Climate Services 30 (2023) 100341

Contents lists available at ScienceDirect

Climate Services

journal homepage: www.elsevier.com/locate/cliser

Forecasting soybean production to enhance climate services for Agriculture in Argentina

Esteban Otto Thomasz [a,*], Kevin Corfield [b], Ana Silvia Vilker [c], Marisol Osman [d,e,f]

[a] Universidad de Buenos Aires, Facultad de Ciencias Económicas, IADCOM, Programa Vulnerabilidad al Riesgo Climático (ProVul), Argentina
[b] Universidad de Buenos Aires, Facultad de Ciencias Económicas, IADCOM, CMA, Programa Vulnerabilidad al Riesgo Climático (ProVul), Argentina
[c] Universidad de Buenos Aires, Facultad, IADCOM, CMA, Programa Vulnerabilidad al Riesgo Climático (ProVul), Argentina
[d] Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales Departamento de Ciencias de la Atmósfera y los Océanos, Argentina
[e] CONICET-Universidad de Buenos Aires, Centro de Investigaciones del Mar y la Atmósfera (CIMA), Buenos Aires, Argentina
[f] CNRS-IRD-CONICET-UBA, Instituto Franco-Argentino para el Estudio del Clima y sus Impactos (IRL3351 IFAECI), Buenos Aires, Argentina

ARTICLE INFO

Keywords:
Soybean yields
Soil water content
Impact evaluation
Forecast

ABSTRACT

Soybean is the most important agricultural commodity produced and exported by Argentina. Water is the most important input for soybean production. Water deficit during critical period and water excess during harvest affect output. Even though it is known that droughts generate important economic losses in soybean production, a standardize model that provides sensibility analysis to explain at what extent water affects output remains a challenge. Therefore, the relation between precipitation levels, the standardized precipitation evapotranspiration index (SPEI) and soil water content was explored, being the soil water content the index with the better performance. By means of a correlation and regression analysis, it was found that in the major number of cases soil water content explains at least 50% percent of the variability in soybean yields, with a maximum of 70% explanatory power in one county. Also, forecast capacity was tested through leave one out validation technique, showing that models are robust enough to provide a one period forecast, as error is mostly explained by the standard deviation of the yield index. The main applications of this finding are related to impact evaluation before final harvesting and the design of index-based insurance.

Practical Implications

Despite the increase in the availability and scope of climate data, the relation of climate variables and crop reaction is not easily found specially when applying the same methodology for large territories, as the case of the geographical scale of the agricultural production data set available in Argentina.

This line of research aims to provide climate indicators that are specifically associated with crops reactions. The indicator tested in this paper, is a unique development because not only calculates the soil water content (which summarize a whole set of climate variables) but calibrates it for specific crops. In this line, the index measures not the quantity of water in the soil, but the quantity that is useful for the crop, in this case, soybean.

Therefore, the main contribution of this paper is to test if the soil water content indicator currently published by the Office of Agricultural Risk of Argentina (ORA) is robust enough to explain the crop performance. If the relation is proven, two main applications are derived, one public and one private:

(i) Implement a standardize and reliable impact evaluation system for the ORA. This system provides assistance to producers in case of losses generated by climate events, but at the moment, there is not a data-driven monitoring system. The emergency declaration is initiated by the request of the local authorities of the affected department, which is time consuming due to bureaucracy and verification process and might not provide equal treatment to different affected areas. The implementation of the new evaluation system contributes to optimize and transparentize the agricultural emergency system
(ii) Serve as basis for index-based insurance design. The soybean area of 17 million hectares makes very costly for insurance companies to make *in situ* verification of the amount of loses. This generates that insurance cost is higher and only large

* Corresponding author.
E-mail address: ethomasz@hotmail.com (E.O. Thomasz).

https://doi.org/10.1016/j.cliser.2023.100341
Received 30 July 2021; Received in revised form 13 October 2022; Accepted 24 January 2023
Available online 6 February 2023

producers take the coverage, leaving smaller and more vulnerable producers with no coverage. Index based insurance is designed to avoid *in situ* inspections, triggering the payment in relation to the value of an index. To do that, the index must have two main conditions (Thomasz and Casparri, 2015). (1)- be highly correlated with the assured asset, in this case soybean output and 2- have certain conditions such as open access and easily measured, objective, transparent, independently verifiable, published periodically and consistent over time. The second set of conditions are achieved by the indicator analyzed in this paper (it is calculated and published periodically by a public institution, free access and with open-source data, with a clear and published methodology, and with enough history to back test the performance). Therefore, the main contribution of this study will be to test condition (1). Evaluating at what extent the index of soil water content can explain or even predict the effect over soybean output constitutes one key element to the development of further applications. The essence of the utility of climate services is not only how they explain or predict climate but how it will affect the human or economic activity.

The results presented are part of an interdisciplinary project (strategic development project) funded by the University of Buenos Aires (UBA) in which the ORA from the Ministry of Agriculture was the beneficiary institution. As was mentioned before, the model documented in this work will be considered by the ORA in the decision-making process related to the management of the agricultural risk associated to climate variability. This office works in collaboration with the research group in the development of this research and was consulted throughout the development of the project.

The findings of this work will then support the planification of contingency measures. In addition, the results obtained in this research are being communicated to the whole agricultural sector through a bulletin issued three times per year.

The estimations of soybean production obtain in this research are based on the relationship between those and the observed summertime water reserve. The good performance observed under a cross-validated approach allows us to implement this model in a real time framework and the communication of the output results to the ORA two months prior to the end of each soybean campaign. Moreover, the results obtained with observed variables motivate us to explore the use of the model with forecast soil water content information instead of observed values to obtain estimations of soybean production as early as December (5 months prior to the end of the campaign). If this second step is successful, it could lead to a better planning of the assistance by the Ministry of Agriculture in case of an emergency.

In the future, this forecast could be provided at county level will be a critical to decide coverage taking (insurance of other hedging tools) by producers.

Introduction

Soybean is the most important agricultural commodity produced and exported by Argentina. At a global scale, Argentina's yearly production of 50 million tons is the third largest in the world after the US and Brazil, representing 16 % of global production. Local production is export-oriented but with an important industrialization chain: beans represent 20 % exports while mill and oil accounts for the 80 % value of sector exports. Argentina concentrates 40 % of international trade of soybean mill (USDA, 2021), being the most important player in that market. At the national level, soybean accounts for 81 % of crop production and 77 % of the sown area during recent years. Soybean and derivates represented on average 24 % of the country's total value of exports between 2003 and 2020. At the subnational scale, it is the main economic activity for many counties across the country (Massot et al., 2016).

Water is the most important input for soybean production. Water deficit during critical period and water excess during harvest affect output. However, a proper impact evaluation on how water affects production remains a challenge in the case of Argentina. Only a few studies on impacts of climate variability and climate change on agriculture have been conducted in Argentina, with a focus on crop response to projected carbon dioxide emissions (Barros et al., 2014; CEPAL, 2014; Murgida et al., 2014; Ortiz de Zarate and Ramayon, 2014). Magrin et al. (2007) studied the impact of climate on yields in the past, until 1999. Letson et al. (2009) studied the effect of a decrease in rainfall in two counties, Pilar and Pergamino. Bert et al. (2006) developed a crop simulation model of climate variability for maize in one county, Pergamino. Barros et al. (2015) explores impacts of fluvial and lingering floods but does not introduce a valuation methodology. More recently, Thomasz et al. (2017) verified that severe and extreme droughts severely affect soybean output in Argentina, designing a valuation methodology at county scale, but lacking a sensibility analysis.

Even though it is known that droughts generate important economic losses in soybean production, a standardize model that provides sensibility analysis to explain at what extent water affects output remains a challenge. Also, a model that can contribute to estimate economic losses from a cash flow perspective, necessary to evaluate the financial liability of adaptation investments, has not yet been developed. Moreover, a model that can also be calibrated to generate forecast to be used for hedging decisions is still underdeveloped. Thus, this work addresses the lack of a reliable, standardized and replicable model that can relate losses in soybean production with water. One of the main objectives of the Argentina Ministry of Agriculture is the early estimation of losses in crop production as a tool to manage the agricultural risk. Then, the availability of a better model for soybean losses can help address this risk. In this sense, taking into consideration the soybean production cycle and the availability of weather information, impact evaluation can be forecasted between two and three months before the end of campaign. Also, a model could be used to forecast future performance at the very beginning of the campaign, providing useful information to taking coverage through insurance or other hedging instruments.

In this context, which indicator of water availability has the most accurate relation with soybean yields at county scale and which is capable of being forecast is ongoing research.

Precipitation has been widely used to relate water variability with soybean production. Among the indexes used to characterize periods of water deficits (droughts) are the Palmer Drought Severity Index (PDSI) and the Standardized Precipitation Index (SPI). However, temperature variability can also induce drought stress through an increase in evapotranspiration. To account for this effect, the Standardized Precipitation-Evaporation Index (SPEI) was developed. The SPEI combines precipitation and potential evapotranspiration and accounts for the intensity and duration of droughts. Soil water content is another useful variable to diagnose the current state of crops since it also includes information on the water demanded by the overlying vegetation. During recent years, the ORA elaborated a water reserve monitoring system, which is calibrated for different crops across the country.[1] This data set, which synthesizes a whole set of critical weather variables, is a more comprehensive index available focused on water availability for crops. The availability of all these indexes to characterized water availability motivates the development of model to link them with soybean production.

In this line, the relation between precipitation levels, the SPEI and soil water content was explored, being the last one the index with the better performance in the first stages of the analysis (Annex 1).

[1] International cases of monitoring index can be found at the *Climate Prediction Center* of the US, at the *European Drought Observatory*, the SISSA of Southern South America and the *Australian Bureau of Meteorology*.

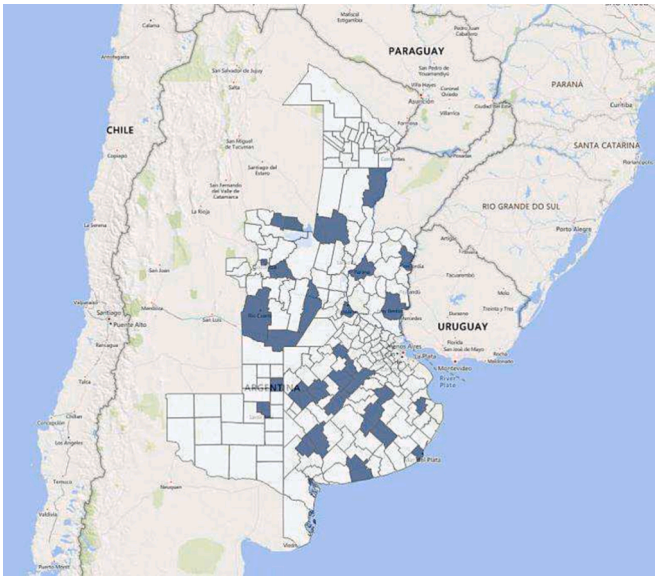


Chart 1. Soybean area and counties with meteorological station.

Table 1
Counties' name and province of location.

Province	Province code	Counties	#
Buenos Aires	BA	9 de Julio, Azul, Bolívar, Coronel Suarez, Dolores, General Pueyrredón, Junín, Las Flores, Olavarría, Pehuajó, Saavedra, Tandil, Trenque Lauquen, Tres Arroyos	14
Córdoba	CO	Córdoba, Marcos Juárez, Presidente Roque Sáenz Peña, Río Cuarto, Río Seco, Río Segundo	6
Entre Ríos	ER	Concordia, Gualeguaychú, Paraná	3
La Pampa	LP	Capital, Maracó	2
Santa Fe	SF	General Obligado, Rosario, San Cristóbal	3

Therefore, the relation between the soil water content and soybean yields will be tested into detail for different time windows and aiming to answer three main questions: i) which time window provides the best relation, ii) which is the sensitivity level between soil water content and soybean yields for each county, and iii) the possibility to use the estimates to perform forecast analysis.

This study is part of a UBA interdisciplinary project "Applications of climate forecast data to manage agricultural risk". In this project, climate scientists and economists have partnered with the ORA to help them manage the climate-driven agricultural risk. Two main applications are derived from results to feed future research: impact evaluation and forecast model. Both elements aim to help decision making regarding fiscal assistance, crop choices and hedging instruments such as index-based insurance.

The paper is organized as follows: section 2 describes the study area, all the data sets of information used and the methodology to estimate the relation between soil water content and soybean yields. Section 3 summarizes results presenting (i) the time span window that maximizes the correlation, (ii) the explanatory power of the relation and (iii) the forecast capacity of the model. Last, some concluding remarks are presented.

Material and methods

Study area

The study area is comprised of the 28 soybean production counties which have a territorial weather station with enough historical agroclimatic data to perform statistical time series analysis (Chart 1).

Counties' name, province of location and number of counties per province are resumed in Table 1.

Weather and climate data

Precipitation

Daily rainfall observations from 28 stations of the Argentina National Weather Service (SMN) distributed throughout Argentina were considered. Monthly accumulated rainfall series for those stations were obtained by summing the daily values. The stations considered were those closest to the soybean production counties used in this study. Evolution of this variable during the soybean season for the top 6 soybean production departments of the sample is presented in Chart 2.

The SPEI

Monthly values of the SPEI for the same 28 stations of Argentina were considered. This index has been proved to be helpful to measure drought intensity and duration as well as to identify the onset and end of drought episodes (Vicente-Serrano et al., 2010). A major advantage of the SPEI over other drought indexes (like the Palmer-Drought Index or the Standard Precipitation Index) is that the SPEI consider the effect of the potential evapotranspiration (PET) on drought severity. Including information of the potential evapotranspiration allows the quantification of the effect of temperature on drought conditions. This is particularly relevant for severe droughts. The data was obtained from the Regional Climate Center for Southern South America. The computation of the SPEI follows Vicente-Serrano et al. (2010) and is the monthly difference between precipitation and PET. Evolution of this variable for the top 6 soybean production departments is presented in Chart 3.

Soil water content

Daily data of soil water content produced by the ORA for the same 28 weather stations ORA were used. The soil water content is obtained through the equation proposed by Forte Lay and Burgos (1983) and includes the definition of evapotranspiration due to vegetation developed by the Food and Agriculture Organization (FAO, 1978). In addition, the definition accounts for the underground and surface runoff and the current state of the soil conditions, in particular the crop sewed and its life cycle. In this study, the soil water content for soybean was considered. To compute the soil water content, daily data of precipitation, maximum and minimum temperature, 10-meter winds, relative humidity and heliophany from conventional weather stations is used. The details of the methodology are summarized in Basualdo (2020). Evolution of this variable for the top 6 soybean production departments is presented in Chart 4.

Agricultural data

Soybean production information consists of yearly data of planted area, harvested area, production level and yields per county from 1970 to 2020. Open data source is the Argentinean Ministry of Agriculture. From the total sample of soybean data, the 28 counties with a territorial-based weather station with enough historical data to perform time series analysis were selected. Data of first-class soybean is used, which represented 83 % of production and 78 % of the sown area during the last 20 seasons. Second class soybean is not analyzed because its development is recent in time and has limited sown area, therefore there is not enough data to perform the statistical analysis presented in this paper. Evolution of this variable for the top 6 soybean production departments is presented in Chart 5.

Soybean yields model

Soybean yields is the predictand variable in this study, which variability can be resumed into two main factors: (i) technological variables

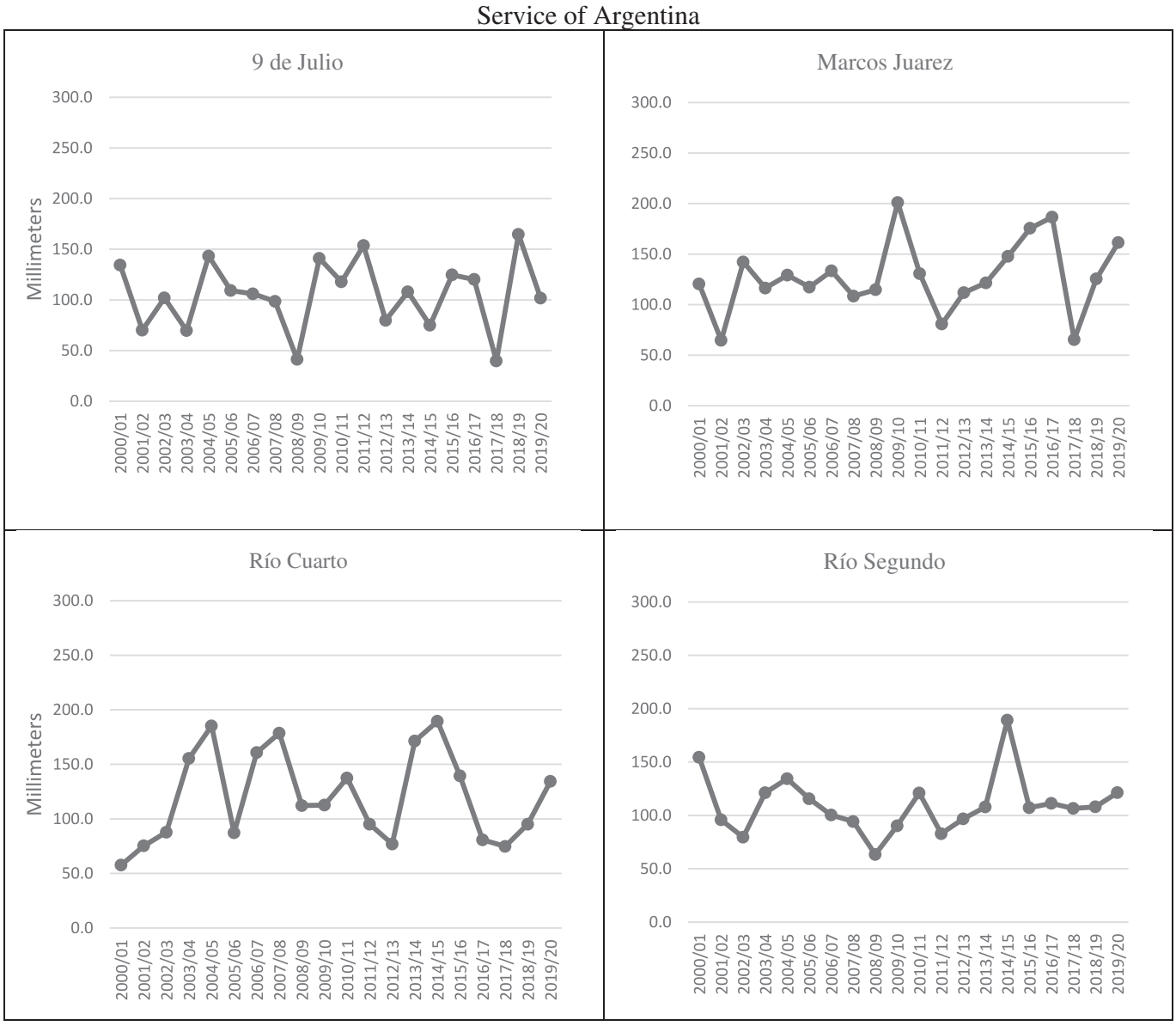


Chart 2. Time series of average rainfall over austral Summer (December, January, and February) for the top-6 soybean production departments. Data source: National Weather Service of Argentina.

such as soil quality, seed genetics, and producer-level management techniques and (ii) climatic variables such as averaged and maximum temperature, accumulated rainfall and others (Thomasz et al, 2019; Lobell and Burke, 2010; Rahman et al., 2005; Paltasingh et al., 2012; Chimeli et al., 2008).

Despite this complexity, yields tend to show a general increase over time, which is commonly referred to as the "trend yield" (Tannura et al., 2008). This phenomenon is present in most of the soybean yields analyzed in this paper (Chart 5). It is accepted that this structure is related to the incidence of technology (Irwin and Good, 2015, Tannura et al., 2008). Considering that there is not available information at department levels of the technology applied each year, is it not possible to perform a multi-regression model. Therefore, the trend yield will be filtered to focus the analysis on the factors that explain variability. To filter the trend, a linear time dependent regression is applied. An alternative sometimes considered is the log-linear trend model, but this model also implies that the range of trend yield deviations in bushels should expand across time which clearly does not happen (Tannura et al., 2008). In Thomasz et al. (2016) the difference between the linear trend and log-linear trend was tested, showing the effect of distortion of yields variability mentioned above. Also, in Thomasz et al. (2017) the linear trend was tested for all 250 soybean departments of Argentina, proving the significance of coefficient in all cases. Therefore, a linear model will be used to de-trend the series.

The de-trended yields series is scaled into an index of relative deviations for better visualization and comparison among counties. Therefore, the final predictand variable to be related to climate predictors is the soybean de-trended yield index, which is constructed as follows:

Current yields are $Y_t = \frac{Q_t}{A_t}$, where Q_t is soybean quantity in tons per county in year t and A_t is area harvest in year t in hectares. From current yield series, a linear model is estimated:

$$\overline{Y_t} = \beta_0 + \beta_1 Y_t + e_t \quad (1)$$

With β_0 as intercept, β_1 as trend and e_t as stochastic error. From the estimated parameters the yearly estimated yield is:

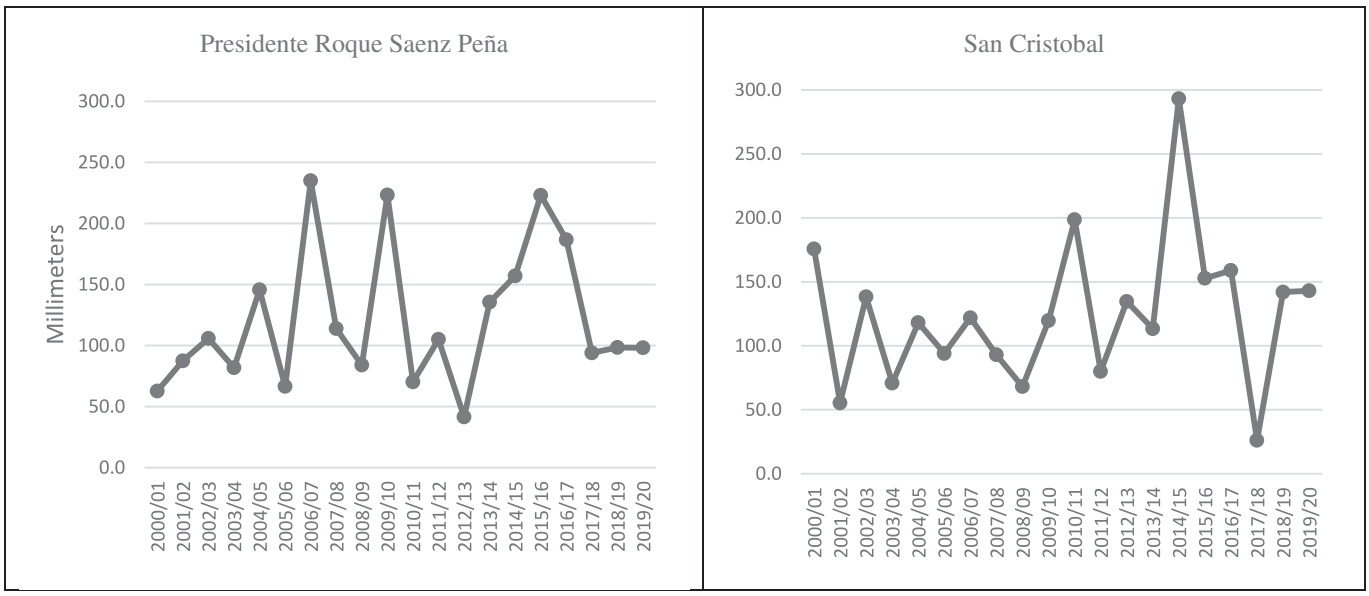


Chart 2. (*continued*).

$$\widehat{Y}_t = \beta_0 + \beta_1 * \mathrm{t} \tag{2}$$

where t is every year of the time series. The de-trended yield $\widetilde{Y}_t$ is calculated as the difference between the observed Y_t yield and estimated yield $\widehat{Y_t}$:

$$\widetilde{Y}_t = Y_t - \widehat{Y}_t \tag{3}$$

The de-tended series is centered in cero with positive and negative values. This work will test to what extent the variability of the de-trended yield is explained by changes in soil water content. For better visualization and comparison among counties, the series is scaled into the following index number:

$$IndexYI_t. = 1 + \left(\frac{\widetilde{Y}_t}{\widehat{Y}_t}\right) \tag{4}$$

Which is the soybean de-trended yield index from which relations with soil water content will be tested.

Soybean yields and soil water content relations tested

Relations between soybean de-trended yields (*soybean index*) and soil water content are tested into three steps:

i. Determination of the optimal critical time span period between soybean index and soil water content by selecting the maximum correlation level through moving windows with different lengths.
ii. Constructing a linear regression with the optimal time span window of soil water content determined in the first step.
iii. Perform a leave one out cross validation technique and report the mean quadratic error to test prediction robustness.

i. Optimal time window

The yield index for a given year is a unique value, while the soil water content values are high frequency with data for the lapse of 10 days. In this framework, soil water content will be averaged for different time windows and then correlations with the yield index will be calculated.

Since the aim is to forecast the yield index for each county, it is necessary to evaluate the critical period of the crop in each zone. Therefore, the optimal time window is determined empirically by the maximum correlation between yield index and soil water content.

The results were obtained by first analyzing correlation between soil water contents and yield index. Then, the maximum level of correlation for each county was determined for different time windows (30, 40, 50 and 60 days), moving each window ten days between December 1st and April 30th (Table 2), which is the soybean growing season. Within the growing season, there are critical periods in which water is more necessary for the crop development (phenology). By moving the start date and length of the soil water content period, we seek to determine the time window with the highest correlation consistent with crop phenology.

Thus, 1278 different values of soil water content for each county are calculated to correlate with the single value of soybean index. The detail of the start and end date of every window is resumed in Annex 2.

The mean of the soil water contents ($\widetilde{hr}_t^{c,s}$) for each county is calculated by changing the amplitude of the window "s" in each period "t":

$$\widetilde{hr}_t^{c,s} = \frac{\sum_s hr_t^{c,s}}{n_a} \tag{5}$$

where:

c = county.
t = year.
s = window amplitude: 30, 40, 50, 60 day.
n_a: number of observations per window: $n_{30} = 3$; $n_{40} = 4$; $n_{50} = 5$; $n_{60} = 6$.

Let equation (4) be the Yield Index and ρ_c the Pearson correlation between Yield Index and soil water content for county "c", aiming to get:

$$\rho_c^* = sup\left\{ corr\left(YI_t^c, \widetilde{hr}_t^{c,s} \right) \right\} \tag{6}$$

The objective is to find the maximum values of the association between the variables for each county in order determine the best time span of soil water content to use as predictor of the linear forecasting model.

It is important to highlight that the determination of the optimal window is one of the main applications of this study. It allows insurance companies or monitoring agencies to specially focus on the time span in which climate values have the highest incidence on output.

ii. Regression model for the yield index per county

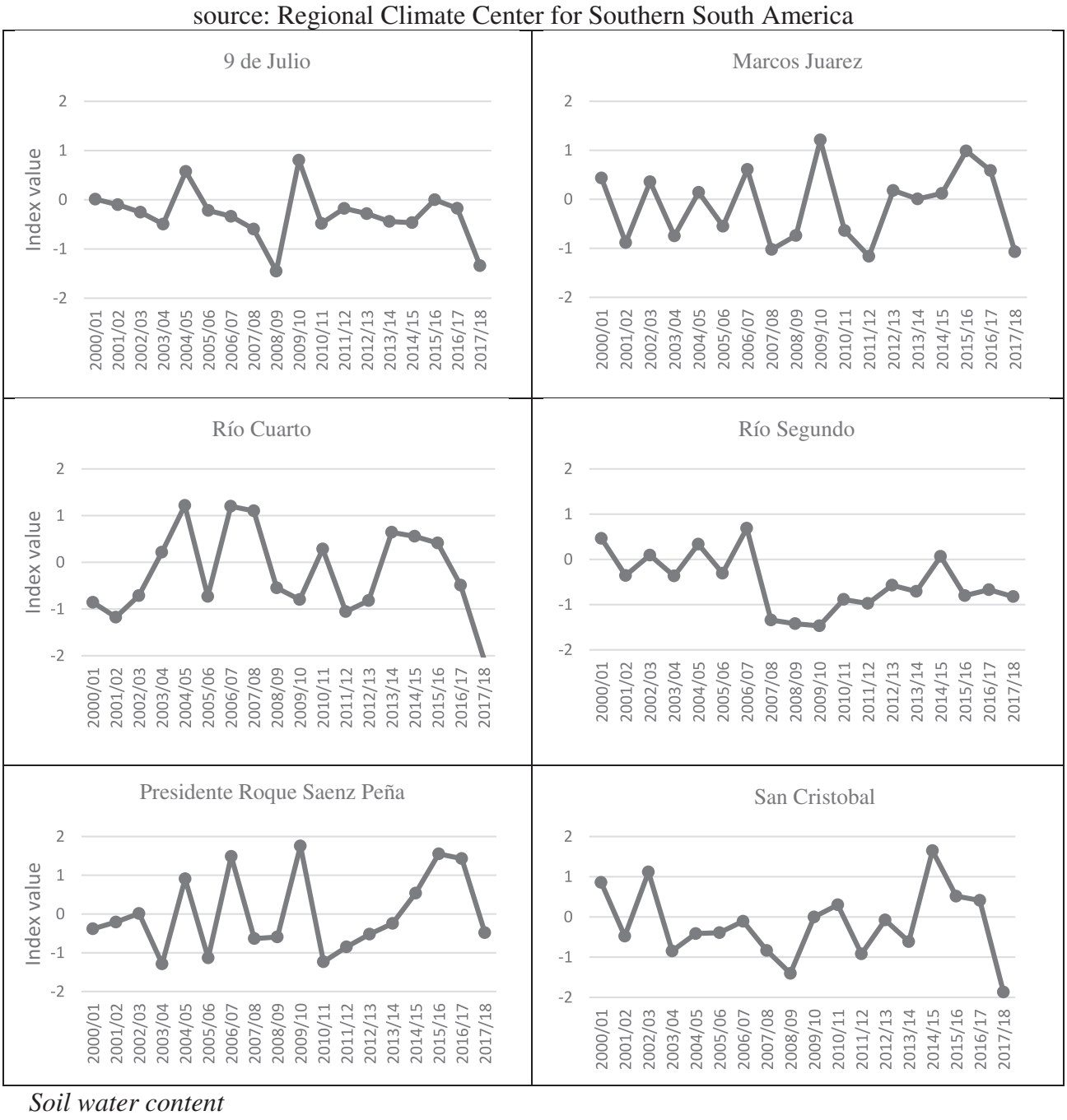


Chart 3. Same as Chart 2 but for Standardized Precipitation-Evapotranspiration Index. Data source: Regional Climate Center for Southern South America.

Once the maximum correlation coefficient ρ_c^* for each county has been obtained by testing different time windows, a simple linear regression was computed using the yield index as the predictand and the average of soil water content corresponding to the maximum level of correlation $\overline{hr_t^c}$ as the predictor. The estimated equation for each county is:

$$YI_t^c = \beta_0 + \beta_1 \overline{hr_t^c} + \varepsilon_t^c \quad \varepsilon_t^c \sim \ iidN\left(0, \sigma^2\right) \tag{7}$$

As is standard in the literature, the goodness-of-fit values and the coefficients of regression are evaluated. Then a series of tests were calculated to determine the behavior of the residuals of the model: (i) the Durbin Watson test to detect serial correlation in the residuals and (ii) the Jarque Bera test to check the normality of the residuals.

iii. Model forecast capacity test: leave one out cross validation.

To test the robustness of the simple models and estimating prediction error, leave one out cross validation is employed. This method consists of randomly defining a part of the observations to fit the model and another part to test it.[2] Since sample size is small, it is convenient to define the data for the model estimation in t-1 observations.

The last step was testing the assessing prediction error with the follow formula:

[2] For more details see Hastie, T. et al. (2009).

Chart 4: Same as Chart 2 but for Soil water content Index. Data source: ORA

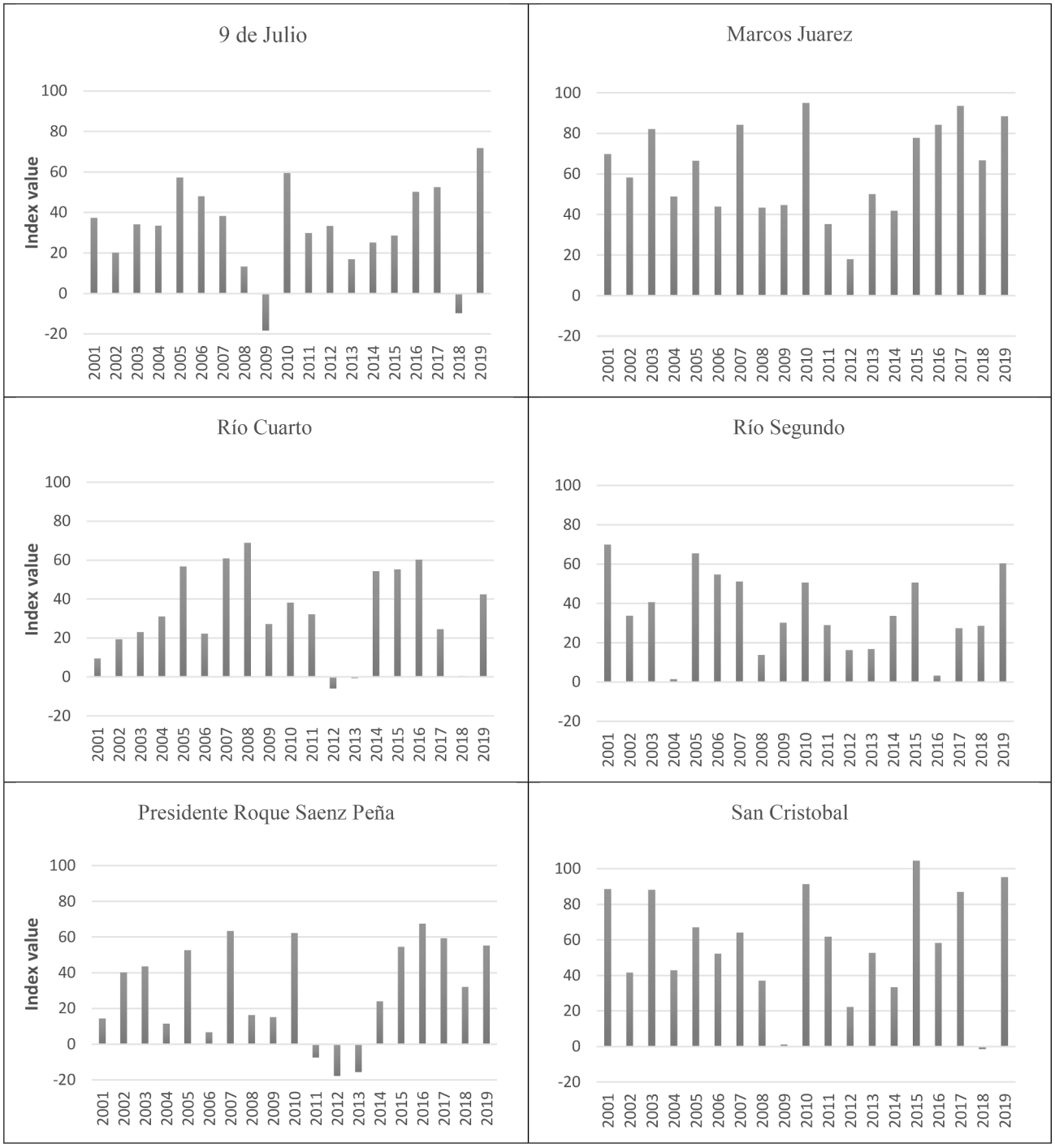


Chart 4. Same as Chart 2 but for Soil water content Index. Data source: ORA.

$$\Lambda = 1 - \frac{RMSE}{\sigma_{YI_t^c}} = 1 - \frac{\sqrt{\sum_{t=0}^{T} \frac{\left(\widehat{YI_t^c} - YI_t^c\right)^2}{T}}}{\sigma_{YI_t^c}} \tag{8}$$

where $\widehat{YI_t^c}$ is the estimated Yield Index for period t and county c, YI_t^c is the observed Yield Index for period t and county c and $\sigma_{ir_t^c}$ is the standard deviation of Yield Index of county c.

If the forecast error is fully explained by the variance of the variable to be forecast, the indicator Λ is equal to zero. The higher the forecast error, the higher the ratio will be greater than unity and the lower the indicator Λ will be less than zero.

As it was said in the introduction, several climatic variables were studied to explain the yield index, such as rainfall in the critical period, minimum, maximum and mean temperature and the SPEI. However, the best results at the county scale were obtained with soil water content, which makes sense since it incorporates different aspects related to climate while adding agronomic variables for each crop. Annex 1 presents the correlation between the yield index and the precipitation, the SPEI and soil water content. Results show that most of the counties present the highest correlation with soil water. Therefore, the results section will show in detail the results for soil water content only.

Chart 5. Time series of soybean yields (in tons per hectare) for the top-6 soybean production departments. Data source: Ministry of Agriculture of Argentina.

Table 2
Total windows by amplitude and number of correlations calculated in each of them.

Window	Start-end date	Total windows	Correlations
30 days	01 December-30 April	13	354
40 days	01 December-30 April	12	336
50 days	01 December-30 April	11	308
60 days	01 December-30 April	10	280

Results

The analysis focuses on the relationship between soybean index and soil water content for different time windows and times dates. Average values of soil water content of 30, 40, 50 and 60 days were tested into moving windows ranging from the 1st December to 30th April obtaining 1278 combinations that were tested for each county (Annex 2). Maximum correlations values with corresponding time window and dates are summarized in Table 3.

Correlation values coefficients are distributed as follows: 4 counties show correlations between 84 % and 80 %, 12 are in the interval between 70 % and 80 %, and 9 are between 50 % and 70 %. The remaining 2 counties are below 50 %. The time span which maximizes correlation is consistent with the phenology of soybean, with an average critical period during January and February across the territory. On the other hand, Tandil exhibits a negative correlation index and Maracó shows the

Table 3
Maximum correlations between soybean yield index and soil water content for time window and date.

Province	County	Correlation	Window	Date
BA	9 De Julio	0.73	40	1 Jan – 10 Feb
BA	Azul	0.79	60	1 Jan – 28 Feb
BA	Bolivar	0.77	60	1 Jan – 28 Feb
LP	Capital	0.62	30	21 Jan – 20 Feb
ER	Concordia	0.77	30	11 Jan – 10feb
CO	Córdoba	0.73	60	1 Jan – 28 Feb
BA	Coronel Suarez	0.76	40	21 Jan – 28 Feb
BA	Dolores	0.84	60	21 Dec – 20 Feb
SF	General Obligado	0.80	60	1 Mar – 30 Apr
BA	General Pueyrredon	0.71	30	21 Dec – 20 Jan
ER	Gualeguaychu	0.63	40	11 Jan – 20 Feb
BA	Junin	0.66	40	21 Mar – 30 Apr
BA	Las Flores	0.77	60	20 Dec 20 Feb
LP	Maracó	0.18	30	21 Jan – 20 Feb
CO	Marcos Juarez	0.55	30	21 Dec – 20 Jan
BA	Olavarria	0.68	60	1 Jan – 28 Feb
ER	Paraná	0.76	40	1 Jan – 10 Feb
BA	Pehuajo	0.62	60	1 Jan – 28 Feb
CO	Presidente Roque Saenz Pena	0.83	40	1 Jan – 10 Feb
CO	Rio Cuarto	0.70	50	10 Jan – 28 Feb
CO	Rio Seco	0.46	30	21 Jan – 20 Feb
CO	Rio Segundo	0.52	30	10 Jan – 10 Feb
SF	Rosario	0.69	50	11 Dec – 10 Feb
BA	Saavedra	0.75	40	1 Feb – 10 Mar
SF	San Cristobal	0.80	60	21 Dec – 20 Feb
BA	Tandil	−0,24	30	10 Dec – 10 Jan
BA	Trenque Lauquen	0.58	60	1 Jan – 28 Feb
BA	Tres Arroyos	0.79	50	11 Feb – 31 Mar

lowest correlation (0.18). According to the resulťs validation given by the ORA, the weak relation between soybean yields and soil water content in these two departments can be explained by some particularities of the territory and climate events, mainly the incidence of frost in Tandil and the water table in Maracó. In sum, at this stage of research, soil water content is not yet the most accurate index to explain yields variability in those counties.

With the values of soil water content of the optimal time windows resumed in Table 4, the regression analysis is performed. Results are summarized in Table 4.

As general result, it can be said that sixty percent of the counties have a coefficient of determination greater than 50 %. In particular, the goodness-of-fit of the models is within the range of 70 % to 64 % for the counties of Dolores (Buenos Aires province) Presidente R. S. Peña (south of Córdoba province) General Obligado (northeast of Santa Fe province) and San Cristóbal (center-west of Santa Fe province). On the other hand, goodness-of-fit between 63 % and 50 % are obtained in the province of Buenos Aires from north to south −9 de Julio, Azul, Bolivar, Coronel Suarez, Saavedra and Tres Arroyos-. In the province of Entre Ríos, also from north to south -Concordia and Paraná-. Finally, in the province of Córdoba, the Córdoba district. Eight counties −32 % of total- have goodness-of-fit between 49 % and 35 %. These include Junín, Pehuajó and Olavarría in the province of Buenos Aires, Río Cuarto in Córdoba, La Capital in La Pampa and Rosario in Santa Fe. The rest of the universe analyzed shows values below 35 %. Territorial location of values of Rˆ2 are summarized in Chart 6.

The results show that in 20 out of 26 counties the explanatory power of the regression is at least 50 %, with ten of them above 60 %, meaning that in general changes in the yield index are explained by 50 % or more by changes in the soil water content. The estimates are robust with respect to the statistical significance of the regression coefficient (p-value), the autocorrelation of the residuals (Durbin Watson test) and the normality of the residuals (Jarque Bera test).

Prediction power of each model is tested through the leave one out cross validation approach. Results of the absolute mean quadratic error and relative to the standard deviation are resumed in Table 5.

The values of the mean squared errors -RMSE- are obtained from the estimation of the yield index for each year by means of leave one out cross validation technique. Small values -tending to cero- indicate a better performance. In the case of the root mean square error over the standard deviation indicator -RMSE/ST- represents the same as the latter one but relative to the variability of the series: the closer to unity indicates that the error converges to the standard deviation of the yield index series. Finally, for 1-(RMSE/ST) the concept is the same although in this case values close to zero would be more consistent and would indicate that the forecast is associated to the natural variability of the series.

Results show that that forecast error average is mainly explained by structural variability of the model and the results are consistent with the goodness-of-fit of the estimated models, i.e., the higher the coefficient of determination of the model the lower the forecast error. Therefore, the model is robust enough to be use as future estimate of soybean yields, given the value of soil water content.

Application of the index during the 2020–2021 soybean campaign

As was mentioned in the introduction this work is part of a project that aims to help the ORA to manage the climate-driven agricultural risk through a data-driven approach. As an example, we briefly describe how the model developed plays a role in the communication of the expected production of the 2020–2021 soybean campaign. Soybean was sown in October 2020 and harvested during May 2021. At different stages of the campaign, different reports were issued with different estimation of the expected production. At the very beginning of the campaign in November 2020, output was forecast by means of seasonal probabilistic rainfall forecast of the DIVAR group (Osman et al., 2021), determining the cumulative probabilities of the distribution of the soybean production divided into terciles (Thomasz et al., 2020). This information is distributed in a report and provides the first estimation of the expected production. At this stage, this information is critical to decide coverage taking (insurance of other hedging tools) by local producers. Since this model is based on precipitation data only, is less accurate than the one introduced in this work. At the beginning of March, with the information of soil water content up to February 2021, the model developed was run and the expected production of the campaign was estimated and reported three months prior to the end of the campaign trough a market-oriented report (Thomasz et al., 2021), distributed by Refinitiv (former Thompson Reuters). This information is used by the ORA for the evaluation of the agricultural risk associated to climate variability but is also of interest for other areas of government and the financial market, given the incidence over foreign exchange, public revenues, and sovereign risk. Soybean production cycle, modelling and applications are resumed in Table 6.

Conclusion

This paper focused on analyzing the relationship between soil water content and soybean yields in 28 agricultural counties with a territorial

Table 4
Estimation of regression models. (**) and (*) denotes parameters significant at the 1% (5%) level.

Province	County	R^2	Slope	Intercept	DW	JB
BA	9 de Julio	0.531	0.0062**	0.8097**	1.789	1.057
BA	Azul	0.627	0.0100**	0.5823**	2.245	0.704
BA	Bolívar	0.597	0.0064**	0.8415**	2.462	1.393
ER	Concordia	0.594	0.0050**	0.7560**	1.161	1.219
CO	Córdoba	0.530	0.0063**	0.7961**	1.927	0.100
BA	Coronel Suárez	0.585	0.0103**	0.8488**	2.041	0.978
BA	Dolores	0.703	0.0124**	0.1895	1.726	0.806
SF	General Obligado	0.646	0.0133**	−0.0541	2.558	1.572
BA	General Pueyrredón	0.508	0.0087**	0.3885*	2.308	0.678
ER	Gualeguaychú	0.400	0.0053**	0.7441**	2.235	3.207
BA	Junín	0.442	0.0088**	0.2857	2.384	1.448
LP	La Capital	0.379	0.0160**	0.9239**	1.692	0.444
BA	Las Flores	0.595	0.0087**	0.6190**	1.376	0.828
LP	Maracó	0.033	0.0035	0.9894**	0.924	10.468
CO	Marcos Juárez	0.304	0.0028*	0.8379**	2.278	0.989
BA	Olavarría	0.462	0.0068**	0.7612**	2.104	2.239
ER	Paraná	0.585	0.0058**	0.7590**	0.940	1.172
BA	Pehuajó	0.381	0.0080**	0.8805**	1.293	0.257
BA	Roque Saenz Peña	0.681	0.0065**	0.8446**	1.397	2.366
CO	Río Cuarto	0.490	0.0068**	0.7894**	2.156	1.452
CO	Río Seco	0.215	0.0097*	0.6539*	0.709	1.274
CO	Río Segundo	0.231	0.0068*	0.8131**	0.827	0.644
SF	Rosario	0.470	0.0046**	0.7432**	2.248	0.178
BA	Saavedra	0.561	0.0091**	0.9347**	1.459	0.917
SF	San Cristobal	0.644	0.0057**	0.6579**	2.254	0.444
BA	Tandil	0.018	−0.0177	2.1878	2.127	138.025
BA	Trenque Lauquen	0.338	0.0084**	0.8840**	1.371	0.766
BA	Tres Arroyos	0.619	0.0068**	0.7921**	1.664	0.483

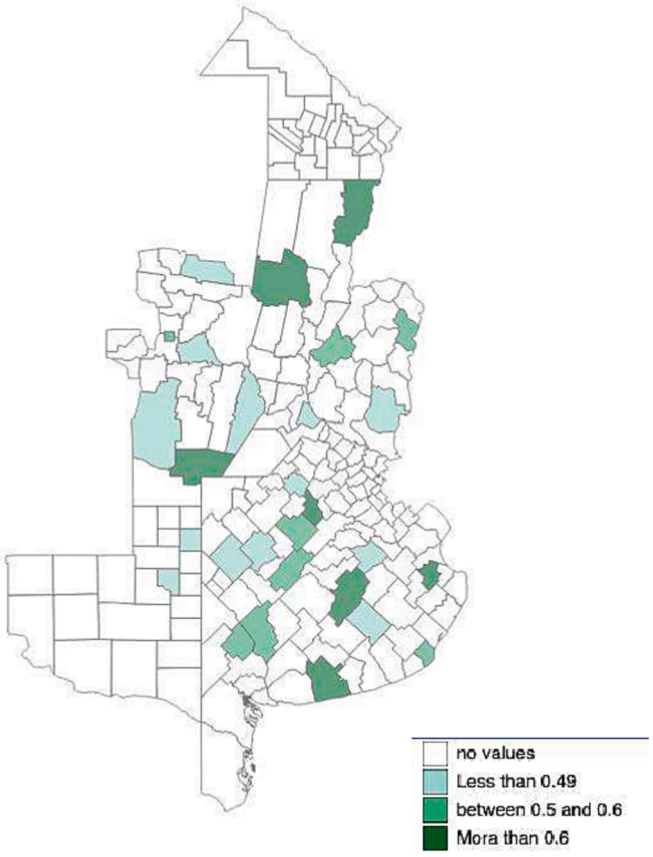


Chart 6. R^2 values per county.

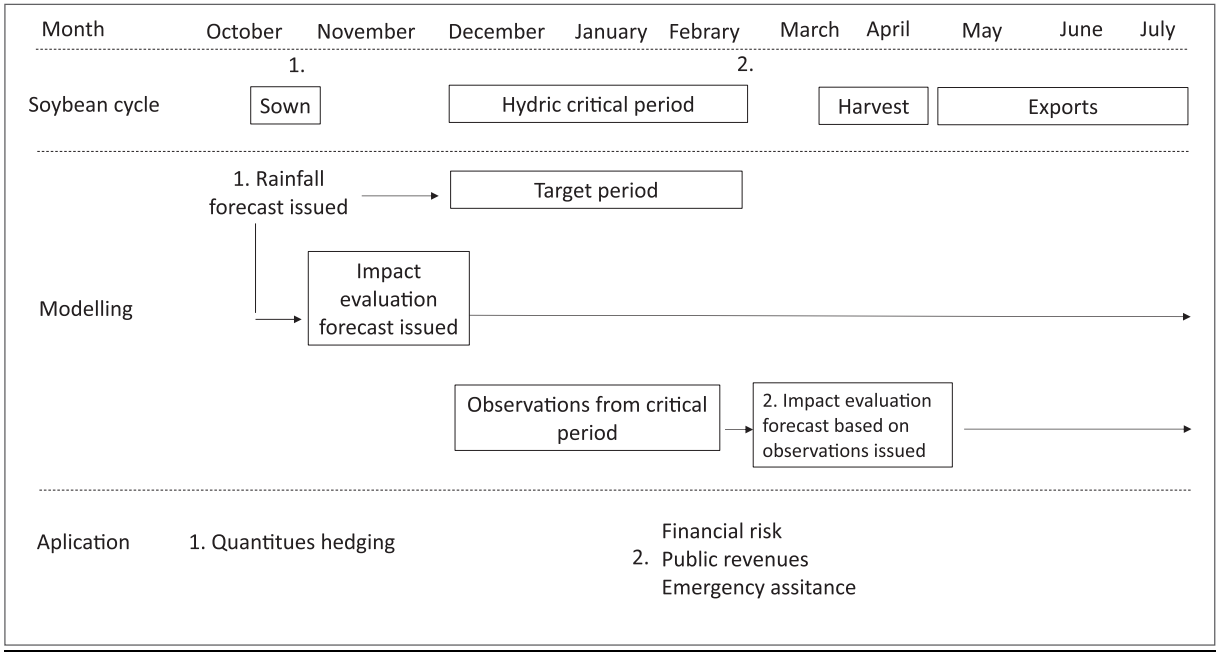
Table 5
Results of the indicators obtained from the application of the leave one out cross validation method.

Province	County	Root Mean Square Error –RMSE (8)-	Root Mean Square Error Standard deviation Ratio- RMSE/ST (8)-	1-(RMSE/ST)
BA	9 de Julio	0.1520	0.7957	0.2043
BA	Azul	0.1402	0.6526	0.3474
BA	Bolívar	0.1273	0.6813	0.3187
ER	Concordia	0.1645	0.7341	0.2659
CO	Córdoba	0.1418	0.7346	0.2654
BA	Coronel Suárez	0.2424	0.7151	0.2849
BA	Dolores	0.1773	0.6778	0.3222
SF	General Obligado	0.3043	0.7488	0.2512
BA	General Pueyrredón	0.1379	0.8061	0.1939
ER	Gualeguaychú	0.1898	0.8275	0.1725
BA	Junín	0.1458	0.9513	0.0487
LP	La Capital	0.4261	0.8641	0.1359
BA	Las Flores	0.2573	0.9104	0.0896
CO	Marcos Juárez	0.1051	0.9176	0.0824
BA	Olavarría	0.1604	0.8241	0.1759
ER	Paraná	0.1707	0.7213	0.2787
BA	Pehuajó	0.2519	0.8781	0.1219
BA	Pres. Roque Saenz Peña	0.1336	0.6080	0.3920
CO	Río Cuarto	0.1656	0.7568	0.2432
CO	Río Seco	0.3976	0.9749	0.0251
CO	Río Segundo	0.2760	0.9660	0.0340
SF	Rosario	0.1294	0.7996	0.2004
BA	Saavedra	0.2478	0.7088	0.2912
SF	San Cristobal	0.1425	0.6453	0.3547
BA	Trenque Lauquen	0.2752	0.8868	0.1132
BA	Tres Arroyos	0.1453	0.6702	0.3298

based weather station with enough data to perform correlation, regression and forecast analysis.

First, the correlation between the first soybean yield index of each county and the soil water content information was studied. This relationship allowed to establish the optimal window with the highest correlation of all the combinations of soil water content data available

Table 6
Schematic of the soybean production cycle, modelling and application stages during a typical soybean campaign (October-May).



during December and April, which represents the soybean growing and harvest period. Results show that January and February play a central role in the relation between water and yields, which is consistent with crop phenology.

With the soil water content window of highest correlation, linear regressions were estimated by county. The results showed that goodness-of-fit is within the range of 70 % to 64 %, and between 63 % and 50 %. Sixty percent of the analyzed counties show these results, while 8 counties −32 %- have goodness-of-fit between 49 % and 35 %. The rest of the analyzed universe values are below 35 %. In sum, in most cases changes in the yield index are explained around 50 % or more by changes in the soil water content.

The main conclusion of the regression analysis is that in the major number of cases soil water content explains at least 50 % percent of the variability in soybean yields, with a maximum of 70 % explanatory power in one county.

Finally, forecast capacity was tested through leave one out validation technique. Results show that models are robust enough to provide a one period forecast, as error is mostly explained by the standard deviation of the yield index: the results are consistent with the goodness-of-fit of the estimated models, i.e., the higher the coefficient of determination of the model the lower the forecast error.

At least two potential applications can be derived from the results. First, the explanatory power and forecast can be used to estimate economic impact of the ongoing campaign. We showed how this information was distributed in the soybean campaign of 2020–2021, providing an estimation of the expected soybean production three months before the end of the campaign, which is useful for emergency agricultural assistance and anticipating impacts on macro and local public finances. Second, the model could be used to design index-based insurance. Having an objective and comprehensive indicator calculated with certified data with enough history, open source and traceability over time complies with most expected requisites for index-based coverage. Despite having other effects that are not explained by the model (such as plagues, cropping strategies and other climate events not captured by the index such as hail and early frost) the explanatory power of at least 50 % is achieved by only one variable and the robustness and statistical properties of the estimated models provide important insights to perform thresholds analysis. In the future, we plan to introduce forecast soil water content data, instead of the observed one, into the model to generate a reliable forecast of soybean production up to 5 months prior the end of the campaign.

Therefore, derived lines of research are related to threshold analysis, evaluating the relation of yields and soil water content during extreme water deficit situations, and the use of the estimated models to evaluate past, present and future economic impact of soil water content variability on soybean production.

This research can be useful to make yield forecasts and can be used as an input for an economic and financial analysis for an investment in adaptation infrastructure.

CRediT authorship contribution statement

Esteban Otto Thomasz: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Kevin Corfield:** Data curation, Formal analysis, Investigation, Methodology, Software, Writing – original draft. **Ana Silvia Vilker:** Formal analysis, Investigation, Methodology, Writing – original draft. **Marisol Osman:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Project administration, Supervision, Writing – original draft.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The research was supported by University of Aires, Argentina (Grant:

PDE 46_2019) and Agencia Nacional de Promocion de la Investigacion, el Desarrollo Tecnologico y la Innovacion, Argentina (Grant: PICT 03537_2018)

Annex 1

Table 1
Correlations of precipitation, SPEI and soil water content with soybean yield index.

	Province	County	Precipitation	SPEI	RH
BA	9 De Julio	45 %	68 %	77 %	
BA	Azul	47 %	65 %	75 %	
BA	Bolivar	59 %	15 %	76 %	
BA	Dolores	80 %	77 %	85 %	
BA	General Pueyrredon	67 %	75 %	76 %	
BA	Las Flores	66 %	83 %	86 %	
BA	Olavarria	51 %	61 %	64 %	
BA	Pehuajo	74 %	85 %	70 %	
BA	Saavedra	72 %	33 %	71 %	
BA	Tres Arroyos	73 %	75 %	63 %	
CO	Marcos Juarez	37 %	50 %	50 %	
CO	Presidente Roque Saenz Pena	70 %	75 %	76 %	
CO	Río Cuarto	65 %	73 %	62 %	
CO	Río Segundo	60 %	67 %	57 %	
ER	Concordia	64 %	63 %	78 %	
ER	Gualeguaychu	32 %	50 %	47 %	
ER	Paraná	63 %	76 %	77 %	
LP	Capital	26 %	30 %	52 %	
LP	Maracó	55 %	58 %	45 %	
SF	General Obligado	60 %	75 %	74 %	
SF	Rosario	44 %	60 %	65 %	
SF	San Cristobal	52 %	67 %	81 %	

The analyzed sample shows that in 14 cases there has been improvement of correlation when using soil water content and 3 cases with values similar to the SPEI. Only 5 cases the SPEI provides better correlation than either IR or precipitation.

Annex 2

Table 1
Start and end date of windows.

30 days window	40 days window	50 days window	60 days window
1 Dec – 31 Dec	1 Dec – 10 Jan	1 Dec – 20 Jan	1 Dec – 31 Jan
10 Dec – 10 Jan	11 Dec – 20 Jan	11 Dec – 31 Jan	10 Dec – 10 Feb
20 Dec – 20 Jan	21 Dec – 31 Jan	21 Dec – 10 Feb	20 Dec – 20 Feb
1 Jan – 31 Jan	1 Jan – 10 Feb	1 Jan – 20 Feb	1 Jan – 28 Feb
10 Jan – 1o Feb	11 Jan – 20 Feb	10 Jan – 28 Feb	10 Jan – 10 Mar
20 Jan – 20 Feb	21 Jan – 28 Feb	20 Jan – 10 Mar	20 Jan – 20 Mar
1 Feb – 28 Feb	1 Feb – 10 Mar	1 Feb – 20 Mar	1 Feb – 31 Mar
10 Feb – 10 Mar	11 Feb – 20 Mar	10 Feb – 31 Mar	10 Feb – 10 Apr
20 Feb – 20 Mar	21 Feb – 31 Mar	20 Feb – 10 Apr	20 Feb – 20 Apr
1 Mar – 30 Mar	1 Mar – 10 Apr	1 Mar – 20 Apr	1 Mar – 30 Apr
10 Mar – 10 Apr	11 Mar – 20 Apr	10 Mar – 30 Apr	
20 Mar – 20 Apr	21 Mar – 30 Apr		
1 Apr – 30 Apr			

Table 2
Number of windows.

Length of window*	**30 days**	**40 days**	**50 days**	**60 days**
Number of moving windows	13	12	11	10

*All windows begin on December 1 and end on April 30.

References

Barros, V.R., Boninsegna, J.A., Camilloni, I.A., Chidiak, M., Magrín, G.O.y., Rusticucci, M., 2015. Climate change in Argentina: trends, projections, impacts and adaptation. WIREs Clim. Change 6, 151–169. https://doi.org/10.1002/wcc.316.

Barros, V., Vera, C., Agosta, E., Araneo, D., Camilloni, I., Carril, A.F., Doyle, M.E., Frumento, O., Nuñez, M., Ortiz de Zárate, M.I., Penalba, O., Rusticucci, M., Saulo, C., Solman, S., 2014. Tercera Comunicación Nacional Sobre Cambio Climático. Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina.

Basualdo, A., 2020. Balance hídrico diario para cultivos específicos. Oficina de Riesgo Agropecuario, República Argentina. http://www.ora.gob.ar/informes/Reservas_de_Agua_Metodologia_balance.pdf.

Bert, F., Satorre, E., Ruiz Toranzo, F., Podestá, G., 2006. Climatic information and decision-making in maize crop production systems of the Argentinean Pampas. Agric. Syst. 88 (2–3), 180–204. https://doi.org/10.1016/j.agsy.2005.03.007.

CEPAL, 2014. La economía del cambio climático en la Argentina. Primera aproximación. Impreso en Naciones Unidas. Santiago de Chile. Recuperado d http://www.cepal.org/es/publicaciones/35901-la-economia-del-cambioclimatico-en-la-argentina-primera-aproximacion.

Chimeli, A.B., De Souza Filho, F.D.A., Holanda, M.C., Petterini, F.C., 2008. Forecasting the impacts of climate variability: lessons from the 97 rainfed corn market in Ceará, Brazil. Environ. Dev. Econ. 13 (02), 201–227. https://doi.org/10.1017/S1355770X07004172.

Food and Agriculture Organization (FAO) (1978). Effective rainfall in irrigated agriculture. Chapter 3, Section 3: Potential Evapotranspiration/Precipitation Ratio Method (India). M-56ISBN 92-5-100272-X. Available at http://www.fao.org/3/x5560e/x5560e00.htm#Contents.

Forte Lay, J. A. y Burgos J. J., 1983. "Verificación de métodos de estimación de la variación del almacenaje de agua en suelos pampeanos". Actas del Taller Argentino-Estadounidense sobre sequías (CONICET-NSF), realizado en Mar del Plata entre el 4 y el 8 de Diciembre de 1978. Editor J. J. Burgos. Buenos Aires, Argentina. Pág. 162-180. Noviembre de 1983.

Hastie, T., Tibshirani, R., Friedman, J., 2009. Model Assessment and Selection. In: The Elements of Statical Learning: Data Mining, Inference and Prediction, 2nd ed. Springer, New York, Springer Series in Statistics, NY, pp. 219–257.

Irwin, S., & Good, D., 2015. Forming Expectations for the 2015 US Average Soybean Yield: What Does History Teach Us?. Farmdoc daily (5): 51. Department of Agricultural and Consumer Economics, University of Illinois at Urbana-Champaign.

Letson, D., Laciana, C., Bert, F., Weber, E., Katz, R., Gonzalez, X., Podestá, G., 2009. Value of perfect ENSO phase predictions for agriculture: evaluating the impact of land tenure and decision objectives. Clim. Change 97 (1–2), 145–170.

Lobell, D.B., Burke, M.B., 2010. On the use of statistical models to predict crop yield responses to climate change. Agric. Forest Meteorol. 150 (11), 1443–1452. https://doi.org/10.1016/j.agrformet.2010.07.008.

Magrin, G., Gay García, C., Cruz Choque, D., Giménez, J.C., Moreno, A.R., Nagy, G.J., Nobre, C., Villamizar, A., 2007. Latin America. Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK, pp. 581–615.

Massot, J., Baez, G., Prieto, K., Petri, G., Argüero, L., Thomasz, E., Gayá, R., Fusco, M., 2016. Agroindustria, innovación y crecimiento económico en la Argentina. Ed.: EDICON. In Spanish.

Murgida, A. M., Travasso, M. I., González, S. y, Rodríguez, G. R., 2014. Evaluación de impactos del cambio climático sobre la producción agrícola en la Argentina. Serie medio ambiente y desarrollo. No. 155. Naciones Unidas. Santiago de Chile, Chile.

Ortiz de Zarate, M. J., Ramayon, J. J. y Rolla, A. L., 2014. Agricultura y Ganadería impacto y vulnerabilidad al cambio climático. Posibles medidas de adaptación. 3era comunicación nacional de la República Argentina a la Convención Marco de las Naciones Unidas sobre cambio climático.

Osman, M., Coelho, C.A.S., Vera, C.S., 2021. Calibration and combination of seasonal precipitation forecasts over South America using Ensemble Regression. Clim. Dyn. 57 (9-10), 2889–2904.

Paltasingh, K.R., Goyari, P., Mishra, R.K., 2012. Measuring weather impact on crop yield using aridity index: Evidence En: Odisha. Agric. Econ. Res. Rev. 25 (2), 205–216.

Rahman, M., Huq, M., Sumi, A., Mostafa, M., Azad, M., 2005. Statistical Analysis of Crop-Weather Regression Model for Forecasting Production Impact of Aus Rice in Bangladesh. Int. J. Statistical Sci. 4, 57–77.

Serrano, V., Beguería, S., López-Moreno, J.I., 2010. A Multi-scalar drought index sensitive to global warming: The Standardized Precipitation Evapotranspiration Index - SPEI. J. Clim. 23, 1696–1718.

Tannura, M.A., Irwin, S.H., Good, D.L., 2008. "Weather, Technology, and Corn and Soybean Yields in the U.S. Corn Belt." Marketing and Outlook Research Report 2008-01, Department of Agricultural and Consumer Economics, University of Illinois at Urbana-Champaign.

Thomasz, E., Casparri, M., 2015. Innovaciones Financieras para Adaptación al Riesgo Climático: el caso de las Coberturas basadas en Índices. Documento de Trabajo del Centro de Investigación en Métodos Cuantitativos Aplicados a la Economía y la Gestión (CMA), Facultad de Ciencias Económicas, Universidad de Buenos Aires. Recovered at://www.economicas.uba.ar/institutos_y_centros/provul/.

Thomasz, E., Massot, J., Rondinone, G., 2016. Is the interest rate more important than stocks? The case of agricultural commodities in the context of the financialization process. Revista Lecturas de Economía, N 85, Universidad de Antioquia. doi: 10.17533/udea.le.n85a04.

Thomasz, E., Vilker, A., & Rondinone, G., 2017. The economic cost of extreme and severe droughts in soybean production in Argentina. Recovered at: https://www.cya.unam.mx/index.php/cya.
Thomasz, E., Eriz M., Vilker, A., Rondinone, G., Corfield K., 2020. Proyecciones soja campaña 2020/21. Reporte Provul 3/2020. Recovered at://www.economicas.uba.ar/institutos_y_centros/provul/.
Thomasz, E., Eriz M., Vilker, A., Rondinone, G., Corfield, K., 2021. Resultados PDE: Proyección y monitoreo campaña soja 2021. Reporte Provul 1/2021. Recovered at: http://www.economicas.uba.ar/institutos_y_centros/provul/.

Web References

Australian Bureau of Meteorology: http://www.bom.gov.au/climate/drought/#tabs=Soil-moisture.
Climate prediction center: https://www.cpc.ncep.noaa.gov/products/Drought/.
European Drought Observatory: https://edo.jrc.ec.europa.eu/edov2/php/index.php?id=1000.
Open data source of agricultural data: https://www.agroindustria.gob.ar/datosabiertos/.
SISSA: https://sissa.crc-sas.org/monitoreo/estado-actual-de-la-sequia/.