Data flow for testing remote-sensed data in crop simulations with the view to increase general availability and accessibility of inputs to crop models and boost scale-out.

Luis Vargas Rojas, Davide Cammarano, Diane Wang, Carolina Rivera, Keith Cherkauer

Keywords: data management, workflow, descriptive statistical methods, remote sensing, crop modelling, wheat, weather data.

Date of document completion: April 15, 2021

Synopsis

The Python steps for the extraction, transformation, and exploratory data analysis process to prepare the minimum weather dataset that the DSSAT crop model requires for wheat crop modeling.

Description

This research project aims to fill knowledge gaps to improve remote sensing (RS) techniques for crop simulation Models (CSMs) towards improving scale-out of models for application across larger populations. The research objective is to determine the minimum dataset of vegetation indices required for accurate winter wheat simulations.

Crop simulations will be performed using the Decision Support Systems for Agrotechnology Transfer (DSSAT,(Jones et al., 2003)). For running the crop model (CM), weather, soil, and management data will be collected from the field trials that are established at CIMMYT's (International Maize and Wheat Improvement Center) experiment station CENEB (Campo Experimental Norman E. Borlaug), near Ciudad Obregón, Sonora, Mexico (27.33N, 109.09W, 38m a.s.l.) during the winter cycles 2021/22 and 2022/2023. Even though most of the data will be collected during the next two winter wheat growing seasons (from November to April), historical weather data from an automatic weather station located in the CENEB research station are available, the station is identified as Block 910 (27.36959N - 109.92892W 38m a.s.l.). Additionally, data from a close station called Block 1101 (27.32233N --110.01590W 38m a.s.l.) will be used to fill missing data. These data are part of the inputs for the DSSAT weather module, whose minimum daily weather requirements are solar radiation, minimum and maximum air temperatures, and precipitation. Moreover, evapotranspiration, humidity, and wind speed are needed to increase the estimation accuracy (Hoogenboom et al., 2019).

This proposed plan for the weather data management will reduce the time used when the data preparation is done manually or using programs based on a graphic interface where task automatization is impossible. The data pipeline includes the extraction, transformation, quality checking, and graphical visualization processes for the weather data variables needed by the DSSAT CM.

Source data overview

The Sonora Mexican state's automatic climatic stations are a net of stations called REMAS (Red de Estaciones Meteorológicas Automáticas de Sonora) located in the agricultural regions of northwest Mexico. Moreover, authorized users can download on-time and historical weather data with a limit of 8650 observations (rows) from the REMAS website (www.siafeson.com/remas). When the user performs a query, each file obtained includes all the available parameters individually. This project used historical data from 2015 (Block 910) and 2017 (Block 1101) to March 2021.Data from Block 1101 are used for filling missing data of Block 910 station.

The data included in the downloaded comma-separated values (CSV) files, recorded with a one-hour frequency, are maximum and minimum air temperature (ºC), solar radiation (W/m2), total rainfall (mm), humidity (%), wind speed (km/h), and evapotranspiration (mm). A CIMMYT's database administrator who has the REMAS user account downloads the data, and for this research, he shared them using the Dropbox file hosting service directory.

Methods

This weather data flow process was developed using Python. It involves three main steps: data preparation, data quality checking, and data presentation:

Data preparation

1. Load the files that contain the weather data observations. This process is performed using the *load_raw_dataframe( )* defined function that executes the following processes:
    a. First, create a list of files in the path registered at a JavaScript Object Notation (JSON) configuration file named *config_file.json*. This process requires the *dropbox* python library and uses the Dropbox token access that is registered in the JSON file.
    b. Each file is opened and concatenated into a single pandas data frame with the DateTime field as the index by using a loop statement and the defined function *get_file_dropbox( )*.
    c. The function returns a pandas data frame that contains all each one-hour recorded information.
2. The data frame obtained is clipped into winter growth seasons from November 1st to April 30th, using the *winter_season_filter( )* function.

Data quality checking methods

1. The first method to review the data quality is making interactive graphs using the Python *plotly.py* library. The graphs are built in a web application using the *dash* Python framework. It also has the function to download the graphs into an image.
2. The other method is to count missing values for each season and station using the *count_missing_values( )* declared function.
3. The missing Block 910 station values were filling with data from Block 1101 station.

Summary statistics and metrics

1. The data are resampled into day observations. Each column is aggregated according to the following parameters:
    a. Mean: humidity (%), wind speed (km/h), solar radiation (W/m2).
    b. Max: maximum temperature (ºC).

   c. Min: minimum temperature (ºC).
   d. Sum: total rainfall (mm), evapotranspiration (mm).
2. This step adds a new row that indicates the season and years to which each day observation belongs. It has the following format: *winter_[November to December year months]-[January to April year months]*.

Data presentation

1. The final data frame is presented on a web application using the *dash* Python framework and plotted with the Python *plotly.py* library.

Graphical data analysis

The graphical analysis shows that automatic Block 1101 and Block 901 weather stations have a similar trend for all the weather measurements during the current winter 2020-2021 growth season (Figure 1). Additionally, the maximum temperature is always higher than the minimum temperature for the data recorded in both stations (Figure 1 and Figure 2). For that reason, it is not necessary to make an extra process to clean that data. Figure 3 shows that Block 1101 station has complete records, and it is possible to use them to fill Block 901 station missing data. Finally, figure 4 shows that there is a variation in the total rainfall (mm) recorded. For that reason, we recommend reviewing the automatic stations to figure out if both stations are measuring this parameter correctly.

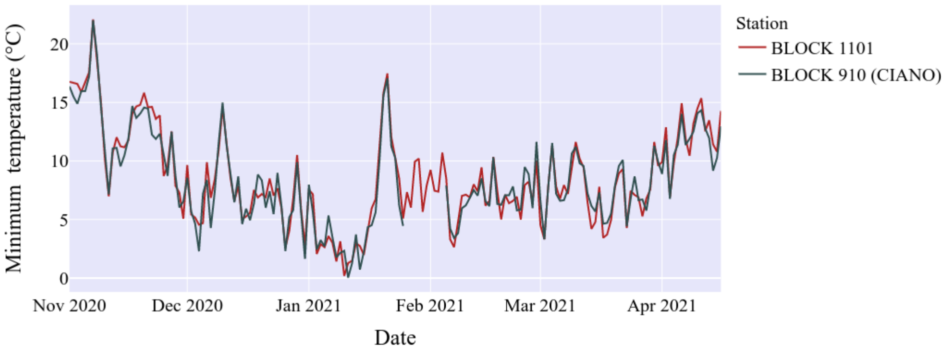

Figure 1. Minimum temperature (°C) recorded from November 1st 2020 to April 15th 2021 on the Block 1101 and Block 901 automatic weather stations.

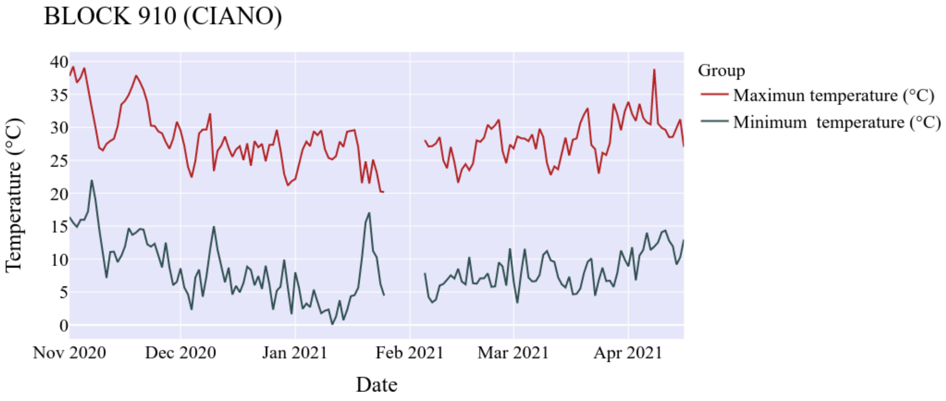

Figure 2. Minimum and maximum temperature (°C) recorded from November 1st 2020 to April 15th 2021 on the Block 901 automatic weather station.

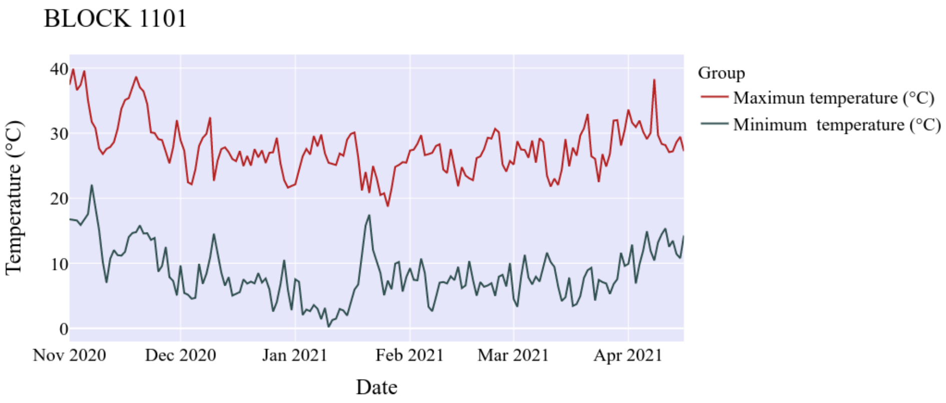

Figure 3. Minimum and maximum temperature (°C) recorded from November 1st 2020 to April 15th 2021 on the Block 1101 automatic weather station.

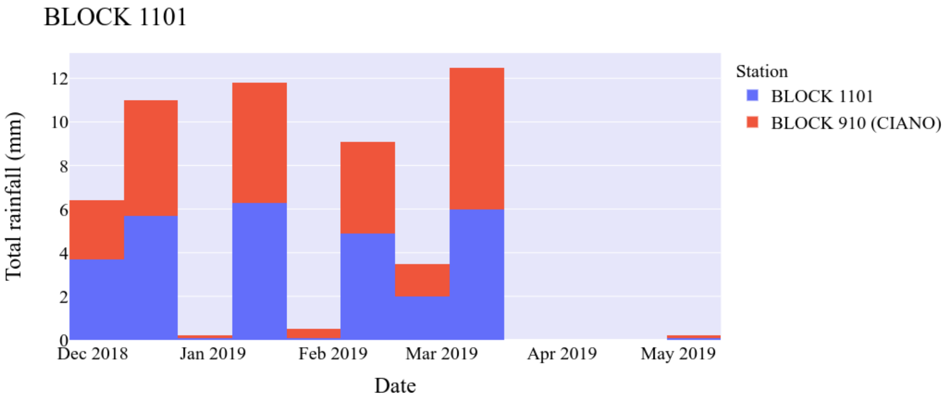

Figure 4. Sum of the total rainfall (mm) recorded from November 1st 2020 to April 15th 2021 on the Block 1101 and Block 901 automatic weather stations.

Data quality checking

The graphical exploratory analysis showed that the data had not observations out normal range. However, 6% of the Block 901 automatic weather station data was missing (Table 1). This data was represented as null. For that reason, there were no additional steps required to find these rows. This data was filled using data from the Block 1101 station. This process filled all the missing data (Table 2).

Table 1. Summary before the quality checking data applied on the Block 901 station for data recorded from November 1st 2020 to April 15th 2021.

Column	Missing values	Total rows	Missing values (%)
Station	0	167	0.00
Maximun temperature (°C)	10	167	5.99
Minimum temperature (°C)	10	167	5.99
Humidity (%)	10	167	5.99

Total rainfall (mm)	10	167	5.99
Solar radiation (W/m2)	10	167	5.99
Wind speed (km/h)	10	167	5.99
Evapotranspiration (mm)	10	167	5.99
Season	0	167	0.00

Table 2. Summary after the quality checking data process applied on the Block 901 station for data recorded from November 1st 2020 to April 15th 2021.

Column	Missing values	Total rows	Missing values (%)
Station	0	167	0.0
Maximun temperature (°C)	0	167	0.0
Minimum temperature (°C)	0	167	0.0
Humidity (%)	0	167	0.0
Total rainfall (mm)	0	167	0.0
Solar radiation (W/m2)	0	167	0.0
Wind speed (km/h)	0	167	0.0
Evapotranspiration (mm)	0	167	0.0
Season	0	167	0.0

Summary statistics and metrics

The DSSAT crop model uses daily weather data. For that reason, the metrics were calculated daily. The mean of each one of the hourly records for each day was calculated for humidity (%), wind speed (km/h), and solar radiation (W/m2). The maximum and minimum temperature (ºC) were obtained from these hourly recorded values of each day. Finally, the records of rainfall (mm) and evapotranspiration (mm) were summarized to obtain daily values. Figure 5 shows an example of a plot obtained from the dashboard that graphs daily calculated values of maximum and minimum temperature.

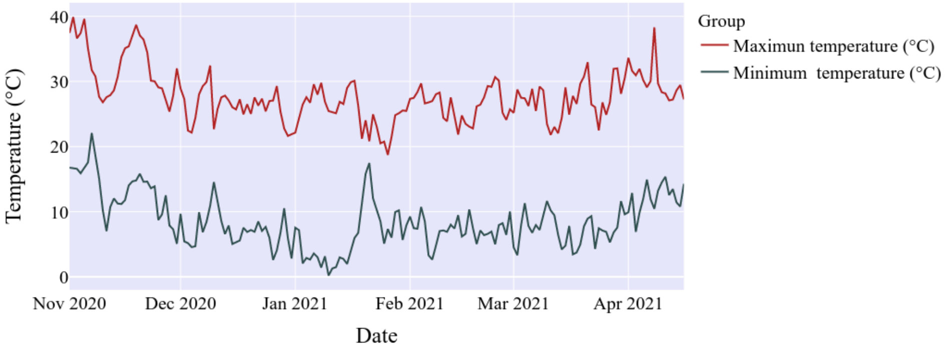

Figure 5. Minimum and maximum temperature (°C) obtained from the hourly records from the Block 901 automatic weather station data.

Using the dashboard features that allow making annotations and shading periods, some graphical analysis was performed to understand how the weather obtained values could affect the winter wheat physiological development. For instance, Figure 6 illustrates that most of the time, maximum and minimum temperature were not under or above the optimal values for sowing and emergency phenological phases (Porter & Gawith, 1999). In addition, Figure 7 shows that during flowering and grain-filling stages several times temperature has been above 31 that could result in thermal heat stress (Rehman et al., 2021).

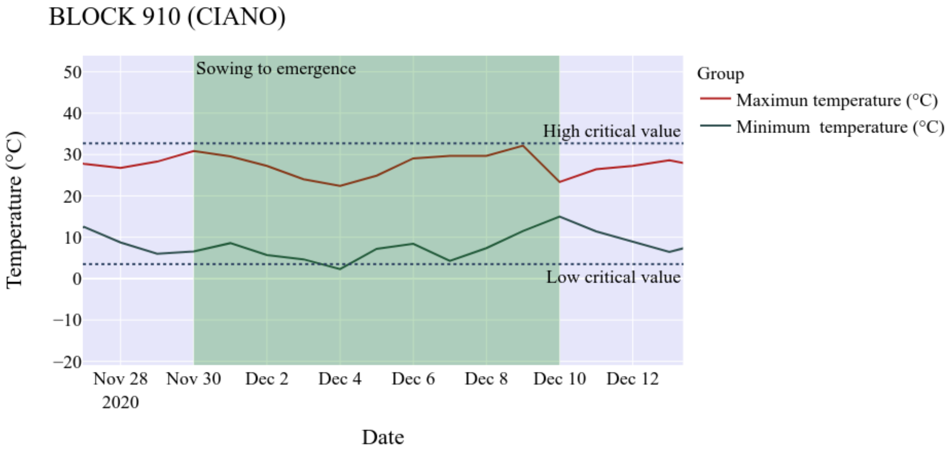

Figure 6. The green shading area represents the sowing to emergence stages of winter wheat planted on the winter 2020-2021 growth season in the plot located at the International Maize and Wheat Improvement experiment station near Ciudad Obregón, Sonora, Mexico. The horizontal lines represent the high and low critical temperature values for crop germination.

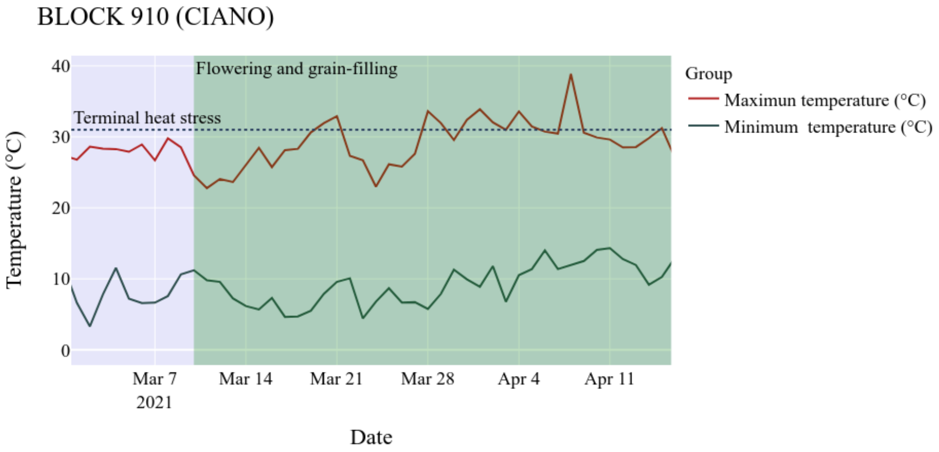

Figure 7. The green shading area represents the flowering and grain-filling stages of winter wheat planted on the winter 2020-2021 growth season in the plot located at the International Maize and Wheat Improvement experiment station near Ciudad Obregón, Sonora, Mexico. The horizontal line represents the temperature value that can cause wheat heat stress.

Conclusions

The Python steps for the extraction, transformation, and exploratory data analysis process allowed us to prepare the minimum weather dataset that the DSSAT crop model requires for wheat crop modeling. Automatic weather stations record the data, and they were downloaded from the Sonora climatic stations' network website. In addition, the cleaning data process showed that the CIMMYT's CENEB research station climatic station dataset has not observations out normal range. Nonetheless, Block 1101 station had approximately 3% of days without data recorded for the 2020-2021 winter growing season. For that reason, these missing values were filled with neighbor station data. The graphical data analysis was performed through an interactive dashboard developed with a Python library. The data were aggregated on a daily basis according to the correct metric that must be applied for each variable. Finally, another interactive dashboard was developed to make a descriptive analysis where weather variables were related to some winter wheat physiology processes.

References

Hoogenboom, G., Porter, C. H., Boote, K. J., Shelia, V., Wilkens, P. W., Singh, U., White, J. W., Asseng, S., Lizaso, J. I., Moreno, L. P., Pavan, W., Ogoshi, R., Hunt, L. A., Tsuji, G. Y., & Jones, J. W. (2019). *The DSSAT crop modeling ecosystem* (Boote & K.J., Eds.; pp. 173–216). Burleigh Dodds Science Publishing.

Jones, J. W., Hoogenboom, G., Porter, C. H., Boote, K. J., Batchelor, W. D., Hunt, L. A., Wilkens, P. W., Singh, U., Gijsman, A. J., & Ritchie, J. T. (2003). The DSSAT cropping system model. *European Journal of Agronomy*, *18*(3), 235–265. https://doi.org/https://doi.org/10.1016/S1161-0301(02)00107-7

Porter, J. R., & Gawith, M. (1999). Temperatures and the growth and development of wheat: A review. In *European Journal of Agronomy* (Vol. 10, Issue 1, pp. 23–36). Elsevier. https://doi.org/10.1016/S1161-0301(98)00047-1

Rehman, H. U., Tariq, A., Ashraf, I., Ahmed, M., Muscolo, A., Basra, S. M. A., & Reynolds, M. (2021). Evaluation of Physiological and Morphological Traits for Improving Spring Wheat Adaptation to Terminal Heat Stress. In *Plants* (Vol. 10, Issue 3). https://doi.org/10.3390/plants10030455