Guidance and Suggestions for Participants of the 2023 EY Open Science Data Challenge

Background

Welcome to the 2023 EY Open Science Data Challenge! This challenge consists of two levels – Level 1 and Level 2. The Level 1 challenge is aimed at participants who are beginners or have intermediate skills in data science. The goal of Level 1 is to predict the presence, or non-presence, of rice crops at a given location. The Level-2 challenge is aimed at participants who are more advanced in data science. The goal of Level 2 is to build a machine learning model that estimates rice crop yield for a given location. Each of these challenges will consider satellite data from the European Copernicus (Sentinel-1 and Sentinel-2) program and from NASA's Landsat program. This document is meant to provide background information, references, and dataset information. The guidance and suggestions presented here should help participants understand how the satellite data can be used to track rice crop location and growth (phenology) and build corresponding machine learning models.

The study of rice crops is not new. Researchers and government organizations routinely use satellite data to identify the location of agriculture crops and forecast yield. Unfortunately, there is no ideal solution that yields perfect results. So, we continue to make progress by taking advantage of increasing amounts of publicly available satellite data and improvements in cloud computing and machine learning. In time, we will get better at understanding agriculture from space. But, this Data Challenge gives YOU the chance to make contributions toward solving world hunger and improving food security!

Satellite Datasets

Satellite data is a unique and valuable tool to study agriculture. In this challenge, participants should consider the use of optical data from Sentinel-2 and Landsat as well as radar data from Sentinel-1. All of these datasets are readily available from the Microsoft Planetary Computer (HERE). Participants can choose one or more of these satellite datasets for their solution.

Quality Landsat data has been around since the early-1980s with a spatial resolution of 30-meters per pixel and a revisit of 16 days for one mission. We currently have two operational missions (Landsat-8 and Landsat-9) which yield 8-day revisits at any given location. The launch of the European Copernicus Sentinel-2 missions in 2015 and 2017 adds new optical data at a higher
10-meter spatial resolution and a revisit every 10 days with one mission and every 5 days with two missions. So, the combination of Landsat and Sentinel-2 missions (4 total) can observe the ground 4 times every 10 days. This coverage combination can yield coincident revisits about 9 times per year but no worse than 5 days between views of the ground. But, the issue with optical data is that it cannot penetrate clouds. So, if a cloud is over any given location, the data is not useable. In the case of Vietnam, clouds are persistent over any one location about 2/3 of the time, so only 1/3 of the data is available in a given year. In addition, our ability to identify these clouds in the data is not perfect, so we often have issues with data that is contaminated by clouds which impacts the results of models. Like I said earlier, we are getting better at using satellite data, but it is not perfect. So, if you decide to use optical data in your models, you need to consider filtering the data to remove clouds. See the section on "sample notebooks" for more information on filtering clouds from optical data.

One of the more exciting advancements in satellite data was the launch of the Sentinel-1 radar missions in 2014 and 2016. While satellite radar has been around since the early 1990’s, the Sentinel-1 system provides time-series radar data at an unprecedented temporal repeat. Since all radar data has the unique capability to penetrate clouds, it has some definite advantages over optical data in regions such as Vietnam with high cloud cover. The Sentinel-1 mission uses C-band radar at 10-meter resolution with a single mission revisit every 12 days. Unfortunately, one of the two Sentinel-1 missions failed in December 2021, so we currently have only one Sentinel-1 mission. One important aspect of the Sentinel-1 data on the Microsoft Planetary Computer is that the data includes Radiometric Terrain Correction (RTC) which accounts for terrain variations that affect both the position of a given point on the Earth's surface and the brightness of the radar return. Without RTC treatment, the hill-slope modulations of the

radiometry threaten to overwhelm weaker thematic land cover-induced backscatter differences. Additionally, comparison of backscatter from multiple satellites, modes, or tracks loses meaning. So, though you might find Sentinel-1 data from other sources (e.g., Google Earth Engine), it may lack certain corrections and have issues with terrain variations.

Analysis Region

This data challenge will use data from the An Giang province in the Mekong Delta in Vietnam. In particular, rice crop yield data was collected for the period of late-2021 to mid-2022 over the Chau Phu, Chau Thanh and Thoai Son districts (Figure 1). This is a dense rice crop region with a mixture of double and triple cropping cycles. For this challenge, all of the training and test data will assume triple cropping (3 cycles per year) with a focus on the Winter-Spring 2021-2022 season (November to April) and the Summer-Autumn 2022 season (April to August) (Figure 2).

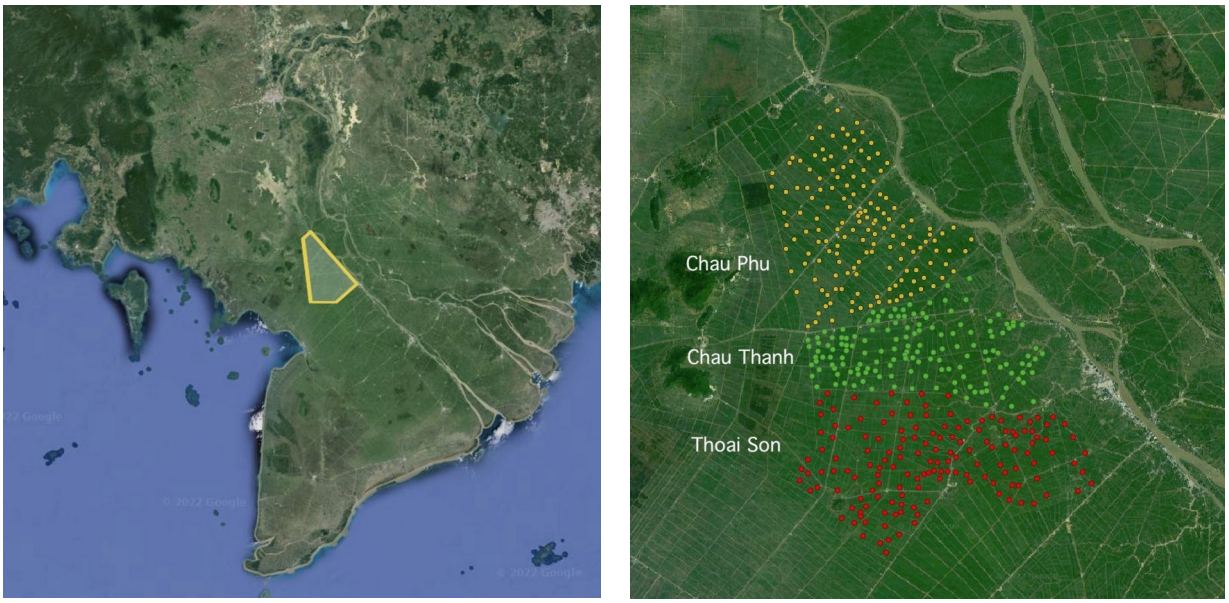

Figure 1. *Rice crop data is available for three central districts (Chau Phu, Chau Thanh and Thoai Son) in the An Giang province of Vietnam. This data is spread evenly across these districts with nearly full coverage of each district.*

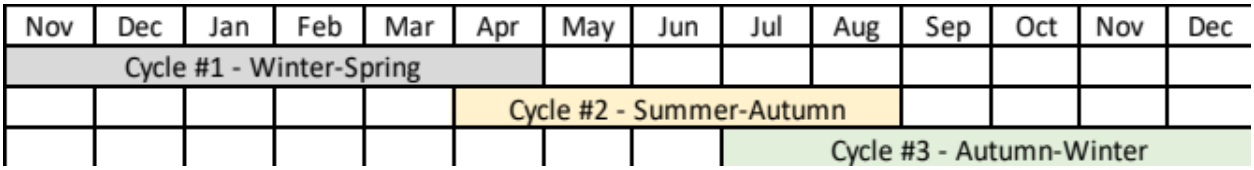

Figure 2*. Many rice crops in the An Giang province of Vietnam have 3 growth cycles per year. The data provided for this challenge is focused on the 1st and 2nd cropping cycle. These time windows should be used to evaluate the rice crop phenology (next section). Planting occurs within the first 2 months of each cycle (depending on location) and harvesting occurs within the last 2 months of each cycle (depending on location).*

Rice Crop Phenology

For this challenge, you will need to consider using satellite data to understand the growth (or phenology) of rice crops. It will be possible to use this data to identify where rice crops are grown and the productivity, or yield, of these crops.

Optical data (e.g., Landsat or Sentinel-2) contains many spectral bands (e.g., Red, Green, Blue, Red-Edge, NIR, SWIR, etc.) that can be related to the presence or growth of rice crops. As you review the references, you will find that researchers often use statistical combinations of these bands, called indices to build models. One of the more common indices for agriculture is the Normalized Difference Vegetation Index (NDVI), but there are many others including Enhanced Vegetation Index (EVI) and Soil Adjusted Vegetation Index (SAVI). Optical data is used to measure the "greenness" of vegetation during the growth cycle. Differences in band or index values can be used to distinguish differences in crop types (Figure 3) and crop yield.

NDVI (Normalized Difference Vegetation Index) = (NIR-Red) / (NIR+Red)

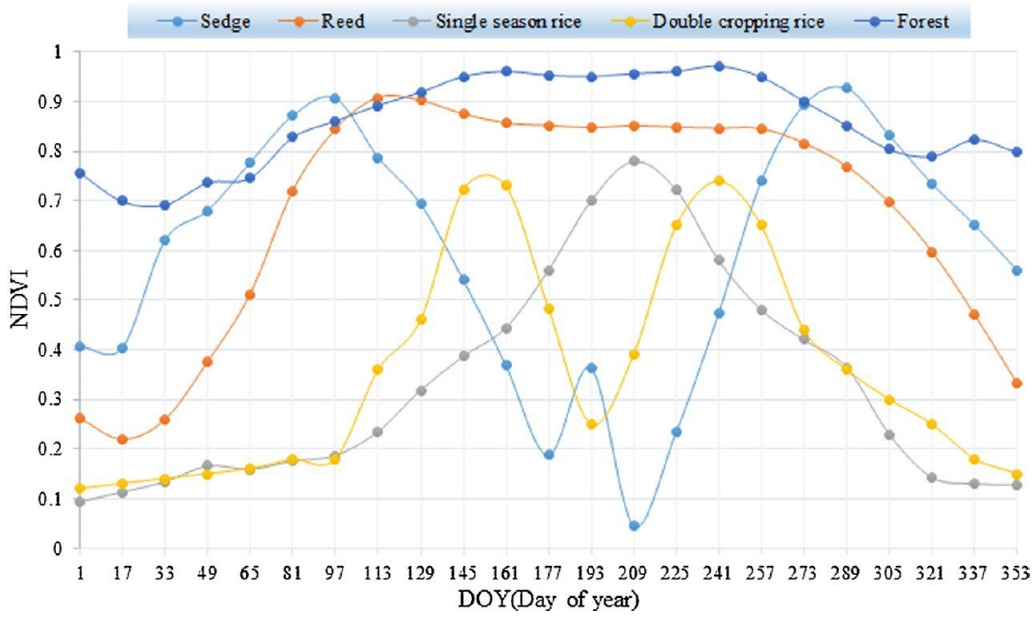

Figure 3. *Sentinel-2 NDVI varies considerably for different crop types. Forests are stable and have high NDVI values. In the case of rice, it is easy to see its variability for double cropping cycles (2 per year, yellow line) and single cropping cycle (1 per year, grey line). Credit: Reference 8.*

In the case of radar data, the Sentinel-1 missions operate in the C-band frequency, corresponding to 5.6 cm wavelength. The radar signal can be transmitted and received at either horizontal (H) or vertical (V) polarization. The Sentinel-1 data contains two basic polarization "bands", VV and VH, with the first letter indicating the transmitted polarisation and the latter the received polarisation. The VV and VH bands give us surface backscatter at any location. This backscatter tells us about the "structure" of the crop, such as its growth progress from small plants to large plants, and then to bare soil after harvesting (Figure 4). See reference [1] for more information about radar backscatter.

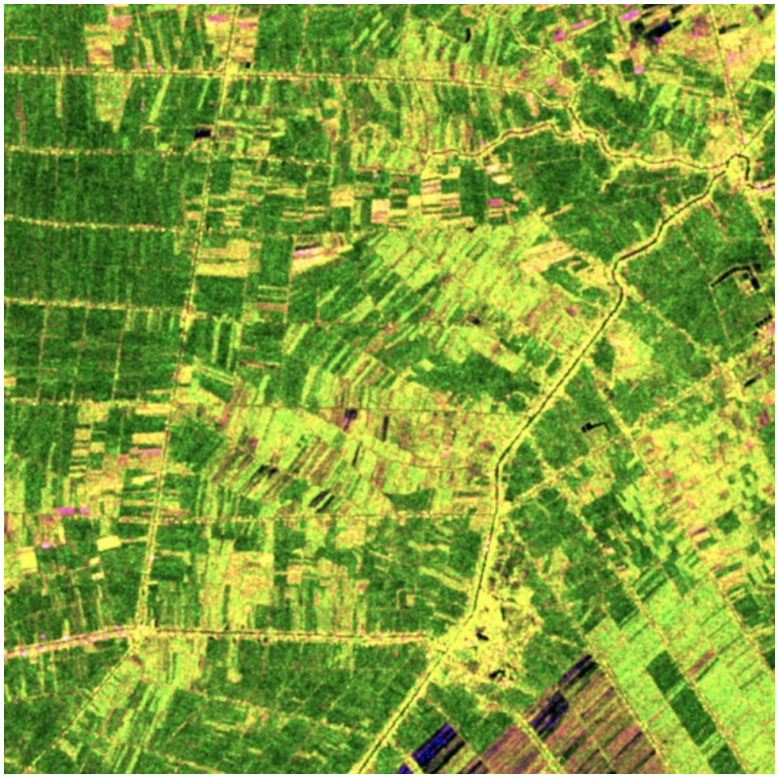

***Figure 4**. RGB plot of Sentinel-1 C-band backscatter intensity (RGB= VV, VH, VV/VH) over the Thoai Son district of the An Giang province in Vietnam. This date (23-Mar-2022) is in the middle of the spring harvest period. The variations in colors are due to backscatter differences for fields with varying growth stages. It is easy to see that these rice regions do not have identical crop cycles for every location.*

VV and VH band backscatter information can also be used to build models for crop detection (Figure 5) and crop yield. Similar to optical indices, it is also possible to develop statistical combinations of the VV and VH band data to improve modeling results. One of the more common indices is the Radar Vegetation Index (RVI) which tends to mimic the properties of optical NDVI, but not exactly. A typical equation for RVI is shown below, but there are other variations of this index used by researchers.

RVI (Radar Vegetation Index) = sqrt (1- VV / (VV+VH)) * 4 * (VH / (VV + VH))

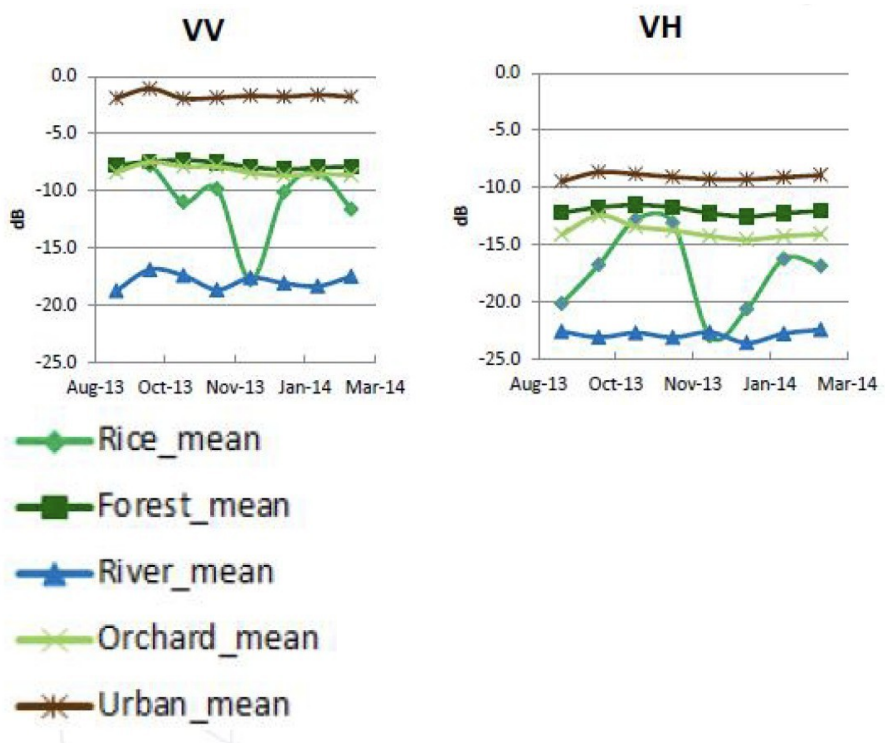

Figure 5: *C-band VV and VH backscatter varies considerably for different crop types. Urban areas have high average backscatter due to complex geometries, forests and orchards have stable and medium backscatter due to consistent annual foliage, water has low backscatter due to specular reflection, and rice has highly variable backscatter due to plant variability over the growth stage. Credit: Lam Dao Nguyen (VNSC)*

The typical growing stages of rice are shown in Figure 6. As discussed previously, it is possible to use optical and radar data to track these growing stages over time. Optical data will tell us about the "greenness" of the plant and radar data will tell us about the "structure" of the plant. For example, peak "greenness" occurs before full plant maturity as the rice grain is formed prior to harvest. In the case of "structure", which can be measured with radar data, we are able to see differences in scattering at the various growth stages. Early stages might see less scattering due to reflections from background soil or flooded fields. The peak flowering stage may see maximized scattering due to dense foliage, whereas, the ripening stage prior to harvest may see a drop in scattering due to rice tassel formation and "layover" of the plant. In the end, there will be differences between the optical and radar phenology, but it is still possible to relate this data to the growth stages and build a good yield model.

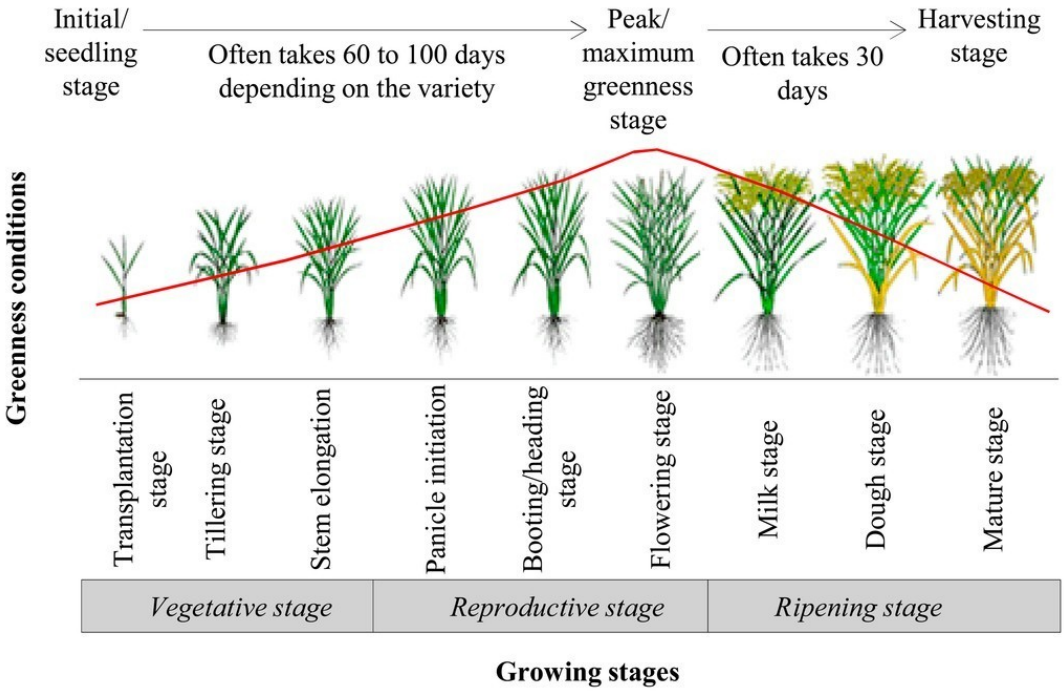

Figure 6. *Rice crop growth cycles can be tracked with optical and radar data over the growing cycle. The response from optical and radar data will differ but such information can be used to build crop detection and yield models. Credit: Mosleh et.al., Application of Remote Sensors in Mapping Rice Area and Forecasting Its Production, MDPI Sensors, 2014.*

Climatic Parameters Impact on Rice Yield

The effects of climate change include rising temperatures, altered precipitation patterns, rising sea levels, saltwater intrusion, and a greater likelihood of extreme weather phenomena like flooding and droughts. To understand rice crop growth, participants may also consider seasonal climate variations.

Climate change is expected to substantially influence the environment and socioeconomic sectors in two areas: water and agriculture. It modifies the temperature and precipitation patterns throughout time and location, directly altering the phenological cycle of plants and reducing their output. A temperature change may significantly impact how much water a crop needs to evaporate.

Rice is sensitive to the effects of both low and high temperatures. Rice cultivation requires a temperature range of 25 to 35 °C, and temperatures below or above this range are detrimental to the crop's growth, physiology, and yield. At the transplant and tillering stages, the highest temperature has a detrimental effect. Rainfall has a favorable effect during the tillering stage but a negative effect during the heading/flowering stage. The spikelet in rice

may become sterile if exposed to extremely high temperatures for a limited length of time. The primary factors contributing to the anticipated yield loss are the shorter growth period, a drop-in photosynthetic rate, and an increase in respiration.

Finding Rice Crops and Forecasting Yield

The focus of this challenge is finding rice crops (Level-1) and forecasting yield (Level-2). This is not a "perfect" process, so here are some tips to get you thinking about approaches for your model.

For the Level-1 challenge, you are given information about the location of rice crops and the locations of non-rice crops (e.g., forest, other vegetation, water). With careful consideration of time series variations in the satellite data bands or statistical combinations (indices) of those bands, it is possible to identify rice crop locations with fairly high accuracy. Figures 3 and 5 gave you examples of variations in bands (Fig 5) and indices (Fig 3) that could be used to classify rice crops versus other non-rice land classes. Notice how these figures show time-series data and not a single point in time. Try out your clever ideas and see if you can get high model accuracy.

For the Level-2 challenge, you are given yield data for specific locations and seasons. You should use those locations to find satellite data over the time window of the season. This will allow you to calculate variations in the satellite data bands or statistical combinations (indices) of those bands. As you plot these data over the crop cycle, you may find patterns or statistical results (e.g., min, max, range, slope, period, etc.) that might be optimized in a machine learning model to yield a good forecasting model. This is the goal of the Level-2 challenge, but it is not an easy process. You might develop these statistical parameters on your own or find software tools to find such parameters. We look forward to seeing your clever ideas and results!

Sample Notebooks

Participants will be given sample notebooks for Landsat, Sentinel-2 and Sentinel-1. These notebooks generate a time series plot of specific bands or indices to demonstrate how such data can be extracted for your model. As this data is prepared, a few things should be considered.

First, a decision should be made regarding the pixels selected at the specific sample location. These specific latitude-longitude locations were identified during data collection and represent a portion of the field that is consistent with the yield data. Choosing a single pixel at that location is not a good choice, as there is likely variability among adjacent pixels. Therefore, it is best to select a "window" of pixels around a given Lat-Lon location. This "window" will "average out" the optical or radar response and yield better results. Care should be taken to avoid a "window" that is too large to cover

adjacent roads or field boundaries. In addition, larger "windows" may also increase computation time.

Second, participants should filter the optical data for clouds. In the case of Landsat, this is done with the "qa_pixel" band. In the case of Sentinel-2, this is done with the "scl" band. These cloud filtering bands are not ideal, so consider approaches that reduce or eliminate outliers in the data.

Figure 7 shows examples of vegetation phenology responses for a small rice field in Vietnam over a 5-month crop cycle. The Sentinel-1 (radar data) plot uses the Radar Vegetation Index (RVI). The Sentinel-2 (optical data) plot used the common Normalized Difference Vegetation Index (NDVI).

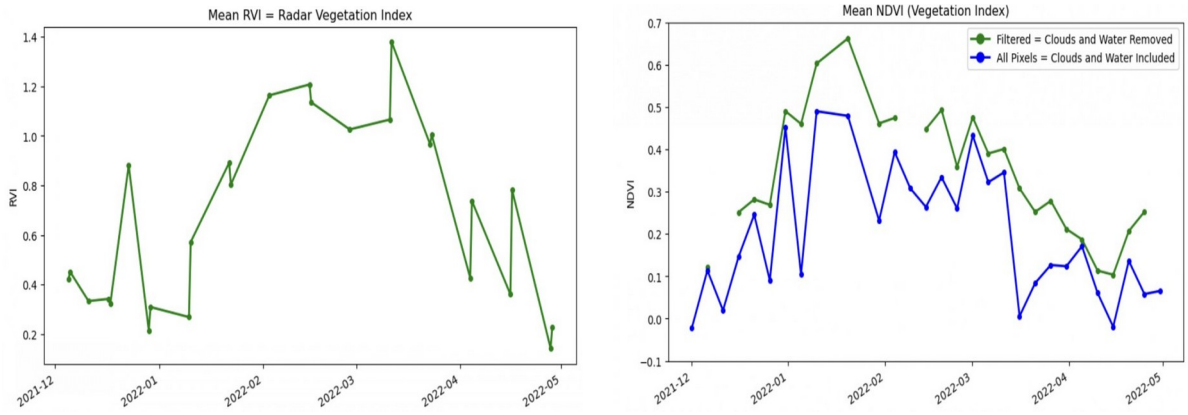

Figure 7. *Mathematical indices can be used to track rice crop growth or phenology. The Sentinel-1 radar response (on the left) tells us about the "structure" of the plant whereas the Sentinel-2 optical response (on the right) tells us about the "greenness" of the plant. The green line for the Sentinel-2 optical data (on the right) is cloud-filtered and the blue line is without cloud filtering. The time series responses from these missions can be used to build yield models.*

Conclusions

Crop identification and crop yield forecasting are growing areas of research due to the availability of free/open time-series satellite data and advancements in machine learning and cloud computing. Your models may end up being a significant contribution to the food security issues around the world. We look forward to seeing your results and wish you the best of luck.

References

[1] Rosenqvist A., 2018. *A Layman's Interpretation Guide to L-band and C-band Synthetic Aperture Radar* - LINK HERE.

[2] Digital Earth Africa - Radar Phenology Using Sentinel-1 (LINK HERE)

[3] Xu et.al., *Paddy Rice Mapping in Thailand Using Time-Series Sentinel-1 Data and Deep Learning Model, Remote Sensing*, 2021.

[4] Bazzi et.al., *Mapping Paddy Rice Using Sentinel-1 SAR Time Series in Camargue, France*, Remote Sensing, 2019.

[5] Wang et.al., *Mapping paddy rice and rice phenology with Sentinel-1 SAR time series using a unified dynamic programming framework*, Open Geosciences 2022.

[6] Jeong et.al., *Predicting rice yield at pixel scale through synthetic use of crop and deep
learning models with satellite data in South and North Korea*, Science of the Total Environment, 2021.

[7] Nazir et.al., *Estimation and Forecasting of Rice Yield Using Phenology-Based Algorithm and Linear Regression Model on Sentinel-II Satellite Data*, Agriculture, 2021.

[8] Yaotong Cai, Hui Lin, and Meng Zhang, *Mapping paddy rice by the object-based random forest method using time series Sentinel-1/Sentinel-2 data*, Advances in Space Research 64, 2019.

[9] Rosenqvist A., *Temporal and Spatial Characteristics of Irrigated Rice in JERS-1 L-band SAR Data,* Int J of Rem Sens, 1999, Vol. 20, No. 8, pp.1567-1587. doi.org/10.1080/014311699212614