



Deliverable 2: Ways data can be integrated into
xSeedScore

Deliverable 3: Visualization of data output

Deliverable 4: Comparison of EO vs current
weather/location predictions

Deliverable 5: xSeedScore EO module

Deliverables 6-7: Multi-frequency Sensor
and Novel Sensor for UAV

Deliverable 8*: Integrated Global Ag Geophysics EO
Data Store - Prototype

Partners in charge: Computomics, SilberGeo

Date: 26.02.2021

Project name	Crop Predictions Take Flight
Acronym	CROPTF
Start date	01.09.2020
End date	31.07.2021
Project Coordinator	Computomics

Document History				
Version	Issue date	Stage	Changes	Contributor
V1	XXX	XXX	XXX	XXX

Disclaimer

Any dissemination of results reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains

Copyright message

© PARSEC consortium, 2020

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgment of previously published material and of the work of others has been made through appropriate citation, quotation, or both. Reproduction is authorized provided the source is acknowledged.

Table of Contents

Introduction	4
Ways the EO data can be integrated into xSeedScore	5
National Oceanic and Atmospheric Administration (NOAA) Weather Data	5
Copernicus Data Store (CDS) – Weather Data	6
Harmonized World Soil Database (HWSD) – Global Stable Soil Properties	6
Copernicus Data Store (CDS) - Global Dynamic Soil Properties (f.e. Soil Moisture)	6
Big Data Toolbox - Satellite Imagery (Rasdaman)	7
Visualization of data output	8
National Oceanic and Atmospheric Administration (NOAA) Weather Data	8
Harmonized World Soil Database (HWSD)	9
Copernicus Data Store (CDS)	11
Comparison of EO versus current weather/location predictions	11
xSeedScore EO module (IP held by Computomics)	14
6-7. Multi-frequency EM Sensor and Novel Sensor for UAV payload (IP held by SiberGeo)	14
Integrated Global Ag Geophysics Data Store (IP held by Landviser)	19

1. Introduction

The performance of a plant for a phenotype, e.g. an agronomic trait such as yield, is influenced by two main factors: the plant's genetics, which is the heritable part that breeders aim to improve, and the environment in which the plant is growing, which is not heritable, not controllable, and highly variable both spatially and temporally. As previously shown by us and in the scientific literature for many crops, environmental effects influence as much as 41% of the phenotype expression (f.e. yield) and is *on par* with the genotype effects (Fig. 1).

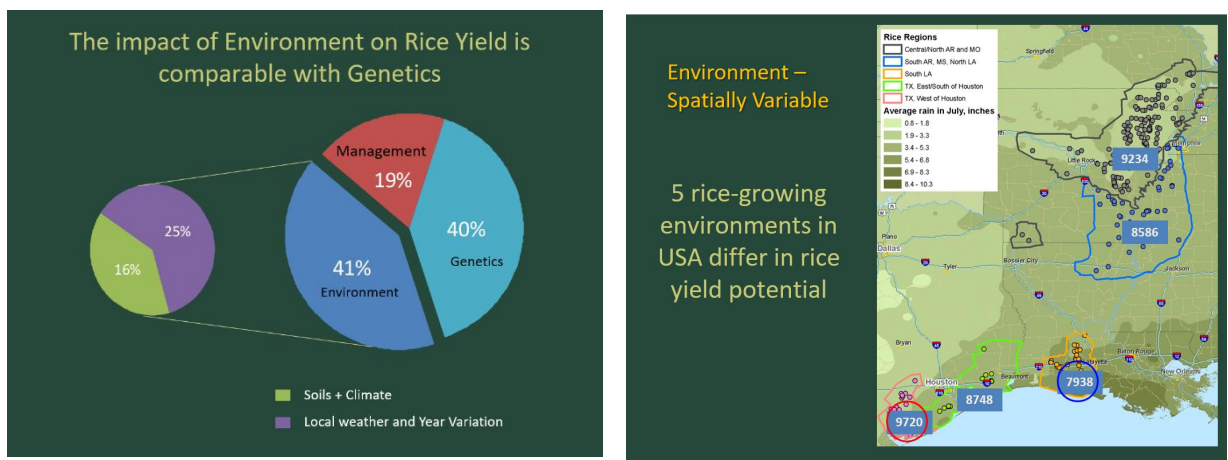


Fig. 1. Quantifying Environmental Impact on Crop Yield. Results from a statistical analysis of >20,000 year-location-variety rice yield field trial results over eleven years showing the major segregation in yield potential of the five areas in the USA: the higher precipitation in July, the lower yield. The field trial data were compiled by Landviser, LLC in 2007 from the field trial data of a private seed company and public Universities, presented at the SSSA-CSA-ASA meeting in 2008, but unpublished. A similar magnitude of Environmental impact on other crop performance can be found in multiple scientific publications.

In a common field test of new plant varieties, each variety is planted in trial plots in different test locations and often also in different years to assess the performance of the variety in different environments. This helps breeders to get a picture of the performance of a specific genetic makeup in different conditions. It also allows controlling for the effects of unforeseen environmental events like extreme weather occurring in one location that biases test results. This is a valuable approach to assess the new varieties performance, however, it is hindered by both availability of environmental (Earth Observation, EO) data, namely weather and soil characteristic, at the specific trial locations during the growing season and by the mechanism to integrate and analyze EO data together with genomic information.

A comprehensive machine learning (ML) model that aims to predict the performance of plant varieties for agronomic traits has to incorporate both sources. The model can then be asked questions about the expected performance of a plant variety in specific target environments and can help breeders select the most promising varieties.

To improve the breeding process and reduce the time-to-market of new crop varieties for Computomic's clients (breeding companies), we investigated the incorporation of environmental data like soil and weather into our existing ML solution xSeedScore additionally to the genetic data.

We used a dataset from an existing breeding program to assess the performance of phenotype prediction of models integrating genetic data and different types of environmental data. The dataset contains genetic data of plant varieties of a breeding population and phenotypic measurements for one trait obtained from planting these varieties in 12 locations over two years.

Environmental data from many available EO platforms, including weather parameters (Copernicus Climate Services, NOAA Historical Weather API), soil properties (global – Big Data Toolbox and HWSD, and regional – USA-SSURGO), satellite imagery (RASDAMAN,

ArcGIS Living Atlas) were evaluated. Technologies to manipulate data and automate data pipelines were developed and the improvement to the phenotype prediction of xSeedSCORE ML solution with incorporated EO data was quantified.

2. Ways the EO data can be integrated into xSeedScore

Objective:

As outlined above, a comprehensive model for predicting plant phenotypes needs to incorporate data that describes plant genetics and the growing environments of the plants. The latter covers the field locations and growing season. Because a growing season spans several months, environmental variables have to be aggregated in some way to meaningfully describe the planting season without having too much detail, e.g. hourly measurements of some variable over several months would be too fine-grained and lead to overfitting of the model.

However, gridded historical weather re-analysis data from Copernicus Climate Change Services are available only either as hourly raw data or monthly aggregates, which are either too fine or too coarse of the **temporal** resolution to be useful to characterize the growing season for most crops. Therefore, we evaluated another source of daily weather data from NOAA – globally available API for the last 40 years and designed our own processes of data aggregation/binning and calculating other valuable parameters, e.g.. daylight length in hours per day at any given location.

The number of input features, i.e. variables that describe a sample, has to be paired to the number of available training samples. An ML model with significantly more features than training samples is prone to overfitting, finding spurious patterns in the data that only occur in this specific training set, but do not reflect a general property.

This also means that we cannot use raw satellite images or similar data, we need the processed, aggregated descriptors derived from them. First, the spatial resolution of satellite imagery could be too coarse (10 m is the lowest available from Sentinel 2, other platforms are much coarser, 30-60 m); second, satellite and even UAV imagery only reflect information about the crop itself, not about properties of the underlying soil, which influences crop performance, and constitutes a pure EO data we are trying to obtain. Therefore, although we evaluated satellite imagery data available through Big Data Toolbox and RASDAMAN, we used Harmonized World Soil Database from FAO and SSURGO Soil Database (USA) which already provides derived soil properties impacting crop growth, although not necessary at the fine enough **spatial** resolution.

In addition, since the limitation of the existing soil data (including suggested through PARSEC In-Situ Data Hub) was expected, SiberGeo and Landviser have tested a new mobile universal soil EC sensor (EM) as payload on the octocopter UAV, results would be presented on upcoming Symposium on Applied Geophysics to Engineering and Environmental Problems (details in Deliverables #6 and #7 at the end of this report).

We have investigated the integration of the following environmental data sources. For additional details, please consult our deliverable report 1 from December 2020.

National Oceanic and Atmospheric Administration (NOAA) Weather Data

We accessed daily minimum, maximum, and average temperatures as well as daily precipitation volumes for the target locations. We have developed custom Python functions that use NOAA's API for this purpose. Using the daily values over the complete planting seasons produces too many datapoints. We investigated two ways to aggregate the values into more usable and informative features.

The first approach separates the planting season into bins and computes average values for the temperature and precipitation values. This has the advantage that this is simple to

implement and that the values are easily interpretable. The time bins can be synchronized with biologically meaningful events, i.e. crop stages, such as planting, emergence, vegetative, reproductive, ripening. A disadvantage can be that extreme events like unusual temperature highs and lows can be averaged out, and information is lost. We tried to mitigate this by using min, max, and average values.

The second approach uses statistical methods such as principal component analysis to reduce the dimensionality of the value vectors while retaining a maximum of the informational content. Here the advantage is that information content is maximized for a given target dimensionality. A disadvantage is that the values are hard to interpret.

We did not see any significant differences between these approaches in preliminary tests, so we opted for the first approach based on the advantages outlined above.

Copernicus Data Store (CDS) – Weather Data

We (out subcontractor Landviser s.r.o.) have taken a course on Copernicus Climate Change Service to evaluate in more detail the weather source available through that platform, both better spatial (0.1 x 0.1 decimal degree grid) and temporal (hourly, since 1950) is available. The CDS ToolBox can be used to flexibly extract and plot just the data needed but not all data sets are available through ToolBox. We used it to develop queries to download the weather parameters for specific location/hour/season and built a UI prototype (deliverable #7).

The limitations are that not all useful data sets, like soil moisture, are available through CDS Toolbox (our adopted methodology to extract those data is described below).

Harmonized World Soil Database (HWSD) – Global Stable Soil Properties

The HWSD is composed of a raster image file and a linked attribute database. The raster covers 221 million grid cells that cover the globe's land territory. Each cell is linked in the database to commonly used soil parameters like:

- organic carbon
- pH
- water
- storage
- capacity
- soil depth
- cation exchange capacity of soil and clay fraction
- Total exchangeable nutrients
- Lime and gypsum contents
- Sodium exchange percentage
- Salinity
- Texture class
- granulometry

For the xSeedScore model, these variables can be accessed by identifying the grid cell a target location belongs to and extracting the variables listed above from the attribute database. The number of values is not high, since they are static and do not change over the course of a planting season, so no further aggregation is necessary. They can be directly included in the ML model.

Copernicus Data Store (CDS) - Global Dynamic Soil Properties (f.e. Soil Moisture)

The Copernicus Climate Change Service Data Store, unlike NOAA historical weather data, also provides historical re-analysis of the daily soil moisture in interpolated grids. We

developed Python functions using the CDS UI to download the values for a given time span and to extract the values for target locations. There also other useful EO data sets available on CDS, but they are not available through UI/API, only as global grid zipped **NetCDF** files one per date, extracting just one number for a specific lat/long for annual time series from each **date.nc** file is very tedious, we have built a standalone Python script to extract such data and incorporated those data to ML model on test data set.

Then, we used a binning approach with average values, similar to those outlined for NOAA weather above.

Big Data Toolbox - Satellite Imagery (Rasdaman)

While Rasdaman and the associated data sources contain EO measures that potentially can be useful for our purposes, there are no suitable data points for the test locations and time spans required for our dataset.

3. Visualization of data output

We will provide customers with a visual overview of the environmental data used in the models for their breeding experiments. The figures in this section show examples of possible visualizations.

National Oceanic and Atmospheric Administration (NOAA) Weather Data

NOAA POWER initiative provides a convenient REST API to query and download a single CSV table of historical daily weather parameters such as min/max/mean temp, precipitation, solar radiation, cloud cover, and wind speed per trial Latitude/Longitude for any location on Earth. The dates span from January 1st, 1981 up to query date minus two days. This format is very useful for linking to the crop field trials for the particular location / growing season as such a query typically occurs for re-analysis after trial harvest, sometimes even a few years after the trial was conducted, and is not concerned about immediate weather data availability/forecast.

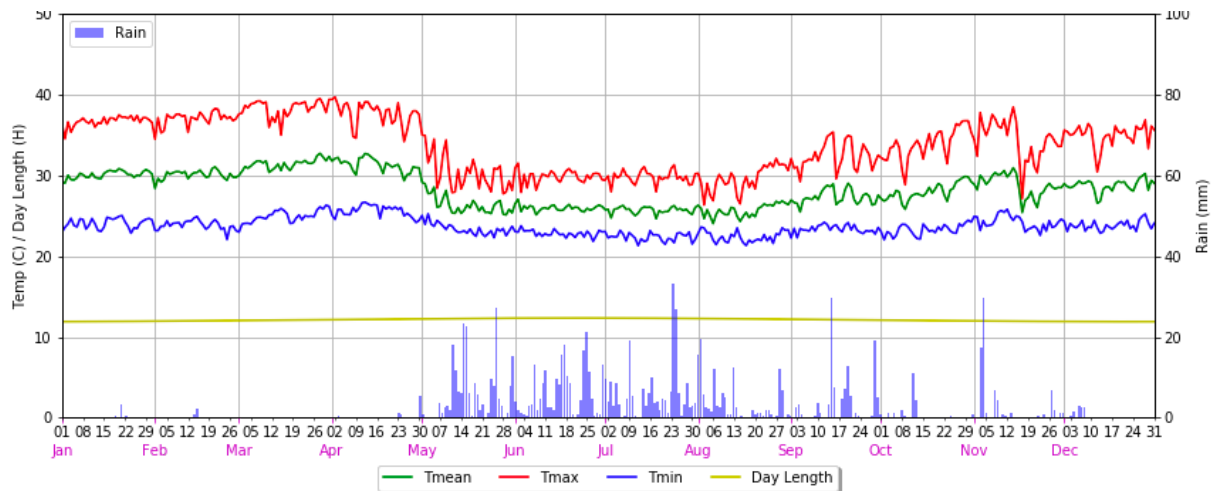


Figure 2. Minimum, maximum, and average temperature and precipitation volumes for an example location and year; calculated Day Light Length for the location latitude.

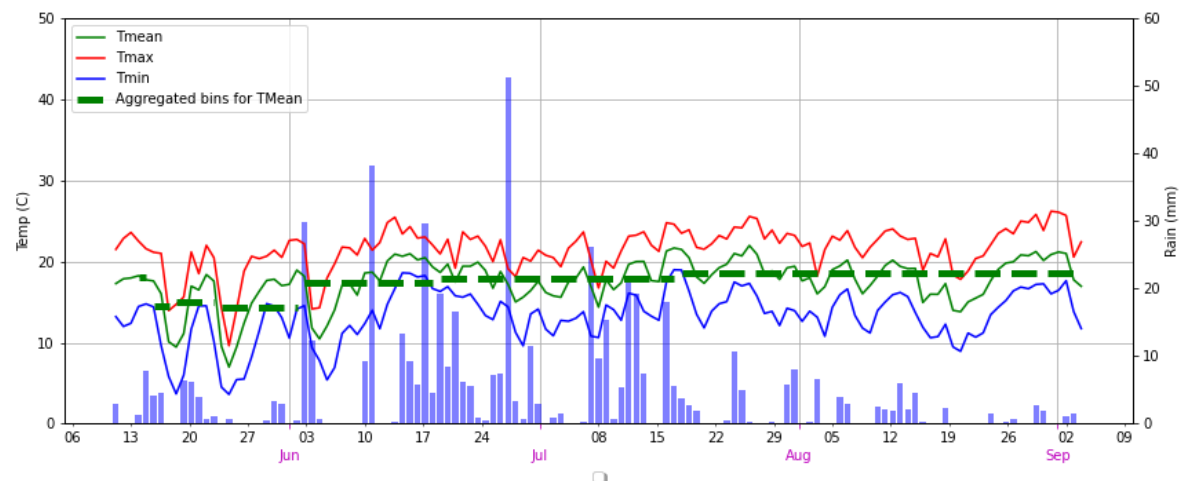


Figure 3. Daily temperatures binned to be included in the ML model, binned data prevent over-fitting of ML model.

Such data were also sliced per crop stage, temperatures can be used for calculating accumulated Growing Degrees, and extremes can be shown on the graph. Such aggregations

and queries are useful for clients (breeding companies) when they want to dive into more details for typical weather pattern at a particular location.

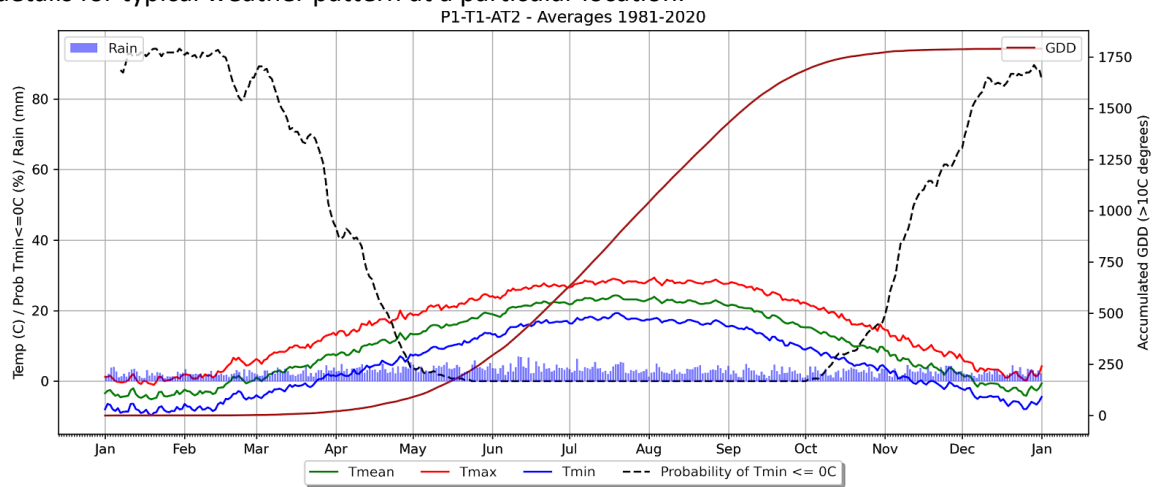


Figure 4. Data showing accumulated GDD at a test location.

The deeper dive to EO data of interest for breeding companies would also include extremes, to evaluate if the crop performance at the trial would be typical at the location or if the variety performance captured at the location can be extrapolated to another location showing similar weather patterns.

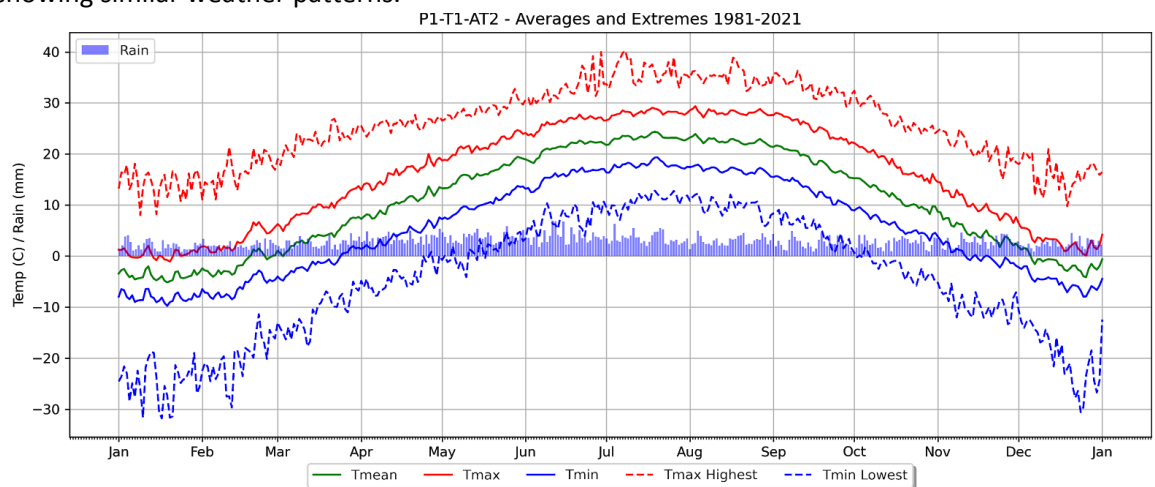


Figure 5. Extremes at the location, per 40 years of available data (NOAA)

Harmonized World Soil Database (HWSD)

Below are some visualizations showing the distribution of soil types from the HWSD using free Map Viewer downloadable with the data set and zoom in depicted with ArcGIS professional software with the field trials data set used for this research.

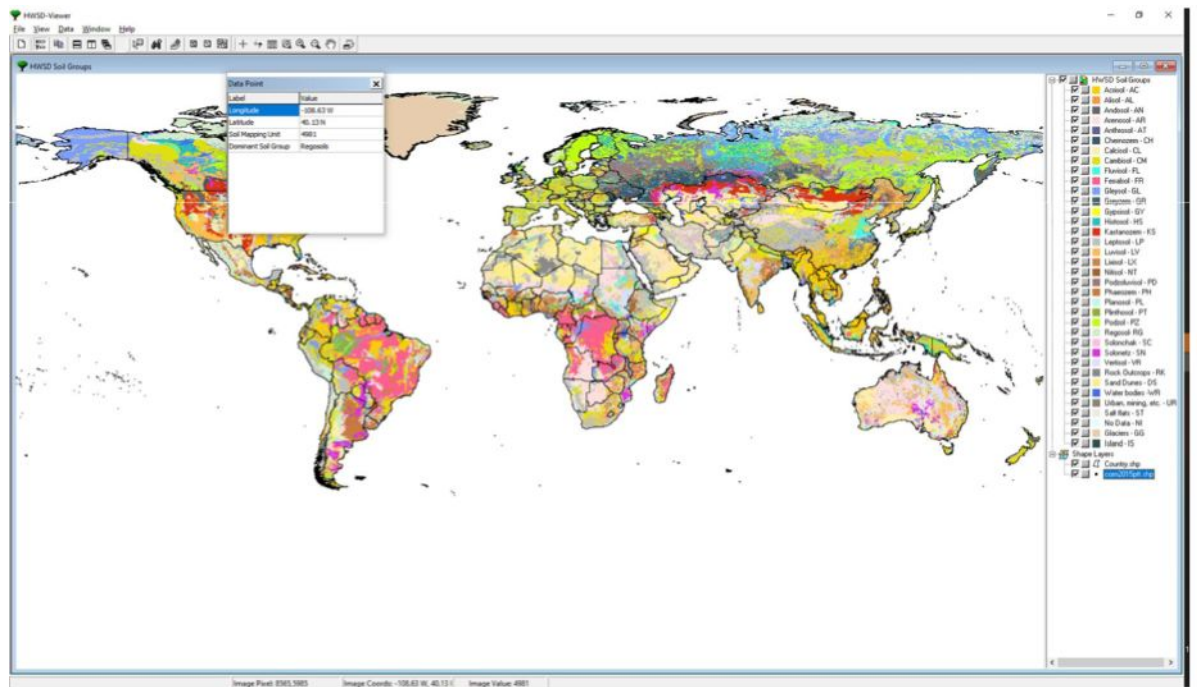


Figure 6: Distribution of the soil types in the HWSD over the global surface.

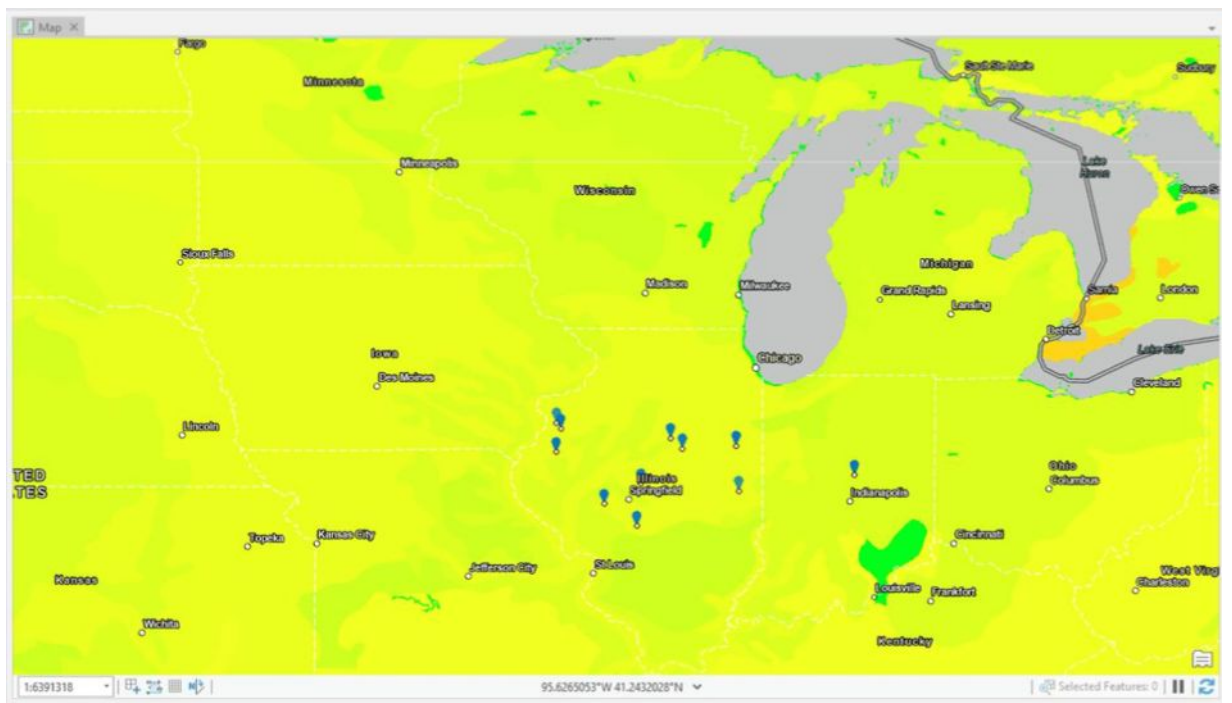


Figure 7: Pins show the locations of fields used in a field trial, colors show the corresponding soil profiles from HWSD.

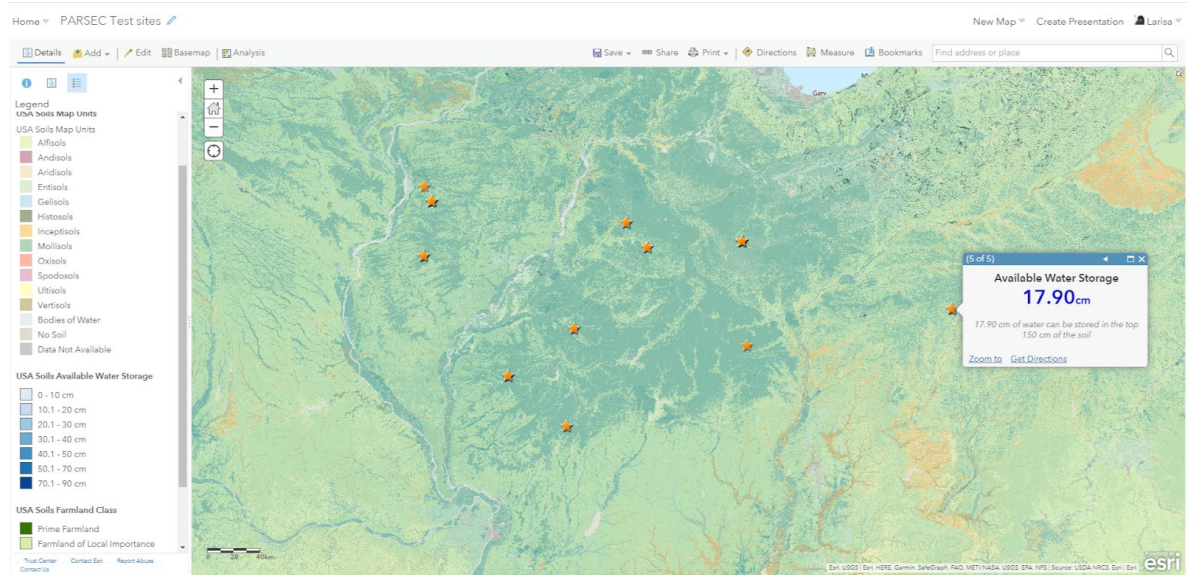


Figure 8: Pins showing the locations of field trials from more detailed SSURGO soil data set

Copernicus Data Store (CDS)

The soil moisture daily values for the test growing season were accessed through CDS ToolBox UI and downloaded as zipped NetCDF data files - global coverage, one **date.nc** file per date. The custom Python script was developed that allowed extracting values corresponding to the selected locations and the time period from the planting to harvest.

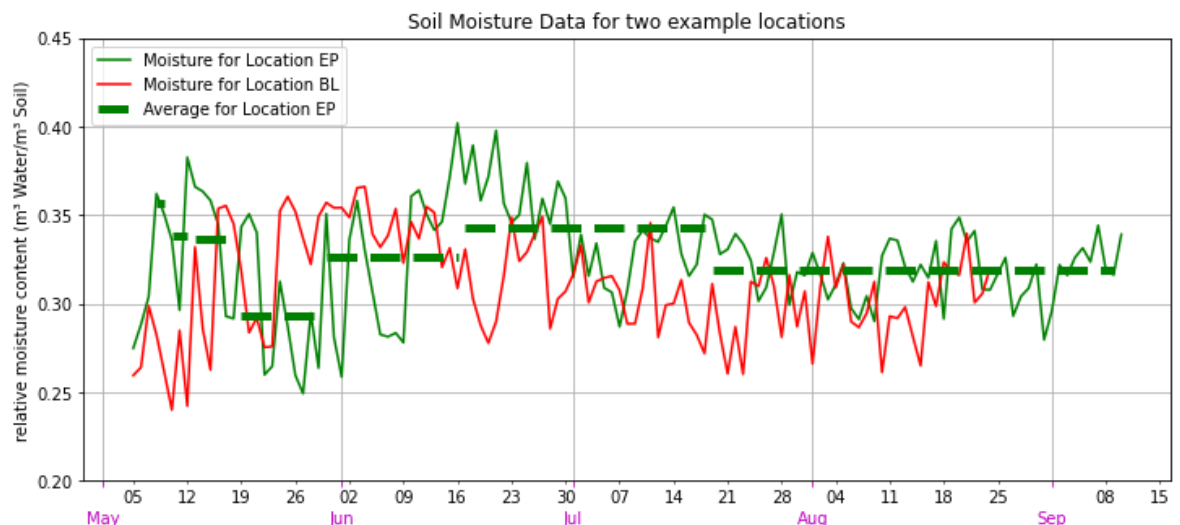


Figure 9: Daily soil moisture binned and visualized similarly to NOAA weather data shown in Fig. 3.

4. Comparison of EO versus current weather/location predictions

We conducted several experiments to investigate the incorporation of EO data into our ML models. All models used the same genetic profiles of the plant populations as input features, and additionally included different subsets of the environmental descriptors outlined in the previous sections.

We used a cross-validation approach to evaluate the predictive performance of the models. Briefly, the data set is split into subsets of the same size, e.g. into five subsets. In an iterative fashion, one subset is set aside as the test set, while the other data is pooled and used to train an ML model. This model is then used to predict values on the left-out test set. The predictions can then be compared to the true values to assess the model's performance. This is repeated so that each subset is left out once.

We also left out all data for one of the locations to see if using EO data allows us to improve prediction performance for locations that the model has not seen in its training data. Predicting the performance of plants in environments where they have not been tested can help breeders in designing field trials. It can also help answer questions about which of their varieties will potentially perform well in new environments, for example, different geographic regions, or different weather patterns, which will be important in a changing climate.

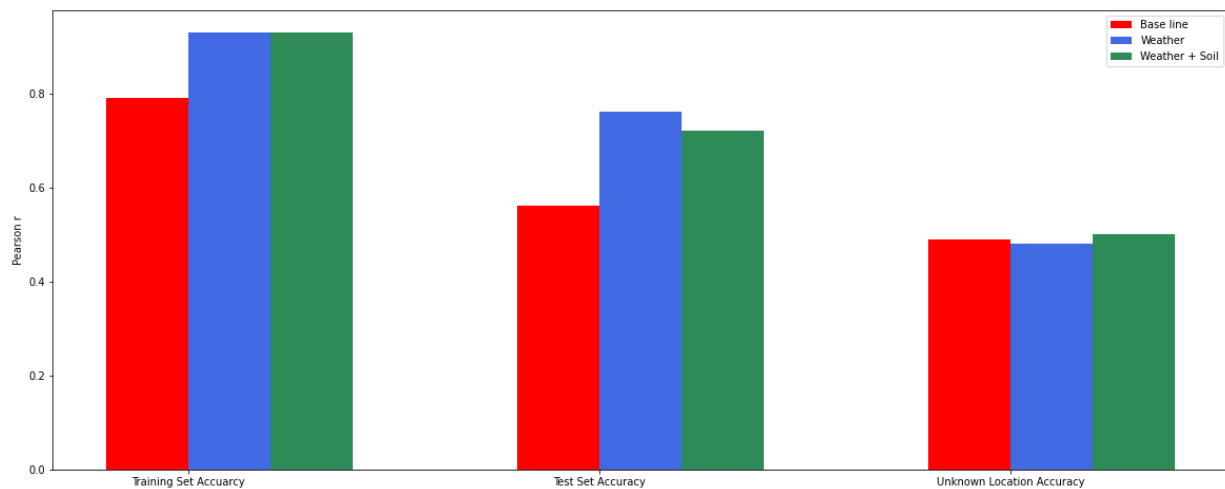


Figure 10: Prediction performance of models using different environmental variables. The y-axis shows Pearson's correlation between measured phenotype values and predicted values. These are average values over the cross-validation folds. Three different models are compared. All of them use genetic descriptors as input features. The red "baseline" model does not use any EO features to describe the growth environments of the plants, only categorical environment IDs were used. The blue "Weather" model includes weather descriptors from NOAA. The green "Weather + Soil" model includes the NOAA weather descriptors and HWSO soil descriptors. Performance is shown for predictions on three data sets: the training set, i.e. data that the model has seen before; the test set, i.e. on plant varieties that the model has not seen before, but the environments they were grown in were included in the training set; and finally, data where the model has not seen the plant varieties nor the environment.

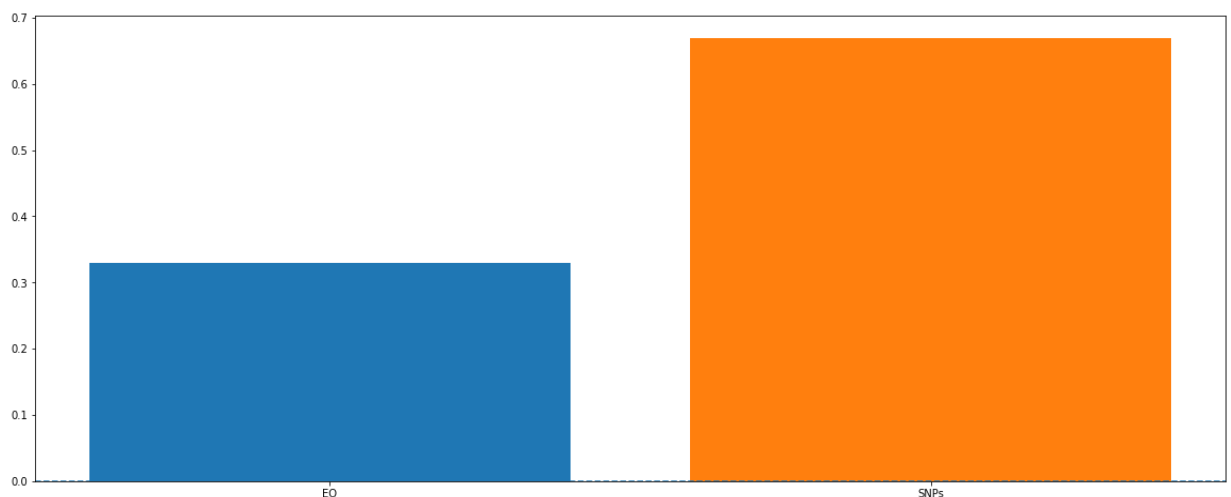


Figure 11: Importance of different data types for predictions. The models we tested have two different types of input features: genetic and EO, which correspond to heritable and non-heritable factors that influence a plant's phenotype. The genetic features are derived from SNP markers. This figure shows the relative importance of each of the two data types in the "Weather" model (see figure 10). Values for the other models with EO data are similar, and therefore not shown. EO variables are responsible for about one-third of a prediction's results, the other two-thirds come from the genetic component.

Figures 10 and 11 show results from our experiments. We can see that using EO data significantly improves prediction performance in both training and test sets, i.e. where the environment is known to the model. This improvement comes mostly from including weather descriptors. Predictions for environments that were not seen in training were not

improved by incorporating EO data. Figure 10 shows that the EO data have a significant influence on the predictions of the model, showing their utility and importance.

Our experiments show that using EO data in our predictive models improves their performance for environments where we have training data. This is a valuable result, as it shows that we can help breeders more accurately determine how a plant variety that they have not yet tested in the field would have performed in any of their environments from field trials. This helps them in deciding which varieties to select for the next steps in their breeding program.

The results show that weather data has a significant impact on prediction performance, while soil data from HWSD does not appear to influence the performance. The twelve locations in our data set have only three different HWSD soil profiles, this is probably too little for the model to discriminate between them and identify soil properties that influence phenotype values. It is well possible that for a dataset with more diverse locations the soil descriptors can improve the model performance.

We had hoped that using EO data allows us to improve predictions in locations not seen in the training data available to us. This would allow breeders to check whether planting a specific variety in a new environment that has never been used for field trials in a breeding program is worthwhile. For our dataset, we could not show that EO data can help here. It is possible that a larger training dataset that covers more and more diverse environments could achieve a model where this task is improved.

5. xSeedScore EO module (IP held by Computomics)

We have developed program functions that allow us to automatically access the EO data sets outlined in section 2. Input to these functions is a set of GPS coordinates corresponding to the locations of test fields and the time range covering the planting seasons. Outputs are the aggregated EO descriptors in a form that allows us to include them in our ML model training functions.

This functionality is flexible and can be extended to cover additional EO data types in the future.

6-7. Multi-frequency EM Sensor and Novel Sensor for UAV payload (IP held by SiberGeo)

In addition, since the limitation of the existing soil data (including suggested through PARSEC In-Situ Data Hub) was expected, SiberGeo and Landviser have tested a new mobile universal soil EC sensor (EM) as payload on the octocopter UAV, results would be presented on upcoming Symposium on Applied Geophysics to Engineering and Environmental Problems (Fig. 12a, 12b). The mobile UAV EC sensor (IP SiberGeo) would provide detailed data on soil properties **at multiple depths** for future breeding trials analysis for Computomics' clients, especially those who work on developing stress-resistant (salinity, drought) and nutrient-efficient (N fertilizer utilization) crop varieties.

1. **Objective:** To create an effective way to map soil electric conductivity/resistivity without influence to the soil itself, with high spatial resolution.

As we know from our previous work with EM frequency-domain portable devices, the maximum height of the device above the ground should not exceed 50 cm. Usually, we keep it 15-30 cm above the ground while walking with it.

The first tests at the Novosibirsk field site, where we've modeled the flight at 50-100 cm height show that our devices are very sensitive to the height, the EM response is significantly changing even at the 10 cm of the height variation.

So, the task formula was added with one very important issue: the flight must be very accurate, keeping the height precisely stable.

2. Hardware and Materials: To make the proper UAV platform we've used DJI S1000 frame with a max payload of 6 kg. However, since DJI protocols are closed and any modification of it is very expensive, we've changed all the control units to PixHawk flight controller and ARDUPILOT platform. To keep the precise height, a lidar sensor was added. The max speed was limited to 1 m/sec, with max angles to 35 degrees.



Figure 12: The drone (UAV).



Figure 13: The drone RosAviation registry mark.

3. **Test Flights:** The first flight with Geoviser onboard shows that the flight is possible, but the Geoviser is very heavy. Geoviser sensor is 5.4 kg, and without a controller that makes the drone flight very difficult to manage. However, the first tests were done, the most important conclusion was: we need 3 m between the drone and the EM profiler to avoid the drone signals influencing measurable soil EM field response. Another conclusion was the need for 4-points hanging suspension gear to avoid swinging the profiler in two directions.

4. **Novel Sensor Harness Developed:** The few attempts of a rigid suspension system modeling show that it will be too heavy and useless for a real soil mapping solution. So, the soft 4-cords with fiberglass beams system was developed.

5. **Novel Sensor Compared with Multi-Frequency Sensor:** The new EM profiling device was developed. The weight was decreased to 3.3 kg (with handle). Both devices are shown in Figure 14.



Figure 14: Geoviser (on top) and customized EM profiler (at the bottom).

6. **Results:** The performed test flights show that:

- the residual snow and water can lead to lidar errors;
- After a 180-degree turning point the system need approx 10 sec to stabilize the flight;
- The data collected by the drone is comparable to the data collected by the usual way with an operator;
- UAV flight is the nice way to calibrate an EM device (zero measurement routine);
- At the direct, linear flight the measured data has very high quality and repeatability.

Field experiment Feb, 24

The field experiment included the flight by 3-points flight plan: takeoff (A), 180-degree turn (B), landing (C). The distance between takeoff and landing points is 5 meters. The GPS record of the track is shown in Figure 15.

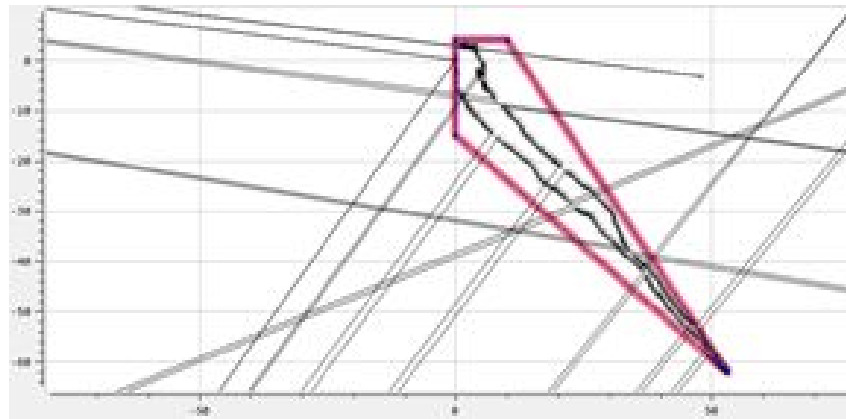


Figure 15: GPS flight tracks.

After the flight, the EM profiler was disconnected from the suspension and the measurement was repeated by the operator (manually).

The record of apparent resistivity is shown in Figure 16.

It can be seen that the drone measurement is essentially more stable, the value of the soil resistivity is very similar. The walkway track is shorter than the drone one because the operator did not walk on the puddles, but the UAV-mounted EM sensor could fly over those and provide complete coverage.



Figure 16: Soil resistivity measurement by the drone and operator.

8. Integrated Global Ag Geophysics Data Store (IP held by Landviser)

The development of Integrated Global Ag Geophysics Digital Data Store (IP Landviser) has been started and would be marketed during the PARSEC Phase II to provide storage, anonymous aggregation, and retrieval of the near-surface soil parameters measured with the new UAV EC sensor (Deliverable 7), other sensors of SiberGeo and Landviser (Deliverable 6), as well as convenient access to global EO data (weather, satellite imagery, soils, and land use) to all clients of Computomics, SiberGeo, and Landviser.

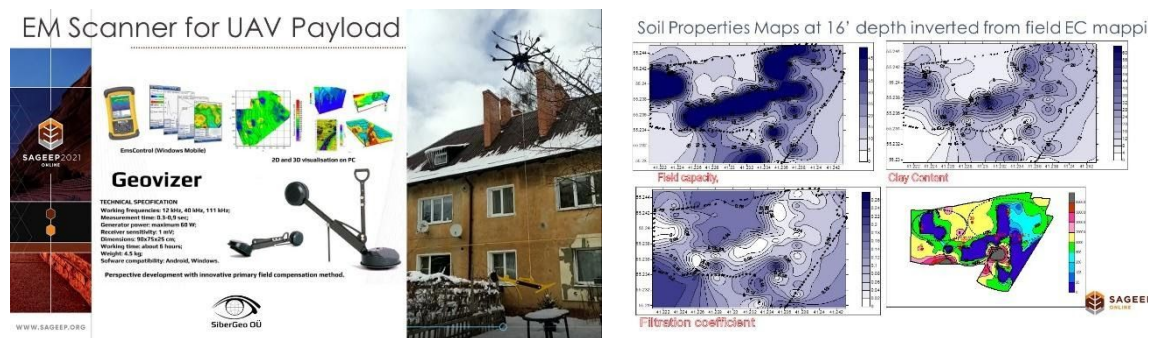


Fig. 17. a. SiberGeo's new airborne EM sensor measuring soil EC at three depths, being developed specifically under PARSEC Business Accelerator.
b. Example of the detailed maps of soil properties at the deeper layer inverted from the soil EC map measured on potato field with LandMapper – EC sensor of Landviser, similar to a new Geovizer UAV.

Landviser has become a Bronze ESRI Business Partner and is using this relationship and GIS software development authorization to build an integrated "Geophysical Data in the Cloud" Repository to make EO data from multiple public platforms readily available through a secure unified portal allowing easily merging such data with Client's proprietary field trial data on crop performance. Such a portal would allow Computomics to quickly and conveniently merge EO data with Genomic information and train the xSeedScore ML module for different Clients (Deliverable 5).

Our 3-steps Approach to Ag Land Survey & Monitor

1. Incorporate Client's Crop Production Data into a Secure Web Mapping Dashboard
2. Gather and Analyze available Soil & Land Use Maps, Satellite Imagery, Climate Variables
3. Map RES/EC at Key Depths with LandMapper and Convert to Maps of Soil Properties
- (4). Incorporate Live Tracking of Severe Weather and Public Health Risk (**Landviser**)
- (5). Install IoT Sensors with LoRa Tower and Network to monitor resources (WP Microsystems)
- (6). Arrange and schedule UAV/MAV imagery collections, UAV spraying (Air Data & Hylío)
- (7). **Supply users and train** on geophysical sensors and software for 2D / 3D subsurface geophysical surveys for soil fertility and crop abiotic stress(SiberGeo, KB Electrometry, TerraZond, AGS/GeoTomo Software)



Figure 18: Landviser's simple yet comprehensive approach to Ag Land Surveying and Monitoring, incorporating EO data and showing our extensive collaboration with other environmental hardware, software, and consulting companies globally.

The EO information from such Portal would also allow breeding companies not only to reduce years and locations for testing new varieties but also provide their customers (farmers) with the insurance and warranties for seed performance and suitability for a particular farm and season (seasonal forecasts would be made available through Copernicus Climate Change Services).

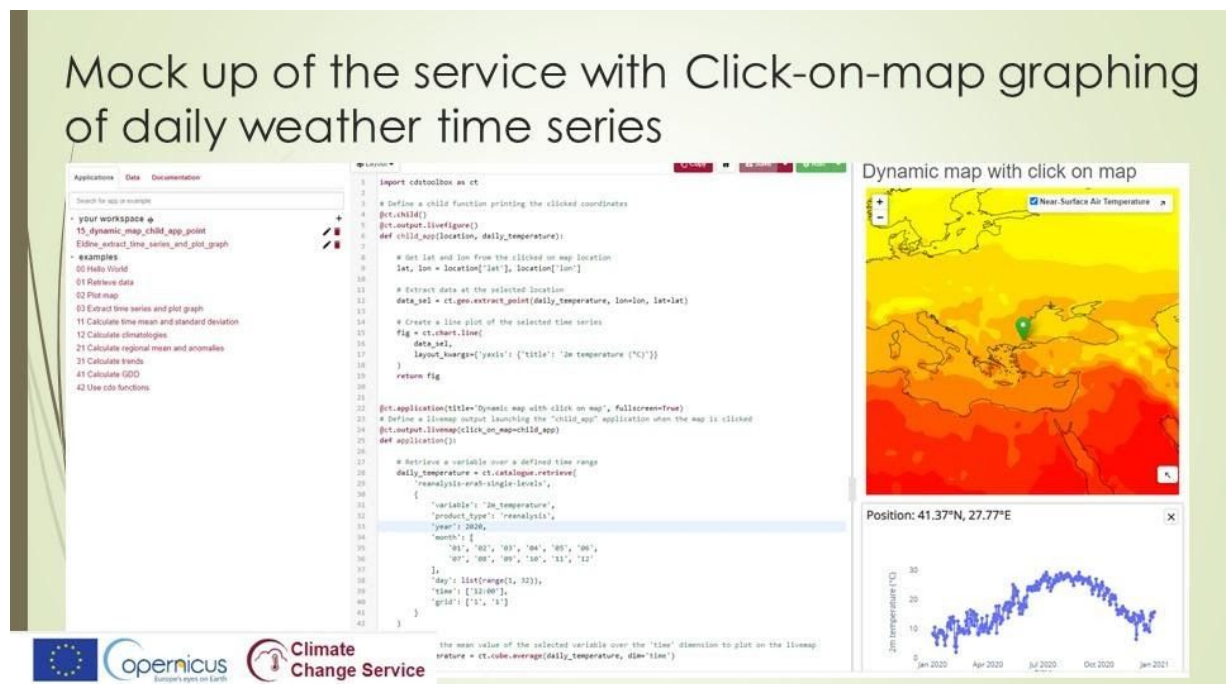


Figure 19: Mockup of UI of the Portal for querying a time-series of Weather Variable per Location of Interest from the Copernicus Climate Change Services.