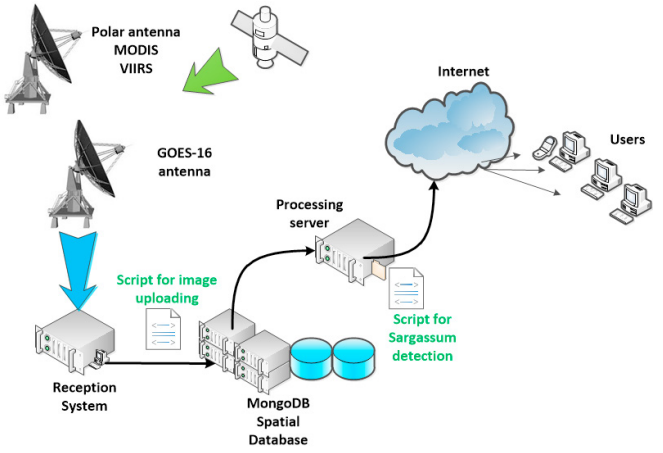

Fig. 1. System architecture

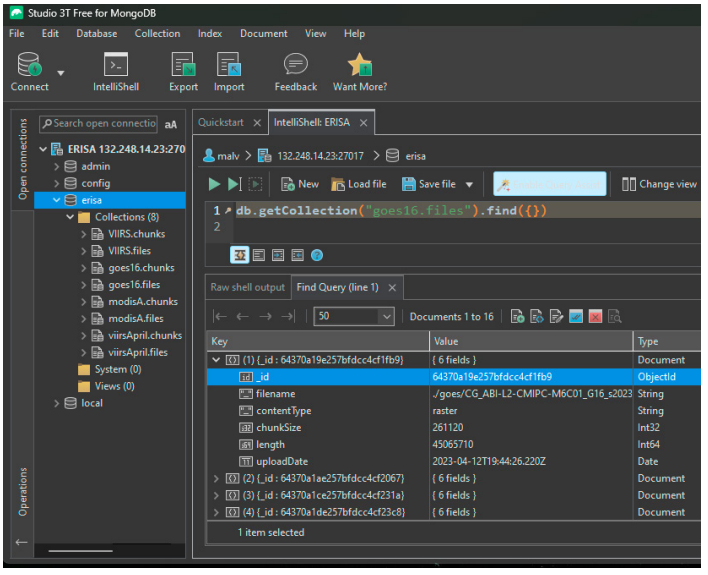

Fig. 2. Studio 3T interface for the spatial database management

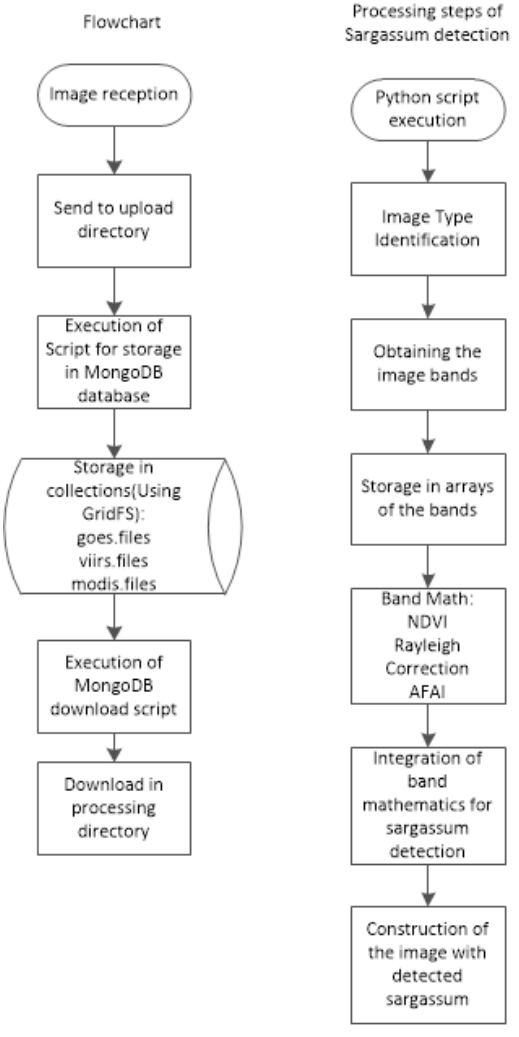

Fig. 3. Processing flowchart

3. Algorithms

Among the essential elements of this geoinformatics system are the algorithms that directly consume the data stored in Big Data MongoDB 4.14.19 documents files. For potential Sargassum presence detection, the following indices and factors were selected: the Alternative Floating Algae Index (AFAI), the Normalized Difference Vegetation Index (NDVI), and the Rayleigh correction factor [16]. A Rayleigh correction is applied to the reflectance to reduce the effect of molecular scattering.

The band algebra for the calculation of the elements for the detection of Sargassum is the following:

Rayleigh_factor = cos((sun_zenith_angle))
ndvi = (band_nir - band_red) / (band_nir + band_red)
afai = (band_red - band_blue) / (band_red + band_blue)

Figures 4-6 illustrate how pymongo and numpy can be used to combine data sets, as well as how to group, aggregate, and analyze data in them. Often, the information that needs to be processed is scattered across multiple data sets, requiring it to be organized and converted to common formats before queries can be launched before working with the required data. In practice, the data needs not always be viewed in summary format. Conversely, some analytical processing may be needed within a row group so that the number of rows in the group stays the same and the arrays that are part of the rasters are ready to implement math operations.

```
#----------------------------------------
#--Universidad-Nacional-Autonoma-de-Mexico--
#--Instituto-de-Geografia------------------
#--Marco Antonio Lopez ----------
#--Python3----------------------------------
#----------------------------------------
import pymongo
import gridfs
import os
import glob
import re
#------------------
#-----Variables-----
#------------------
COLECCION = input("Enter the name of the Collection where they are stored:")
#------------------
#--Start Script--
#------------------
from pymongo import MongoClient
client = MongoClient('xxx.xxx.xx.xxx',username='user',password='xxxxxxx',
            authSource='admin',authMechanism='SCRAM-SHA-256')
db = client.erisa
#print(db.list_collection_names())
fs = gridfs.GridFS(db, collection=COLECCION)
directory = './goes'
for filename in glob.glob('./goes/*.nc'):
        file = open(filename, 'rb')
        fid = fs.put(file,filename=filename,content_type='raster')
print('##################################################')
print(f'Upload successful in the collection:{COLECCION}')
```

Fig. 4. Script: uploadGOESMongo.py

```
#----------------------------------------
#--Universidad-Nacional-Autonoma-de-Mexico--
#--Instituto-de-Geografia------------------
#--Marco Antonio Lopez ----------
#--Python3----------------------------------
#----------------------------------------
import pymongo
import requests
import os
from pymongo import MongoClient
import gridfs
from gridfs import GridFS
client = pymongo.MongoClient('xxx.xxx.xx.xx',username='user',password='xxxx',
                    authSource='admin',authMechanism='SCRAM-SHA-256')
db = client.erisa
print(db.list_collection_names())
print('These are the collections:')
print(db.list_collection_names())
print('##################################################')
print('Select one')
COLLECTION = input("Name of Collection GOES: ")
fs = GridFS(db)
output_directory = '/home/malv/detectSARGASSUM/goesDOWNLOAD/'
os.makedirs(output_directory, exist_ok=True)
# Get all documents in the collection
# Get the collection in MongoDB
collection_name = COLLECTION + '.files'
collection = db[collection_name]
fs = gridfs.GridFS(db, collection = COLLECTION)
cursor = collection.find()
for document in cursor:
        if document['contentType'] == 'raster':
                file_id = document['_id']
                filename = document['filename']
                filename = filename[7:]
                output_path = os.path.join(filename)
                download_location = output_path
                out_data = fs.get(file_id).read()
                download_location = "/home//detectSARGASSUM/goesDOWNLOAD/"
                + filename
                output = open(download_location, 'wb')
                output.write(out_data)
                output.close
                print(f'File has been downloaded {filename}
            successfully in {download_location}.')
print(f'Download Successful in: {output_directory}')
```

Fig. 5. Script: downloadGOESMongo.py

By providing a rich set of methods, data containers, and types, xarray and numpy are packages that make Python a powerful language for data processing and analysis. Relatively complex data queries are allowed with the help of intuitive and easy-to-use features, combining and reshaping original data sets as needed. Among its advantages over other programming languages, Python is easy to read. It is largely based on common keywords of the English language. When all the modular features are used, it's quite possible to produce a block of code that can literally read like a book.

Algorithms were built for the three sensors, considering the characteristics of each one of them, but the structure of the mathematical calculations is similar.

```
#-----------------------------------------
#--Universidad-Nacional-Autonoma-de-Mexico--
#--Instituto-de-Geografia------------------
#--Marco Antonio Lopez ----------
import xarray as xr
import numpy as np
import matplotlib.pyplot as plt
import cv2
from datetime import datetime
import ephem
# Cargar archivo netCDF del sensor GOES-16
ruta_archivo_banda2 = 'CG_ABI-L2-CMIPC-M6C02_G16_s20231002026172_e20231002028549_c20231002030404.nc'
ruta_archivo_banda3 = 'CG_ABI-L2-CMIPC-M6C03_G16_s20231002026172_e20231002028547_c20231002030441.nc'
ruta_archivo_banda5 = 'CG_ABI-L2-CMIPC-M6C05_G16_s20231002026172_e20231002028551_c20231002030454.nc'
# Cargar datos de las bandas del GOES-16
banda2 = xr.open_dataset(ruta_archivo_banda2)
banda3 = xr.open_dataset(ruta_archivo_banda3)
banda5 = xr.open_dataset(ruta_archivo_banda5)
# Obtener la fecha de una de las bandas
fechaImagen = banda5['t'].values
fechaImagen = str(fechaImagen)
# Obnetener los Datafiles(matriz) de cada banda
banda2 = banda2['CMI'].values
banda3 = banda3['CMI'].values
banda5 = banda5['CMI'].values
############################### sun_zenith_angle
date_string = fechaImagen.split('.')[0]
# Convert string to datetime object using strptime()
dt = datetime.strptime(date_string, "%Y-%m-%dT%H:%M:%S")
# Extract day, month, year, hour, minute, and second from datetime object
day = dt.day
month = dt.month
year = dt.year
hour = dt.hour
minute = dt.minute
second = dt.second
# Format
formatted_date_time = f"{day:02d}-{month:02d}-{year} {hour:02d}:{minute:02d}:{second:02d}"
lat = 00.0  # Latitud en grados
lon = -75.0  # Longitud en grados
observer = ephem.Observer()
observer.lat = str(lat)
observer.lon = str(lon)
# Crear un objeto Sun para representar al sol
sun = ephem.Sun()
# Establecer la fecha y hora para la cual se desea obtener el Angulo de cenit del sol
fecha_hora = formatted_date_time
observer.date = fecha_hora
# Calcular el Ã¡ngulo de cenit del sol en grados
sun.compute(observer)
zenith_angle = ephem.degrees(sun.alt)
```

```
############################################ Raleigh
sun_zenith_angle = zenith_angle_deg
raleigh_factor2 = np.cos(np.deg2rad(sun_zenith_angle))
band_corrected2 = banda2 / raleigh_factor2
band_corrected2 = cv2.resize(band_corrected2,(5000, 3000) )
raleigh_factor3 = np.cos(np.deg2rad(sun_zenith_angle))
band_corrected3 = banda3 / raleigh_factor3
raleigh_factor5 = np.cos(np.deg2rad(sun_zenith_angle))
band_corrected5 = banda5 / raleigh_factor5
band_blue = band_corrected2
band_red  = band_corrected3
band_nir  = band_corrected5
############################################ AFAI - NDVI
# Alternative Floating Algae Index (AFAI)
afai = (band_red - band_blue) / (band_red + band_blue)
#  Normalized Difference Vegetation Index (NDVI)
np.seterr(divide='ignore', invalid='ignore')
ndvi = (band_nir - band_red) / (band_nir + band_red)
# Umbral para detecciÃ³n de sargazo
umbral_afai = 0.15  # Umbral para AFAI
umbral_ndvi = 0.2  # Umbral para NDVI
# Mask for sargassum
mascara_afai = np.where(afai > umbral_afai, 1, 0)
mascara_ndvi = np.where(ndvi > umbral_ndvi, 1, 0)
mascara_sargazo = np.logical_and(mascara_afai, mascara_ndvi)
# Plot
plt.figure(figsize=(12, 6))
plt.subplot(2, 2, 1)
plt.imshow(band_blue, cmap='gray')
plt.title('Blue band')
plt.colorbar()
plt.subplot(2, 2, 2)
plt.imshow(band_red, cmap='gray')
plt.title('Red Band')
plt.colorbar()
plt.subplot(2, 2, 3)
plt.imshow(afai, cmap='jet', vmin=-1, vmax=1)
plt.title('AFAI')
plt.colorbar()
plt.subplot(2, 2, 4)
plt.imshow(ndvi, cmap='jet', vmin=-1, vmax=1)
plt.title('NDVI')
plt.colorbar()
plt.tight_layout()
plt.figure(figsize=(8, 8))
plt.imshow(mascara_sargazo, cmap='gist_earth')
plt.title('Sargassum detection')
plt.colorbar()
plt.show()
```

Fig. 6. Script: detectSargazoGOES.py

4. Results

AFAI and NDVI maps were derived from a GOES-16 satellite scene (Figure 7). Potential Sargassum presence was then derived on the basis of AFAI and NDVI indices (Figure 6) for each sensor type. Figures 8-10 illustrate different scenes of the Mexican Caribbean coast. These results show similitudes with the usual distribution of Sargassum detected by Sentinel-2 data, although they tend to overestimate the actual presence of Sargassum. In principle, the algorithms are able to detect Sargassum on the Mexican coast with a periodicity that would allow a near real time monitoring of potential Sargassum presence, as well as a periodical verification / correction using MODIS-Aqua and VIIRS images. In the implemented algorithms, a rendering function can perform classification tasks directly from raw input data (such as images, sound, or text), using multiple layers to gradually extract higher-level information from the data. The study could be consolidated using more data from the images, also establishing a storage of the products generated for the three sensors. The detection products can be made available on the online platform for the purpose of comparison.

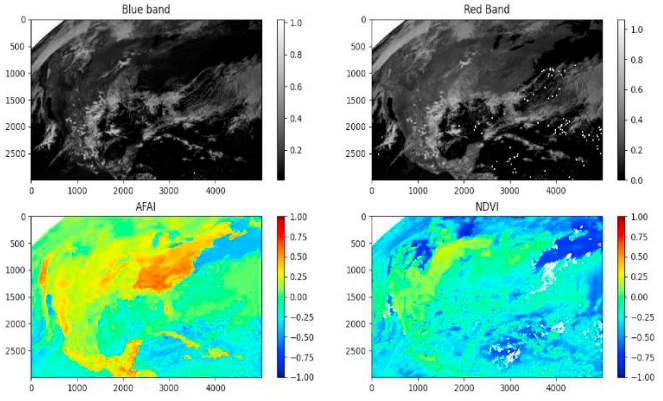

Fig. 7. AFAI and NDVI images derived from GOES

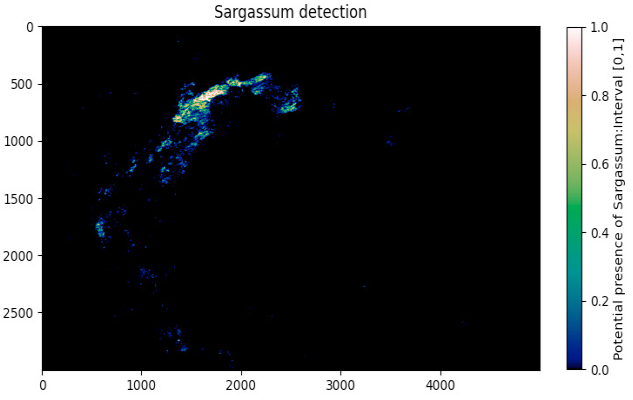

Fig. 8. Potential Sargassum presence map derived from GOES

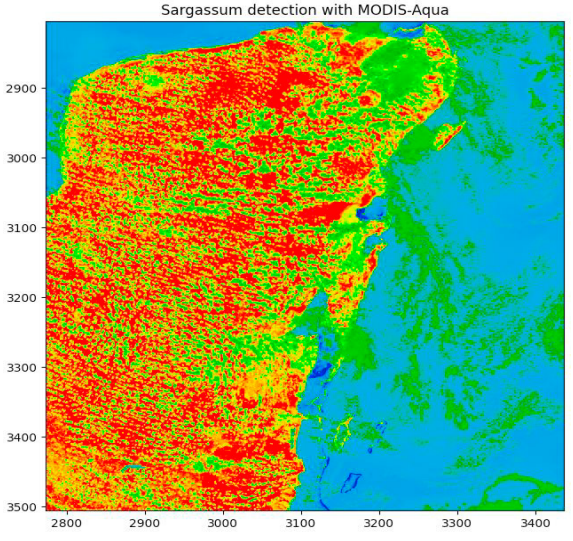

Fig. 9. Potential Sargassum presence map derived from MODIS-Aqua

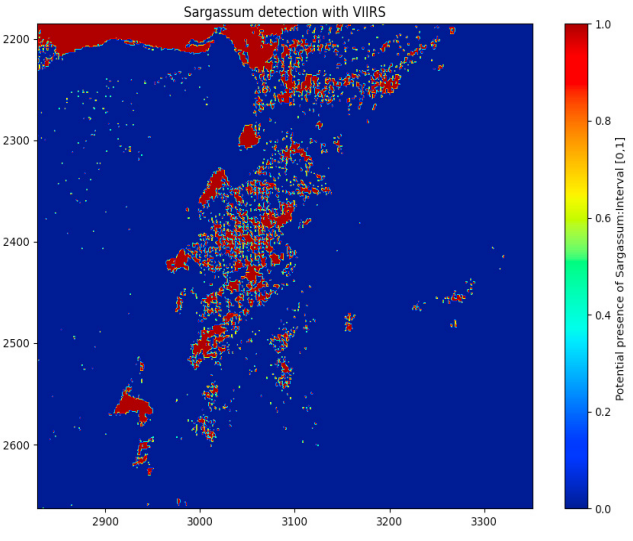

Fig. 10. Potential Sargassum presence map derived from VIIRS

5. Conclusions

A geo-informatics system was designed and implemented in the National Laboratory for Earth Observation in Mexico (LANOT, UNAM) for acquiring, storing, and managing near real time images from the GOES-16, VIIRS and MODIS satellital platforms. Preliminary Sargassum distribution maps were shown at coarse (above decametric) spatial resolution, but at high temporal frequency, in view of feeding propagation and landing models of Sargassum in the Caribbean Sea.

The proposed system provides a solution to some common problems of storing the typically large volumes of spatial data produced by receiving satellite data. The algorithms proposed for preliminary Sargassum detection and displaying results are based on software tools such as Python and GeoServer. The algorithms presented are perfectible to improve the detection of Sargassum. Various stages of testing and refinement of the database parameters are expected to be required in order to achieve satisfactory performance. A fair amount of pre-processing is needed to reduce further noise and artefacts from the satellite images before meaningful information can be extracted from them. This will involve removing unsuitable bands, masking unwanted features, correcting for atmospheric effects, and transforming data into a different image space to reduce dimensionality. Geometric correction is also a consideration, along with proper preparation of reference spectra. A major output of this study is the functional implementation platform that contributes with higher periodicity data with respect to the existing data at the moment in LANOT. This geoinformatics prototype may be used to monitor developmental stages throughout the sargassum growing season and may help provide better estimates of the impact on the Mexican coasts. An interesting perspective of this work is to establish a technological infrastructure for the implementation of Machine Learning and Deep Learning algorithms to gradually improve the Sargassum detection and modelling process. This is currently under development at LANOT. In the face of more frequent environmental crisis in the subtropical belt, the understanding of environmental damage substantially depends on the technical capability of subtropical countries for monitoring and modeling the threat. In the case of Sargassum, existing international cooperation networks (e.g. GEOSS) are able to do little unless the affected subtropical countries come up with home-made information and platforms to share with others.

Acknowledgements

We are grateful for the National Laboratory consolidation fund “LANOT-2023” (funding from the CONACYT Scientific and Technological Board in Mexico).

References

[1] Ceballos G, Ehrlich P, Barnosky A, Garcia A, Pringle R, and Palmer T. (2015). “Accelerated modern human–induced species losses: Entering the sixth mass extinction”. Science Advance. Vol 1, No.5.

[2] Olguin E, Leal R, Alzate L, Domínguez J and Tapia R. (2022). “Environmental impact of Sargassum spp. landings: an evaluation of leachate released from natural decomposition at Mexican Caribbean coast.” *Environmental Science and Pollution Research* 29: 91071-91080

[3] Fraga J, and Robledo D. (2022). “Covid-19 and Sargassum blooms: impacts and social issues in a mass tourism destination (Mexican Caribbean).” Maritime Studies 21: 159-171

[4] Osorno-Covarrubias, Couturier S, and Ricárdez M. (2015). “Global environmental sustainability: the role of geography and its recent hybridations”. *Boletín de la Asociación de Geógrafos Españoles*, 69(10): 509-513.

[5] Minghelli A, Chevalier C, Descloitres J, Berline L, Blanc P, and Chami M. (2021). “Synergy between Low Earth Orbit (LEO)—MODIS and Geostationary Earth Orbit (GEO)—GOES Sensors for Sargassum Monitoring in the Atlantic Ocean.” *Remote Sensing* 13 (8): 1444.

[6] Barton T, and Muller C. (2023). “Apply Data Science. Introduction, Applications and Projects.” USA: Springer

[7] Hassan A, Karimi and Karimi B. (2018). “Geospatial Data Science Techniques and Applications.” USA: CRC Press

[8] López Vega, M. A., Couturier, S., and Barrera González, K. Y. (2015). “Design scheme for a spatial database of climatic and environmental variables in Mexico, integrating Big Data Technology”, *Procedia Computer Science*, 55C: 503-513

[9] Couturier S, Osorno Covarrubias, Magaña Rueda, Martínez Zazueta I, and G. Vázquez Cruz (2017). “Prototype of the Mexican spatial data infrastructure for climate raster models and satellite imagery (“VISTA-C”)”. *Earth and Environmental Science*, *IOP Publishing*, 54.

[10] López Vega M, Couturier S, and J.A. López (2016). “Integration of NoSQL databases for analysing spatial information in Geographic Information Systems”. *Computational Intelligence and Communication Networks, IEEE*, 75: 351-355

[11] Jiang Z and Shekhar S. (2017). “Spatial Big Data Science. Classification Techniques for Earth Observation Imagery.” USA: Springer

[12] Moniruzzaman A, and Akhter S. (2013). “NoSQL Database: New Era of Databases for Big data Analytics Classification, Characteristics and Comparison.” *International Journal of Database*: Theory and Application 6 (4).

[13] Fischer M, and Getis A. (2010). Handbook of Applied Spatial Analysis. Software Tools, Methods and Applications: 125-135.G USA:Springer

[14] Yeung A, and Brent G. (2007). “Spatial Database Systems.” Canada: Springer

[15] López Vega, M.A, Couturier S, and D. G. Hernández Rivera (2018). “Raster data storage in Big Data infrastructures: Implementing the GridFS Tool for a Mongo DB Database” *Realidad, datos y espacio, revista internacional de estadística y geografía*, 9 (1): 72-81.

[16] Mengqiu W, and Chuanmin H. (2016). “Mapping and quantifying Sargassum distribution and coverage in the Central West Atlantic using MODIS observations.” *Remote Sensing of Environment* 183: 350-367