Forecasting crop yields through climate variables using mixed frequency data. The case of Argentine soybeans | Magdalena Cornejo | Revista Económica La Plata, Vol. LXVII, Nro. 1 | Enero-Diciembre 2021 | ISSN 1852-1649 | https://doi.org/10.24215/18521649e022 | https://revistas.unlp.edu.ar/economica | Facultad de Ciencias Económicas, Universidad Nacional de La Plata | La Plata | Buenos Aires | Argentina

REVISTA
ECONÓMICA
LA PLATA

Forecasting crop yields through climate variables using mixed frequency data. The case of Argentine soybeans

Pronósticos del rendimiento de los cultivos a través de variables climáticas usando datos en frecuencias mixtas. El caso argentino

Magdalena Cornejo

ABSTRACT

This article evaluates the value of information on climate variables published in advance and at a higher frequency than the target variable of interest—crop yields—in order to get short term forecasts. Aggregate and disaggregate climate data, alternative weighting schemes and different updating schemes are used to evaluate forecasting performance. This study focuses on the case of soybean yields in Argentina. Results show that models including high frequency weather data outperformed particularly during the three consecutive campaigns after 2008/09 when soybean yield decreased almost by 50%. Furthermore, forecast combinations showed a better forecasting performance than individual forecasting models.

Keywords: yields, forecasting, climate, mixed-frequency, soybeans.

RESUMEN

Este artículo evalúa el valor de utilizar información sobre variables climáticas publicadas con anticipación y con una frecuencia superior a la variable objetivo de interés —los rendimientos de los cultivos— con el fin de obtener pronósticos a corto plazo. Se utilizan datos climáticos agregados y desagregados, esquemas de ponderación alternativos y diferentes esquemas de actualización para evaluar el desempeño de las predicciones. Este estudio se centra en el caso de los rendimientos de la soja en Argentina. Los resultados muestran que los modelos que incluyen datos meteorológicos de alta frecuencia obtuvieron mejores resultados, particularmente durante las tres campañas consecutivas después de 2008/09, cuando el rendimiento de la soja disminuyó en casi un 50%. A su vez, las combinaciones de pronóstico mostraron un mejor desempeño que los modelos de pronóstico individuales.

Palabras claves: rendimientos, predicciones, clima, frecuencia mixta, soja.

Recibido: 08/10/2020. Aceptado: 25/06/2021
Clasificación JEL: C53, Q10

Magdalena Cornejo: Escuela de Gobierno, Universidad Torcuato Di Tella, Argentina y Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina. e-mail: mcornejo@utdt.edu.

I. INTRODUCTION

Farmers, government and traders' decisions would benefit from obtaining more accurate crop yield forecasts using short horizons. For instance, harvesting decisions are often based on incomplete information on crop yields which depend on current agricultural and meteorological conditions, particularly during the last part of the plant growth cycle. However, one important limitation to forecast crop yields is that the final figures are usually available on a national level and an annual basis only. Instead, a great deal of weather information, which is paramount to agricultural production, is published on a more frequent basis than agricultural information. Weather information, which includes temperature and precipitation data as the most relevant measures, may help improve the forecast accuracy of crop yields in the short term by mixing these higher frequency data with lower frequency data on crop yields.

Different factors may explain crop yield in the long run, e.g. the global and local effects of climate change or technological innovations such as the use of modified seeds, fertilizers or changes in management practices that can boost crop yields. These long-run effects are analyzed for the case of soybean yields in Argentina in Marcellino (2002). Notwithstanding, climate variables are expected to be a crucial determinant of yield variation in the short run. This article, in line with a growing body of literature, focuses on the contribution of climate variables to forecast crop yields.

In particular, precipitation and temperature have been recognized as major climate factors affecting crop growth on national and global scales (Lobell and Burke, 2009; Schlenker and Roberts, 2009). Water stress during the flowering stage of the plant may affect seed weight, resulting in large seed weight variations. Nonetheless, many recent studies indicate that changes in temperature are more important than changes in rainfall, at least at the national and regional levels (Reilly and Schimmelpfennig, 2000; Schlenker and Lobell, 2010). Furthermore, crops are more sensitive to extremely high temperatures, in particular, during the plant growth cycle.

Thus, the primary goal in our current study is to assess the value of information on (climate) variables published in advance and at a higher frequency than the target interest variable (crop yields) in order to get short-term forecasts. Different approaches are considered using aggregate and disaggregate climate data, as well as alternative weighting schemes—simple averages or Mixed Data Sampling (MIDAS) regressions—and ways of updating—rolling estimations—to evaluate their forecasting performance. We focus on the case of soybean yields in Argentina, the third worldwide producer and exporter.

The article is organized as follows. The next section reviews the literature and presents the different approaches used to forecast crop yields. Section 3 describes the data and forecast design. Section 4 analyzes the individual forecasting performance, while Section 5 evaluates forecast combinations. Finally, Section 6 presents our conclusions.

II. FORECASTING CROP YIELDS USING CLIMATE DATA

Crop yield forecasting has been a matter of global concern given the need to increase food and renewable energy production to cope with a rapid global population growth. In particular, global warming awareness has sparked renewed interest in studying this topic over the last decade.

As Lobell and Burke (2010) state, a common approach is to use statistical models trained on historical yields and some simplified weather measurements, such as growing season average temperatures and precipitations. According to Ray, Gerber, MacDonald, and West (2015), climate-driven temperature variations, precipitation or their interaction explain a third of global crop yield variability.

Different weather measures have been suggested to explain and forecast crop yields. It has been usually considered that the best crop yield predictor is some measure of extreme heat during the plant growth cycle. Using aggregated weather data during the plant's growing season, Schlenker and Roberts (2009) found that extreme high temperatures are always harmful to crop growth. Temperatures above 30°C for soybeans and 29°C for corn are very harmful. They found that crop losses on the hottest days drive much of the tem-

perature effect. Furthermore, many studies indicate that temperature extremes can be critical to reducing yields, especially if they coincide with the flowering stage of the crop (Auffhammer, Ramanathan, and Vincent, 2012; Welch, Vincent, Auffhammer, Moya, Dobermann, and Dawe, 2010; Wheeler, Craufurd, Ellis, Porter, and Prasad, 2000).

Using data from 1980 to 2003, Lobell, Cahill, and Field (2007) analyzed the relationship between 12 major Californian crop yields and monthly temperatures and precipitations before and during growing seasons. However, as they acknowledged, monthly climate variables mask daily extremes that can have significant effects on yield. Nonetheless, disaggregating climate data to finer temporal scales will not necessarily improve model performance. Therefore, dealing with mixed frequency data is considered a key issue in terms of modelling (annual) crop yields based on climate variables disaggregated at a finer temporal scale.

Thus, this article assesses different time-frequency approaches to evaluate if using disaggregate climate data to finer (weekly) temporal scales help to improve crop yield forecasts. All these model, which include climate variables, will be compared to a benchmark that does not rely on them: an AR(1) model as expressed in Equation (1).

$$y_t = \alpha + \phi y_{t-1} + \varepsilon_t \qquad (1)$$

where represents the crop yield variations, measured as the log-difference of drop yields, which are observed on an annual basis.

Comparing the performance of alternative forecasting models based on weather information with respect to this AR(1) benchmark will allow us to evaluate the relative forecast improvements from introducing climate variables.

There are different ways of modelling with mixed-frequency data. This article evaluates different weighting schemes of high-frequency information (from simple averages to MIDAS regressions).

In the first approach, as usually done in the empirical literature, climate variables are aggregated to yearly values (using information from an year prior to the harvesting period) by taking means or through the sum of all values in the case of cumulative variables (such as rainfall) and they are added to the AR(1) model as shown in Equation (2).

$$y_t = \alpha + \phi y_{t-1} + x_t'\beta + \varepsilon_t \qquad (2)$$

where includes a set of climate variables such as maximum temperature, number of days with maximum temperature above a threshold of 29°C, 30°C or 31°C, cumulative rainfall, and number of days without rain. Given that most regressors have a strong correlation and that the in-sample period has only 30 observations (years), those climate variables will be individually included in the model and also, their joint evaluation will be assessed by using LASSO (Least Absolute Shrinkage and Selection Operator). LASSO (Tibshirani, 1996) is a shrinkage and selection method for linear regression that fits a linear model such as in Equation (2) by least squares with a bound on the sum of the absolute values of the coefficients. Such restriction is multiplied by a parameter λ that slows or accelerated the penalty. In this study, λ selected on a rolling basis using a block form of -fold cross-validation (CV) given that we are working with time series data.

Using yearly weather information may be useful to predict crop yields variations as the soil conditions (prior to the beginning of the planting season) are usually affected by extreme weather events such as high cumulative rainfall that will condition the plant's growth and yield. However, this low-frequency models regress crop yield variations on climate variables on an annual basis (t), without exploiting high-frequency variations that may help to improve the model's forecast performance.

Secondly, in order to evaluate if forecasting models based on recent climate data outperform those obtained by using the full year data, the climate information is restricted to the last five or seven weeks before the start of the harvest season.

$$y_t = \alpha + \phi y_{t-1} + \tilde{x}_t'\beta + \varepsilon_t \quad (3)$$

where includes a set of climate variables that restrict the information to the last five or seven weeks before the harvest begins (e.g. the maximum temperature in the last months prior to the harvest). Once again, the climate variables will be individually included in the model and jointly evaluated by LASSO.

Therefore, aggregating climate data restricted to more recent information will allow us to assess if short-run (Equation 3) or long-run (Equation 2) weather information helps improve our forecasts. Moreover, relevant historical climate information may differ according to the weather measure considered.

Finally, a mixed-frequency approach is also followed, the so-called MIDAS (mixed-data sampling) approach. The MIDAS regression models, developed by Ghysels, Santa-Clara, and Valkanov (2004, 2006), economize on the number of parameters to be estimated by fitting a flexibly and parsimoniously parametrized lag polynomial of the response of the lower frequency dependent variable (here, crop yield variations) to the higher frequency data (here, climate variables). Weekly climate indicators are earlier available than the final figures of yields. Thus, weekly climate indicator is typically available within the year for which no yield figure is available. A simple MIDAS model is:

$$y_t = \beta_0 + \beta_1 B\left(L^{1/m};\theta\right)x_t^m + \varepsilon_t \quad (4)$$

where y_t represents the crop yield variations which is observed once between t-1 and t (yearly) and x_t^m the climate variables observed m times in the same period (i.e. weekly or m=52). $B\left(L^{1/m};\theta\right) = \sum_{k=0}^{K} B(k;\theta)L^{k/m}$ and $L^{(1/m)}$ is a lag operator such that $L^{1/m}x_t^m = x_{t-1/m}^m$; and the lag coefficients in $B(k;\theta)$ of the corresponding lag operator are parameterized as a function of small-dimensional vector of parameters .

Different polynomial parametrizations have been suggested and employed in empirical work. In this article, an Exponential Almon Lag, as expressed in Equation (5), is used as the weighting function of lag values.[1]

$$B(k;\theta) = \frac{e^{\theta_1 k + \cdots + \theta_Q k^Q}}{\sum_{k=1}^{K} e^{\theta_1 k + \cdots + \theta_Q k^Q}} \quad (5)$$

The function is known to be quite flexible and can take various shapes with only a few parameters, that is, even if it is restricted to a two-parameter case, $\nu = \log(Y_t)$ (see Ghysels et al., 2004, for further explanation on the flexibility of the Exponential Almon Lag). However, as stated by Jansen, Jin, and de Winter (2016), the efficiency gains of this approach come at the cost of potential efficiency losses, if the implied restriction on the lag dynamics between the weekly indicators and yearly crop yield variations happen to be invalid.

[1] A Beta weighting function was also used, but it had a lower forecasting performance than the Exponential Almon lag. Although not reported, results are available upon request.

The optimal exponential Almon lag structure of the MIDAS model was first selected for each climate regressor in terms of the selected information criteria (Akaike and Schwarz criteria) over the in-sample period. Results showed that the optimal lag length to be considered is 7 weeks for the maximum temperature and cumulative rainfall, and 5 weeks for the number of days without rain and the number of days with different maximum temperature thresholds (29°C, 30°C and 31°C).

Furthermore, to assess if the weighting scheme (the exponential Almon lag polynomial) in the MIDAS representation was more accurate than the low-frequency data estimations, their forecasting performance was compared with respect to each climate regressor's simple average using the same optimal lag length. If the MIDAS representation outperforms the later, it will imply that a non-linear weighting scheme of climate variables helps forecast crop yields.

The next section describes the data and introduces the forecast design.

III. DATA AND FORECAST DESIGN

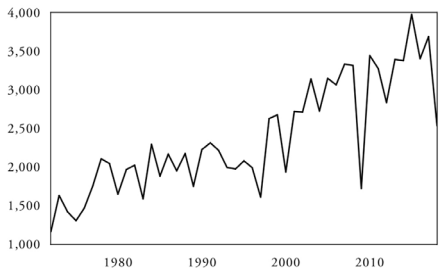
III.1 THE CASE OF SOYBEANS IN ARGENTINA

Argentina is the world's third largest producer and exporter of soybeans and the world's top exporter of soymeal and soybean oil. Soybean was almost an unknown crop in the agricultural landscape in the early 1970s and it has rapidly gained in popularity mainly due to the increased global demand for food, animal feed and biofuel.

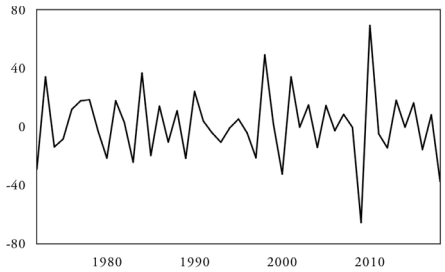
This study focuses on the nucleus or central zone of soybean production in Argentina (northern part of Buenos Aires, southern part of Santa Fe, eastern part of Córdoba and western part of Entre Ríos). This area comprises 34 counties ("departamentos" or "partidos") that account for 42% of the national production and only 38% of the total harvest area over 1971/72 to 2017/18. During this period, the nucleus zone has had average soybean yields above 25% from the national average, with an annual increase of 4%.[1]

Figure 1: Soybean yields (level and growth)

Panel A: Soybean yield (kg/ha)



Panel B: Soybean yield annual growth (%)



Note: date label indicates the year in which the campaign ends.
Source: Dirección Nacional de Estimaciones Agrícolas.

Despite the observed long-run trend, soybean yield growth has shown large variations within this period as shown in figure 1.

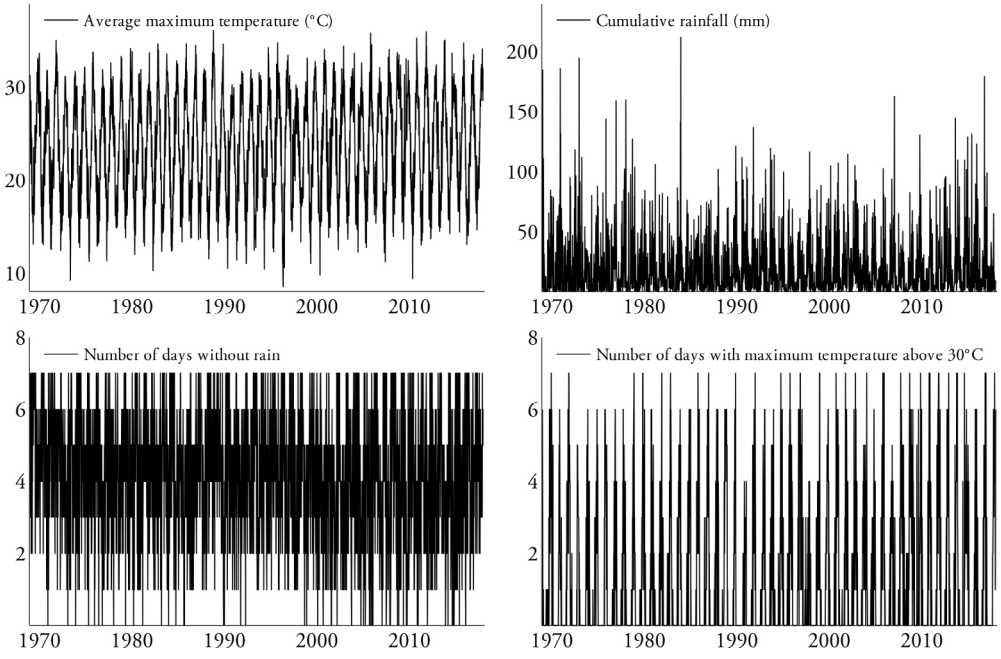
[2] The 34 counties considered are: 20 from the north of Buenos Aires province (Alberti, Baradero, Bragado, Campana, Chacabuco, Chivilcoy, General Arenales, Junín, Leandro N. Alem, Luján, Mercedes, Pergamino, Pilar, Ramallo, Rojas, Salto, San Nicolás, San Pedro, Suipacha and Zárate), 2 from the east of Córdoba province (Marcos Juárez and Unión), 3 from the west of Entre Ríos province (Colón, Diamante and Victoria) and 9 from the center and south of Santa Fe province (Belgrano, Caseros, Constitución, General López, Iriondo, Rosario, San Jerónimo, San Lorenzo and San Martín).

III.2 DATASET

In order to obtain short-term forecasts of annual soybean yield variations, our study is based on the sample period ranging from 1971/72 to 2017/18 (47 annual observations). In Argentina, the soybean campaign typically starts in October and finishes in May or June, when the harvest is finished. At the national and county scale, soybean yields are published by the Dirección Nacional de Estimaciones Agrícolas. The historical crop yield figures considered in this study were released on November 27th, 2018.

Weather data used in this study include: cumulative rainfall, average maximum temperature, number of days without rain, and number of days with maximum temperatures above 29°C, 30°C and 31°C. The different threshold temperature measures showed similar results, therefore only the forecasting performance of the number of days with maximum temperature above 30°C is reported. Even if these climate variables are released by six meteorological stations located in the nucleus zone at a daily frequency,[3] their mean weekly values are considered which are weighted by each county share in the total soybean planted nucleus area. Even if weekly climate averages mask daily extremes that can have strong effects on crop yields, daily climate variables reduce the reliability of the dataset due to data inaccuracies (e.g. missing values) that may be cancelled out at broad scales.

Figure 2 shows the historical evolution of the climate variables considered in the forecast exercise.



Sources: National Weather Service (SMN) and National Institute of Agricultural Technology (INTA).

III.3 FORECAST DESIGN

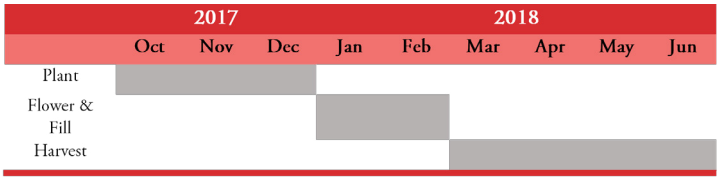
Using 1971/72 – 2000/01 as the estimation sample, a rolling[4] pseudorealtime forecast exercise of soybeans yields was performed over the period 2001/02 to 2017/18, with a rolling 30year window.

[3] The six meteorological stations, which have complete historical weather data, are located in Pergamino, Junín, Marcos Juárez, Rosario, San Pedro and Iriondo. The first four stations belong to the National Weather Service (SMN), whereas the last two are from the National Institute of Agricultural Technology (INTA).

[4] Although not reported, recursive forecasts were also performed, but they never showed a better performance than the rolling forecasts. Results are available upon request.

For each forecast, the origin was at the end of the flowering and filling stages (end of February), and the final figures of soybean yields were forecasted for the corresponding out-of-sample campaign. Figure 3 shows the soybean production calendar for Argentina's nucleus zone for the 2017/18 campaign.

Figure 3: Soybean crop calendar



Source: own elaboration.

As explained in Section 2, the forecasting models were estimated using different frequency data.

Table 1 summarizes the estimated individual forecasting models using different frequency data. All models included an autoregressive term and passed diagnostic tests. Insample estimations for the initial window are reported in the Appendix in Tables A1 and A2.

Given that soybean yields may be represented as stationary around a deterministic linear trend and that all climate variables are found to be stationary,[5] all forecasting models regard the log-difference of yields as the dependent variable. Nevertheless, given that our interest is in forecasting soybean yield levels (measured in kg/ha), the growth rates (calculated as the first difference of the natural logarithm of yields) were converted to median levels in order to evaluate their forecasting performance. That is, median forecasts are easily derived from the inverse transformation as if $y = \log(Y_t)$, then $\hat{Y}_{T+h|T} = \exp(\hat{y}_{T+h|T})$.[6]

Table 1: Alternative forecasting models for soybean yields growth

Model	Frequency	Acronym	Climate regressor
Benchmark	Annual	AR(1)	None
Long-sample estimations	Annual	$maxtemp$	Maximum temperature
		$maxtemp > 29$	Number of days with $maxtemp$ above 29°C
		$maxtemp > 30$	Number of days with $maxtemp$ above 30°C
		$maxtemp > 31$	Number of days with $maxtemp$ above 31°C
		$rainfall$	Cumulative rainfall
		$no - rain$	Number of days without rain
		all	All climate variables selected by LASSO
Short-sample estimations	Weekly	$maxtemp$	Average maximum temperature
		$max29$	Number of days with $maxtemp$ above 29°C
		$max30$	Number of days with $maxtemp$ above 30°C
		$max31$	Number of days with $maxtemp$ above 31°C
		$rainfall$	Cumulative rainfall
		$no - rain$	Number of days without rain
		all	All climate variables selected by LASSO
MIDAS	Weekly	$maxtemp$	Average maximum temperature
		$max29$	Number of days with $maxtemp$ above 29°C
		$max30$	Number of days with $maxtemp$ above 30°C
		$max31$	Number of days with $maxtemp$ above 31°C
		$rainfall$	Cumulative rainfall
		$no - rain$	Number of days without rain

Source: own elaboration.

[5] Although not reported, unit root test results are available upon request.

[6] The median and mean level forecasts are identical if no functional transformation is used.

The individual performance of each model is compared in order to evaluate forecasting gains from using high frequency data (Section 4). Then, those models are compared with different model and forecast combination strategies (Section 5).

IV. INDIVIDUAL FORECAST PERFORMANCE

This section compares the forecasting accuracy of the different estimated models. Results are divided into two subsections: individual forecasting comparisons, and the timevarying forecasting ability.

IV.1 RMSE AND MAPE

In this section, the root mean squared error (RMSE) and the mean absolute percentage error (MAPE) are used to evaluate the quality of point forecasts over the 2002/03 to 2017/18 period, as shown in table 2. The comparison between the individual forecasting models is expected to produce different results using either of those measures. These differences may arise not only because of the normalization of the MAPE, but also because the implicit quadratic loss function of the RMSE measure gives relatively higher weight to large errors than MAPE does.

Table 2: Forecasting performances

Climate regressor	Model	RMSE	MAPE
None	Benchmark: AR(1)	610.34	14.83
$maxtemp$	Long-sample (annual data)	705.59	19.62
	Short.sample (weekly data)	590.69	15.48
	MIDAS	612.17	16.17
$maxtemp > 30$	Long-sample (annual data)	629.58	16.97
	Short.sample (weekly data)	603.06	16.29
	MIDAS	738.80	20.82
$rainfall$	Long-sample (annual data)	636.80	16.73
	Short.sample (weekly data)	707.93	19.48
	MIDAS	685.28	19.81
$no - rain$	Long-sample (annual data)	**521.38***	**14.02**
	Short.sample (weekly data)	572.86	14.21
	MIDAS	801.46	22.28
all (by LASSO)	Long-sample (annual data)	673.35	17.07
	Short.sample (weekly data)	596.53	15.13

Note: ***p_i.01, **p_i.05, and *p_i.10. The best forecasting model is in bold, and the best model specification within each climate regressor is highlighted.
Source: own elaboration.

From table 2, we can note that there are forecasting gains when taking averages of the weekly maximum temperature during the growing cycle of the plant than considering the annual data. This result is also found when using the maximum temperature threshold of 30°C. These results are in line with many studies

that indicate that temperature extremes can be critical to reducing yields, especially if they coincide with the flowering stage of the crop (Auffhammer et al., 2012; Welch et al., 2010; Wheeler et al., 2000). However, when using weather data associated to precipitations, that is, the cumulative rainfall and the number of days without rain , the model based on annual data outperforms the rest. Among all climate regressors, the number of days without rain using annual data shows forecasting gains.

For each climate variable, the predictive accuracy of each model was compared against the rest. Significance is indicated when the forecasting model outperforms the rest. According to Diebold-Mariano tests, average forecasting gains are detected only when using a quadratic loss function for the models based the number of days without rain prior to the harvest.

It is worth noting that these results are valid only on average for the whole out-ofsample period and may differ over time. Focusing solely on the average performance of the model may result in a loss of information and possibly lead to incorrect forecast selection decisions. Therefore, the following subsection conducts a fluctuation test to evaluate the relative forecasting performances of the two models with the best average performance for each variable.

IV.2 RELATIVE FORECASTING ABILITY

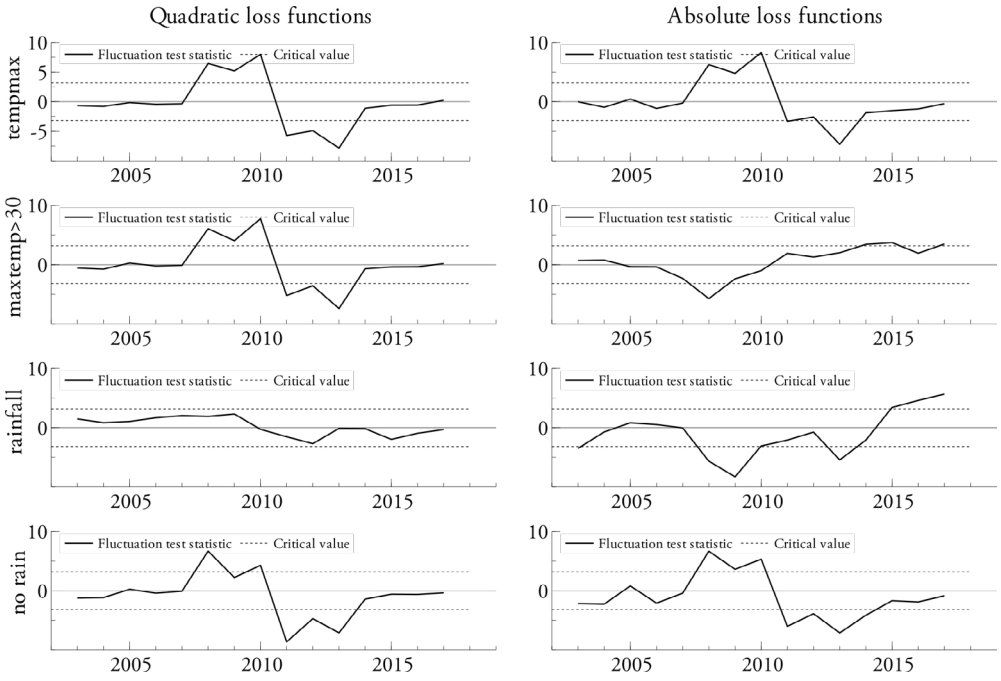
The out-of-sample period was quite unstable, particularly during the 2008/09 campaign when crop production was severely affected by a historical drought. Given the observed instability of the out-of-sample period, it is also worth evaluating the evolution of the different models' relative performances. To assess the time-varying forecasting performance of two competing models, the fluctuation test developed by Giacomini and Rossi (2010) was applied. In our case, we evaluate the local relative forecasting performances of the best forecasting model (within each climate regressor, as highlighted in Table 2) with respect to the benchmark, the AR(1) model.

We defined a (quadratic or absolute) loss function $L(L_{i,t} = e_{i,t}^2$ or $L_{i,t} = |e_{i,t}|)$ for each forecasting model . The out-of-sample relative performance of the models was given by: $\Delta L_{j,t} = L_{j,t} - L_{b,t}$ where corresponds to the best forecasting model within each climate variable and to the benchmark model.

The local relative loss for two models is evaluated as the sequence of out-of-sample loss differences over centered rolling windows of size (in our case,). In this case, a quadratic and an absolute loss functions were considered when comparing the best forecasting model (in terms of their average performance) with respect to the benchmark.

The results of the fluctuation test are shown in figure 4. In each case, the graph reports both the fluctuation test statistic and the two-sided critical value at 5% (the dotted lines). Positive values, above the critical value, indicate that the benchmark forecasting model is better than the forecasting model with the best average performance during the outofsample period, while negative values indicate the opposite.

Figure 4: Fluctuation test results (best model vs benchmark)



Source: own elaboration.

Figure 4 shows that forecast gains are time-varying between each forecasting model and the benchmark. The best forecasting model that accounts for the number of days without rain showed statistically significant differences arise particularly during the three following campaigns after the 2008/09 campaign when soybean yield decreased almost by 50% in the nucleus zone. It is also worth noting that there are no systematic forecasting gains with each model, the benchmark model showed significant gains in some cases priori to the 2009 crisis.

V. FORECAST COMBINATIONS

Forecast combinations are usually found to produce better forecasts than individual models. Moreover, as Timmermann (2006) indicates, simple combinations that ignore correlations between forecast errors often dominate more refined combination schemes aimed at estimating the theoretically optimal combination weights.

Combining forecasts from alternative single models can be beneficial in the presence of misspecification or instabilities (Clark and McCracken, 2010; Hendry and Clements, 2004). Furthermore, as high frequency climate variables are usually highly collinear, forecast combination of several univariate models can also be considered as a potential solution to dealing with multicollinearity issues as suggested by Andreou, Ghysels, and Kourtellos (2013). Moreover, the combination of methods was the king of the recent M4 Competition. According to Makridakis, Spiliotis, and Assimakopoulos (2018), of the 17 most accurate methods, 12 were "combinations" of mostly statistical approaches.

Therefore, table 3 presents the pooling of the best individual models: the long-sample data of the number of days without rain and the short-sample data of the maximum temperature during the plant's flowering stage, using the mean of individual forecasts, and also the inverse weights of the RMSE and MAPE obtained from the forecasts made during the previous campaign. Those forecasts are also compared with a model which combines the regressors used in the individual models *(maxtemp and no rain).* All models include an autoregressive term and the in-sample estimations are presented in the Appendix.

Table 3: Forecast combination performances

Model	Weighted by	RMSE	MAPE
	Mean	533.08	13.42
Pooling	Inv. RMSE	548.59	14.95
	Inv. MAPE	543.30	14.54
maxtemp & no-rain		510.28	14.21

Source: own elaboration.

When pooling the two best individual models, the mean pooling showed the best forecast performance in terms of RMSE and MAPE. However, when combining, in a single model, annual data on the number of days without rainfall and weekly data on the number of days with maximum temperature above 30°C during the plant flowering stage forecasting gains are found in terms of MAPE, but RMSE still indicates that the mean pooling outperforms. According to the DieboldMariano tests, their differences are not statistically significant.

Overall, forecast combinations showed a better forecast performance than their individual counterparts, with a statistically significant difference detected at a 10% when comparing the mean pooling with respect to the individual models.

VI. CONCLUSIONS

This article addressed the usefulness of exploiting high frequency climate information to forecast annual crop yields in the short run. We have focused on soybean yields in the nucleus zone of production in Argentina, the world's third largest producer and exporter.

Different approaches were considered using aggregate and disaggregate climate data as well as alternative weighting schemes (simple averages or MIDAS regressions) to evaluate their forecasting performance. Climate data included cumulative rainfall, average maximum temperature, number of days without rain, and number of days with maximum temperatures above 29°C, 30°C and 31°C.

Average forecasting gains were detected in models based on maximum temperature data. In particular, models based on weekly data corresponding to the maximum temperatures during the plant growth phase outperformed the long-sample estimation. This result is in line with many studies that indicate that temperature extremes can be critical to reducing yields, especially if they coincide with the flowering stage of the crop.

By testing their local relative performance, we also found that the forecasting gains are time-varying. Models including high frequency weather data outperformed particularly during the three consecutive campaigns after the 2008/09 campaign when soybean yield decreased almost by 50%.

Finally, although significant at a 10% level, forecast combinations showed a better forecasting performance than individual forecasting models.

Therefore, forecasting using different frequencies could be worthwhile to explore since it could be useful in some cases such as forecasting crop yields using temperature data. There is no need to use a given weighting scheme as models based on simple averages of high frequency data, or forecast combinations of different frequencies can have better forecasting performance.

REFERENCES

Andreou, E., E. Ghysels, and A. Kourtellos (2013). Should macroeconomic forecasters use daily financial data and how? *Journal of Business & Economic Statistics 31*(2), 240–251.

Auffhammer, M., V. Ramanathan, and J. R. Vincent (2012). Climate change, the monsoon, and rice yield in India. *Climatic Change 111*(2), 411–424.

Clark, T. E. and M. W. McCracken (2010). Averaging forecasts from VARs with uncertain instabilities. *Journal of Applied Econometrics 25*(1), 5–29.

Ghysels, E., P. Santa-Clara, and R. Valkanov (2004). The MIDAS touch: Mixed data sampling regression models. *University of North Carolina and UCLA Discussion Paper.*.

Ghysels, E., P. Santa-Clara, and R. Valkanov (2006). Predicting volatility: getting the most out of return data sampled at different frequencies. *Journal of Econometrics 131*(1-2), 59–95.

Giacomini, R. and B. Rossi (2010). Forecast comparisons in unstable environments. *Journal of Applied Econometrics 25*(4), 595–620.

Hendry, D. F. and M. P. Clements (2004). Pooling of forecasts. *The Econometrics Journal 7*(1), 1–31.

Jansen, W. J., X. Jin, and J. M. de Winter (2016). Forecasting and nowcasting real GDP: comparing statistical models and subjective forecasts. *International Journal of Forecasting 32*(2), 411–436.

Lobell, D. B. and M. Burke (2009). *Climate change and food security: adapting agriculture to a warmer world*, Volume 37. Springer Science & Business Media.

Lobell, D. B. and M. B. Burke (2010). On the use of statistical models to predict crop yield responses to climate change. *Agricultural and Forest Meteorology 150*, 1443–1452.

Lobell, D. B., K. N. Cahill, and C. B. Field (2007). Historical effects of temperature and precipitation on California crop yields. *Climatic Change 81*, 187–203.

Makridakis, S., E. Spiliotis, and V. Assimakopoulos (2018). The M4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting 34*, 802–808.

Marcellino, M. (2002). Forecasting pooling for short time series of macroeconomic variables. *Oxford Bulletin of Economics and Statistics 66*, 91–112.

Ray, D. K., J. S. Gerber, G. K. MacDonald, and P. C. West (2015). Climate variation explains a third of global crop yield variability. *Nature Communications 6*, 5989.

Reilly, J. and D. Schimmelpfennig (2000). Irreversibility, uncertainty, and learning: portraits of adaptation to long-term climate change. *Climatic Change 45*(1), 253–278.

Schlenker, W. and D. B. Lobell (2010). Robust negative impacts of climate change on african agriculture.

Environmental Research Letters 5(1), 014010.

Schlenker, W. and M. J. Roberts (2009). Nonlinear temperature effects indicate severe damages to us crop yields under climate change. *Proceedings of the National Academy of Sciences 106*(37), 15594–15598.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

Timmermann, A. (2006). *Forecast Combinations*. North Holland.

Welch, J. R., J. R. Vincent, M. Auffhammer, P. F. Moya, A. Dobermann, and D. Dawe (2010). Rice yields in tropical/subtropical Asia exhibit large but opposing sensitivities to minimum and maximum temperatures. *Proceedings of the National Academy of Sciences 107*(33), 14562–14567.

Wheeler, T. R., P. Q. Craufurd, R. H. Ellis, J. R. Porter, and P. V. Prasad (2000). Temperature variability and the yield of annual crops. *Agriculture, Ecosystems & Environment 82*(1-3), 159–167.

APPENDIX INITIAL IN-SAMPLE ESTIMATIONS

Table A1: Initial in-sample estimations (1971/1972 – 2000/01)

Variable	Long-sample (annual)					Short-sample (weekly)					Combination (annual & weekly)
	(1)	(2)	(3)	(4)	(5) LASSO	(1)	(2)	(3)	(4)	(5) LASSO	(1)
$maxtemp$	-0.10				0.12*	-0.08***				-0.07	-0.08***
	(0.07)				(0.09)	(0.02)				(0.04)	(0.02)
$maxtemp > 29$					-0.002					0.002	
					(0.004)					(0.007)	
$maxtemp > 30$		-0.007***					-0.01***			-0.01*	
		(0.002)					(0.003)			(0.006)	
$maxtemp > 31$					-0.005					0.02***	
					(0.005)					(0.008)	
$rainfall$			0.0005**		0.0003			0.001***		0.001***	
			(0.0002)		(0.0002)			(0.0003)		(0.0002)	
$no - rain$				-0.005*	-0.002				-0.007*	-0.004	-0.004
				(0.003)	(0.003)				(0.004)	(0.003)	(0.002)
Δy_{t-1}	-0.60***	-0.60***	-0.39**	-0.52***	-0.41**	-0.48***	-0.52***	-0.24*	-0.48***	-0.33***	-0.45***
	(0.16)	(0.14)	(0.16)	(0.15)	(0.18)	(0.13)	(0.13)	(0.14)	(0.16)	(0.12)	(0.13)
$constant$	2.28	0.44***	-0.44**	1.11*	-2.29	2.47***	0.45***	-0.30***	0.62*	2.27***	3.07***
	(1.59)	(0.13)	(0.18)	(0.57)	(2.14)	(0.67)	(0.11)	(0.08)	(0.31)	(1.20)	(0.75)
$\hat{\sigma}$	0.174	0.152	0.161	0.169	0.153	0.148	0.146	0.136	0.169	0.099	0.143
AR 1-2 test	2.92	0.74	1.66	3.46	0.39	0.91	0.01	4.18	2.82	0.37	0.78
	[0.07]	[0.49]	[0.21]	[0.05]	[0.68]	[0.42]	[0.99]	[0.03]	[0.08]	[0.70]	[0.47]
ARCH 1-1 test	6.46	0.98	7.78	0.89	1.48	16.33	8.28	3.24	0.003	0.25	2.68
	[0.02]	[0.33]	[0.01]	[0.35]	[0.23]	[0.00]	[0.01]	[0.08]	[0.96]	[0.62]	[0.11]
Normality test	0.12	0.38	1.59	0.00	0.54	0.05	1.23	2.38	0.53	6.89	0.43
	[0.94]	[0.83]	[0.45]	[1.00]	[0.76]	[0.97]	[0.54]	[0.30]	[0.77]	[0.03]	[0.81]
Heteroskedasticity	0.37	0.45	0.58	1.28	0.51	0.95	0.65	0.32	0.08	0.35	0.59
	[0.82]	[0.77]	[0.68]	[0.30]	[0.88]	[0.45]	[0.63]	[0.86]	[0.98]	[0.97]	[0.73]
RESET test	1.82	1.45	2.58	1.20	2.02	1.40	1.97	2.17	3.03	0.81	1.61
	[0.18]	[0.25]	[0.10]	[0.32]	[0.16]	[0.26]	[0.16]	[0.13]	[0.07]	[0.46]	[0.22]

Note: standard errors are reported in parentheses and p-values in brackers. *** p< .01, ** p< .05, * p< .10
Source: own elaboration.

Table A2: Initial in-sample MIDAS estimations (1971/72-2000/01)

Climate regressor:	$maxtemp$	$maxtemp > 30$	$rainfall$	$no - rain$
β_1^{MIDAS}	-0.09***	-0.12**	0.01***	-0.07***
	(0.02)	(0.06)	(0.002)	(0.02)
β_2^{MIDAS}	0.49	0.13	0.17***	1.29**
	(0.36)	(0.23)	(0.06)	(0.51)
Δy_{t-1}	-0.52***	-0.56***	-0.21***	-0.49***
	(0.10)	(0.13)	(0.08)	(0.08)
$constant$	2.67***	0.49*	-0.32***	0.28***
	(0.61)	(0.28)	(0.05)	(0.09)
σ	0.138	0.144	0.126	0.160