CEREAL YIELD FORECASTING WITH SATELLITE DROUGHT-BASED INDICES, WEATHER DATA AND REGIONAL CLIMATE INDICES USING MACHINE LEARNING IN MOROCCO

El Houssaine BOURAS [1 2 †], Lionel JARLAN [2], Salah ER-RAKI [1 3], Clément ALBERGEL [4], Riad BALAGHI [5] and Saïd KHABBA [3 6]

[1] *ProcEDE, Cadi Ayyad University, Morocco (bouras.elhoussaine@gmail.com)*

[2] *CESBIO, University of Toulouse, France (lionel.jarlan@ird.fr)*

[3] *CRSA, University Mohammed VI Polytechnic, Morocco (s.erraki@uca.ma)*

[4] *CNRM, France (clement.albergel@esa.int)*

[5] *INRA, Rabat, Morocco (riad.balaghi@gmail.com)*

[6] *LMFE, Cadi Ayyad University, Morocco (khabba@uca.ma)*

[†] *Now at Department of Physical Geography and Ecosystem Science, Lund University, Sweden*

Abstract: The aim of this work was to develop early prediction models for cereal yields in Morocco, based on previous works that showed high linkage between cereal yields and various datasets including weather data (rainfall and air temperature), regional climate indices (North Atlantic Oscillation), and drought indices derived from remote sensing observation. The prediction models were developed using several machine learning algorithms. The results show that combining data from multiple sources outperformed models based on one dataset only. In addition, the remote sensing drought indices are a major source of information for cereal prediction when the forecasting is carried out close to harvest (2 months before), while weather data and, to a lesser extent, climate indices, are key variables for earlier predictions. The best models can accurately predict yield in January (4 months before harvest) with an $R^2 = 0.88$ and RMSE around 0.22 t.ha^{-1}.

Keywords: Crop yield forecasting; Machine learning; Remote sensing drought indices; Climate indices.

Résumé : L'objectif de ce travail est de développer des modèles de prévision précoce de rendement des céréales au Maroc, sur la base de travaux antérieurs qui ont montré un lien élevé entre le rendement des céréales et divers ensembles de données, y compris les données météorologiques (précipitations et température de l'air), les indices climatiques régionaux (Oscillation Nord-Atlantique), et les indices de sécheresse dérivés de l'observation par télédétection. Les modèles de prédiction ont été développés en utilisant plusieurs algorithmes d'apprentissage automatique. Les résultats montrent que la combinaison de données provenant de plusieurs sources a fourni de meilleurs résultats que les modèles basés sur un seul ensemble de données. En outre, les indices de sécheresse issus de la télédétection constituent une source d'information majeure pour la prévision des céréales lorsque la prévision est effectuée pas longtemps avant la récolte (2 mois avant), tandis que les données météorologiques et, dans une moindre mesure, les indices climatiques, sont des variables clés pour les prévisions antérieures. Les meilleurs modèles peuvent prédire avec précision le rendement en janvier (4 mois avant la récolte) avec un $R^2 = 0{,}88$ et un RMSE d'environ 0,22 t.ha^{-1}.

Mots-clés : Prévision du rendement des cultures ; Machine learning ; Indices de sécheresse par télédétection ; Indices climatiques.

Introduction

Climate change will affect global agricultural production in the future (Asseng *et al.*, 2015) and it will threaten food security in several regions of the world including the Mediterranean areas, which have long been identified as a hotspot of climate change (Lionello and Scarascia, 2018). In addition to the change expected in average climate characteristics, including temperature and precipitation, the increased frequency of extreme events may further reduce agricultural production. Indeed, drought can be responsible for a loss in agricultural production of 10-35% depending on its intensity, duration, and spatial extent (Kogan, 2019).

The frequency and intensity of drought periods will increase in the future (Vicente-Serrano *et al.*, 2020). In this context, accurate seasonal forecasting of crop yields is an important decision support tool to predict import needs as early as possible. In addition, it provides critical and timely information to enable farmers to make quick decisions to increase yields through improving agricultural practices during the growing season. Also, it allows to model global and local market prices (Peng *et al.*, 2016).

Crop growth models forced by seasonal weather forecasts and empirical regression-based models are the main widely applied methods to forecast crop yield (Basso and Liu, 2019). Crop growth models are able to describe crop growth and yield response to weather condition, soil, and management practices (Jones *et al.*, 2017). Thus, these models provide a good estimation of final crop yield, when the input variables and parameters are available throughout the growing season. The uncertainty of weather forcing data during, the period between the forecast date and end of the crop growing season is one of the main limitations when using these models to forecast the crop yield (Lawless and Semenov, 2005). On the other hand, a further challenge, is to provide the model with an accurate description of the crop, soil and the management practices through the numerous input parameters (pedology, information on crop type and variety, land use, sowing date, etc.). Given these main limitations, the majority of the national agriculture department use empirical regression-based models to forecast yield over large areas. These models rely on the use of some selected variables or indicators of environmental conditions (agrometeorological, and/or remotely sensed data) as independent variables to forecast crop yield (Balaghi *et al.*, 2008; Kogan *et al.*, 2013). The performance of the empirical models on forecasting crop yield is related to the availability of datasets (Martinez *et al.*, 2009). Generally, the empirical models are simple and need fewer parameter settings compared to crop growth models. In addition, as the quantity and the quality of observed data have increased in recent years, these models forecast crop yield with reasonable accuracy (Kogan *et al.*, 2013).

In this context, the objective of this work is to develop early forecasting models for cereals yield in Morocco at the provincial scale using different data source and machine learning approaches.

1. Materials and methods

1.1. Study area

In this study, we focus on the main cereal cropping region of Morocco (Figure 1a). Morocco is located at the southern edge of the mid-latitude storm track with a semi-arid climate (Driouech *et al.*, 2010). The climate is influenced by the Atlantic Ocean, the Mediterranean Sea and the Sahara, together with very steep orography in the Atlas region. Most of the precipitation falls during winter and spring from the beginning of November until the end of April (Driouech *et al.*, 2010). Cereals are one of the country's main crops. It is cultivated both in rainfed and irrigated fields, depending on access to water supply and climate conditions. Cereals can be sown as early as November 1st if significant rainfall occurs. Nevertheless, a persistent drought at the beginning of the growing season can delay sowing until January 15th often leading to a production loss through a decrease of the wheat cropped areas as many farmers are used to wait for regular rainfall events to seed at the beginning of the season. Harvesting usually takes place around the end of May (Balaghi *et al.*, 2008). Cereal production in Morocco exhibits high inter-annual variability due to uncertain rainfall and recurrent drought periods, and this variability is expected to increase in the future due to the impact of climate change (Bouras *et al.*, 2019).

1.2. Methodology

The data on cereal yield for 15 provinces were acquired from the Economic Services of the Ministry of Agriculture, 4 groups of provinces with similar cereal yield interannual variability are identified through a classification using the kmeans based on the correlative distance. The target variable is cereal yield and the potential predictors are satellite-based drought indices, weather data (rainfall and temperature) and climate indices derived from atmospheric and oceanic variables. Table 1 lists all the raw used datasets with their sources. For the satellite-based drought indices, the three widely used indices were selected, which are: the Vegetation Condition Index (VCI), the Temperature Condition Index (TCI) (Kogan, 1995) and the Soil Moisture Condition Index (SMCI) (Zhang and Jia, 2013). The VCI, TCI and SMCI are the normalized anomalies of

Normalized Difference Vegetation Index (NDVI), Land Surface Temperature (LST) and soil moisture (SM). Regarding the climate indices we have selected the North Atlantic Oscillation (NAO), the Scandinavian Pattern (SCA) and the tow leading modes of Sea Surface Temperature (SST), which are correlated with wheat yield in Morrocco (Jarlan *et al.*, 2014). All these indices were camputed at monthly scale during the study periode from 2000 to 2017.

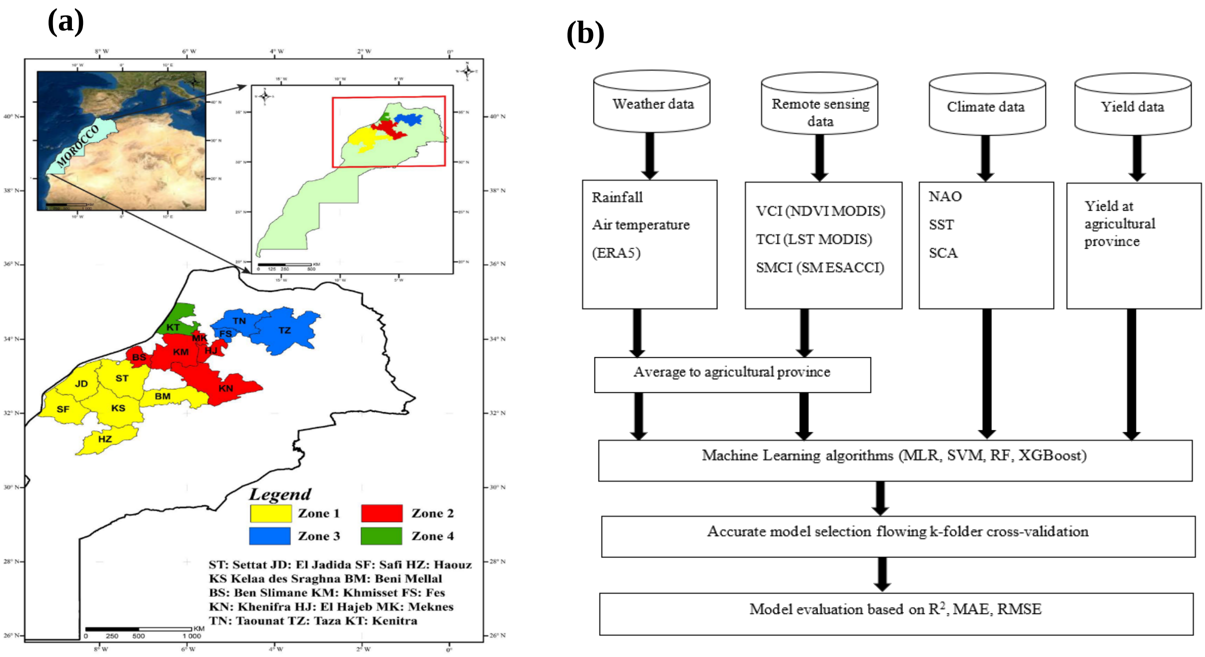
(a)
MOROCCO
Legend
Zone 1
Zone 2
Zone 3
Zone 4
ST: Settat JD: El Jadida SF: Safi HZ: Haouz
KS Kelaa des Sraghna BM: Beni Mellal
BS: Ben Slimane KM: Khmisset FS: Fes
KN: Khenifra HJ: El Hajeb MK: Meknes
TN: Taounat TZ: Taza KT: Kenitra
KT
TN
TZ
FS
MK
KM
HJ
BS
ST
JD
SF
KS
BM
KN
HZ
0 250 500 1 000 KM
(b)
Weather data
Remote sensing data
Climate data
Yield data
Rainfall
Air temperature
(ERA5)
VCI (NDVI MODIS)
TCI (LST MODIS)
SMCI (SM ESACCI)
NAO
SST
SCA
Yield at agricultural province
Average to agricultural province
Machine Learning algorithms (MLR, SVM, RF, XGBoost)
Accurate model selection flowing k-folder cross-validation
Model evaluation based on R², MAE, RMSE

figure 1: The study areas with the 15 provinces and results of the four classifications (see text) (a) and the main inputs data and methodology used in this study (b).

In order to build the seasonal forecasting models, we relied on Multiple Linear Regression, and three non-linear machine learning algorithms extensively used for crop yield prediction (van Klompenburg *et al.*, 2020) which are: Support Vector Machine (SVM), Random Forest (RF) and eXtreme Gradient Boost (XGBoost). An overview of the methodology is represented in the flowchart of Figure 1b.

Table 1. Summary of the raw characteristics of the data sets used for cereal yield prediction.

Category	Product	Variable	Spatial Resolution	Temporal Resolution	Source
Crop Yield		Crop yield	Province level	Yearly	Ministry of agriculture of Morocco
Remote sensing	MOD13A2	NDVI	1 km	16-Day	https://lpdaac.usgs.gov
	MOD11A1	LST	1 km	Daily	
	ESA CCI SM	SM	25 km	Daily	https://www.esa-soilmoisture-cci.org/
Weather	ERA5	Temperature, Rainfall	30 km	Daily	https://www.ecmwf.int/en/forecasts/dataset
Climate		NAO, SCA, SST		Monthly	https://psl.noaa.gov/

2. Results

2.1. Choose of inputs data sets

In order to identify the best combination of input data among the satellite-based drought indices, the weather data and the climate indices, the forecasting models of cereal yields were developed using the different combinations of inputs data from October to April (about 1 month prior to harvest) in April. The statistical metrics for the different combination of input dataset and for the different methods are reported at Table 2. On average, the more input datasets are considered, the better prediction performances are achieved. In addition, all statistical metrics are improved when adding a dataset and this is also true for all the tested methods. The results showed that the yield variability is reasonably explained with satellite-based drought indices only with R² values ranging from 0.67 (for MLR) to 0.81 (for XGBoost), and RMSE from 0.66 t.ha^{-1} (for MLR) to 0.44 t.ha^{-1} (for XGBoost). By combining satellite-based drought indices and weather data the performances of all models are improved by 2-7% for R² and by 25-30% for RMSE. The best statistical metrics are obtained by combining the three data sets with a further improvement of the statistical metrics by about 11-45% for RMSE and 4-10% for R² depending on the used method.

Table 2. Statistical metrics (RMSE, MAE and R²) of the forecasting models for the three combination of input data from October to April (all correlation coefficients are significant at the 99% level).

Inputs data	Models	RMSE (t. ha^{-1})	MAE (t. ha^{-1})	R²
Satellite-based drought indices only	MLR	0.66	0.57	0.67
	SVM	0.54	0.43	0.78
	RF	0.46	0.35	0.80
	XGBoost	0.45	0.34	0.81
Satellite-based drought indices and weather data	MLR	0.46	0.39	0.72
	SVM	0.40	0.31	0.80
	RF	0.34	0.24	0.84
	XGBoost	0.37	0.25	0.86
Satellite-based drought indices, weather data and climate indices	MLR	0.41	0.31	0.75
	SVM	0.25	0.21	0.88
	RF	0.22	0.19	0.92
	XGBoost	0.20	0.16	0.95

2.2. Model performance as a function of lead time before harvest

The performance of the forecasting models using the three datasets are evaluated as a function of the leading time prior to harvest from January to March (from 4 to 2 months before harvest). The RMSEs and R² of the models are plotted as a function of the lead time at figure 2 to investigate the prediction accuracy. In addition, the relative importance of each data set is reported in Table 2.

The closer to harvest the forecast is carried out, the better the performance metrics as shown by the increase of the correlation coefficient and the drop of RMSE when going from January to March (Figure 2). The best method whatever the lead time is XGBoost as already shown followed closely by RF based approaches. The models based on XGBoost explain 88, 92 and 96% of yield variability (RMSE of 0.41, 0.34 and 0.22 t.ha^{-1}) for a forecasting from January, February and March, respectively. By contrast, the poorest results are obtained with MLR with a strong gap of metrics with regards to the non-linear machine learning approaches (R² is below to 0.75 for MLR while the correlations for the non-linear methods are above 0.90). While a slight improvement of the model metrics is observed when going from January to February, considering predictors in March leads to a significant jump in the metrics with RMSE close to the international standard of 0.20 t. ha^{-1} for the XGBoost method and, to a lesser extent for RF model. This is probably related to the very high correlation between NDVI around the crop development peak in March and wheat yields that was already shown by various authors (Belaqziz *et al.*, 2014; Jarlan *et al.*, 2014) giving a large weight to VCI at this time. The dominating importance of the satellite drought indices in March for the model based on XGBoost support this assumption (Table 3).

Table 3. The importance of different inputs variables for cereal yield predition using XGBoost model at the national scale in January, February and March.

Inputs data	January	February	March
Satellite-based drought indices only	20%	35%	73%
weather data	60%	53%	16%
climate indices	20%	12%	11%

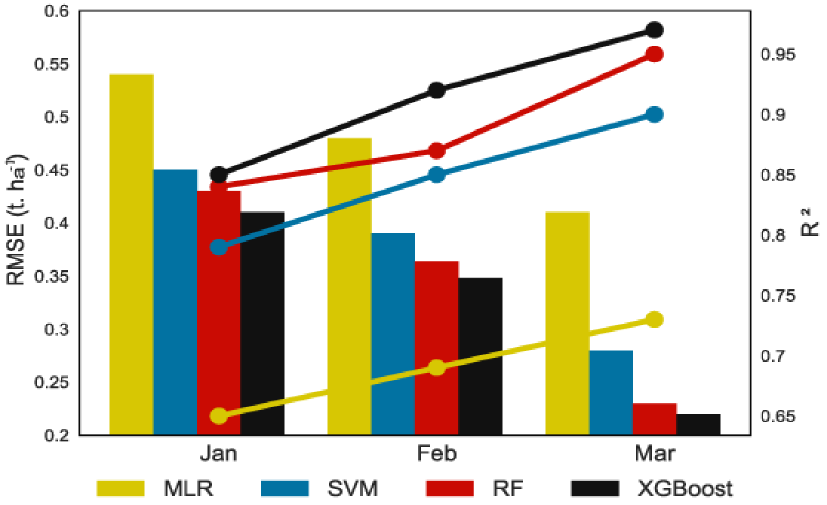

figure 2: The temporal progression (between January to March) of the model performance at the national scale, expressed by R² (lines) and RMSE (bars), based on the four methods (MLR, SVM, RF and XGBoost) and all inputs data.

Other striking comments can be drawn by analyzing the importance of the three data sets: (1) the weather data dominates largely in January and, to a lesser extent, in February while a strong shift is observed in March when satellite drought indices take the lead over the two other data sets. This is in agreement with the already observed high correlation between yields and precipitation around emergence in October and November, and between yields and temperature in December during the tillering stage (Jarlan *et al.*, 2014); (2) likewise, the importance of climate indices decreases with the lead time and their contribution is the lowest of the three data sets apart in January when it contributes to 20% like the satellite drought indices. Indeed, the highest correlation with yields was found in December and January for NAO and SCA, respectively while the correlations with the SST leading modes peak in October and February for "Atlantic Niño" and Atlantic Tripole modes, respectively. In addition, linkages between climate indices, in particular based on SST, and yields occur through teleconnection meaning that the relationships are indirect. This means that when good quality precipitation and temperature data are available, they should be preferred to climate indices as they provide more direct information on growing conditions; (3) satellite drought indices play a dominating role for an early prediction in March only when they contribute up to 73% on the prediction accuracy. Nevertheless, significant contribution is observed in February (35%) and in January (20%). This is because VCI and TCI was found to be significantly correlated to final yields in January and February, and because SMCI is significantly related to yields as early as October around the emergence stage (Bouras *et al.*, 2020). Indeed, high moisture at the upper soil layers at this time facilitate and favorites the emergence and significant rainfall event during October-December promotes the farmer to seed leading to an increase of yields production (Balaghi *et al.*, 2013).

Conclusion

The results presented in this study clearly showed that combining satellite-based drought indices, weather data and climate indices is better predictors of cereal yield, and integrate these data into machine learning algorithms can provide useful tools to early forecast of cereal yield in Morocco. And it can be used as source of timely information needed to decision making during the growing season. However cereal yields may be related to other factors that were not considered in our study, such as sowing date, soil properties, and other management aspects. In particular, the sowing dates can shift the growing season with regards to the average growing period from November to May considered in this study. Finally, it is necessary to combine the empirical models developed in this study with a crop growth model in order to include climate change impacts on crop yield forecasting.

Acknowledgement: This work was carried out within the framework of the Joint International Laboratory TREMA (https://www.lmi-trema.ma/). This work was founded by the ERANETMED03–62 CHAAMS project, the ACCWA project (grant agreement no: 823965). The H2020 PRIMA ALTOS project, MISTRALS/SICMED2, PHC Toubkal #39064WG/2018 and PRIMA-IDEWA project are also acknowledged for additional funding.

Bibliography

Asseng, S., Ewert, F., Martre, P., Rötter, R.P., et al., 2015: Rising temperatures reduce global wheat production. *Nature Climate Change*, **5**, 143-147.

Balaghi, R., Tychon, B., Eerens, H., Jlibene, M., 2008: Empirical regression models using NDVI, rainfall and temperature data for the early prediction of wheat grain yields in Morocco. *International Journal of Applied Earth Observation and Geoinformation*, **10** (4), 438-452.

Basso, B., Liu, L., 2019: Seasonal crop yield forecast: Methods, applications, and accuracies. *Advances in Agronomy*, **154**, 201-255.

Belaqziz, S., Mangiarotti, S., Le Page, M., Khabba, S., Er-Raki, S., Agouti, T., Drapeau, L., Kharrou, M.H., El Adnani, M., Jarlan, L., 2014: Irrigation scheduling of a classical gravity network based on the Covariance Matrix Adaptation - Evolutionary Strategy algorithm. *Computers and Electronics in Agriculture*, **102**, 64-72.

Bouras, E.H., Jarlan, L., Er-Raki, S., Albergel, C., Richard, B., Balaghi, R., Khabba, S., 2020: Linkages between rainfed cereal production and agricultural drought through remote sensing indices and a land data assimilation system: A case study in Morocco. *Remote Sensing*, **12**, 1-35.

Bouras, E., Jarlan, L., Khabba, S., Er-raki, S., Dezetter, A., Sghir, F., & Tramblay, Y. 2019: Assessing the impact of global climate changes on irrigated wheat yields and water requirements in a semi-arid environment of Morocco. *Scientific Reports*, **9**, 1-15.

Driouech, F., Déqué, M., Sánchez-Gómez, E., 2010: Weather regimes-Moroccan precipitation link in a regional climate change simulation. *Global Planet Change*, **72** (1-2).

Jarlan, L., Driouech, F., Tourre, Y., Duchemin, B., Bouyssié, M., Abaoui, J., Ouldbba, A., Mokssit, A., Chehbouni, G., 2014: Spatio-temporal variability of vegetation cover over Morocco (1982-2008): Linkages with large scale climate and predictability. I*nternational Journal of Climatology*, **34**, 1245-1261.

Jones, J.W., Antle, J.M., Basso, B., Boote, K.J., Conant, R.T., Foster, I., Godfray, H.C.J., Herrero, M., Howitt, R.E., Janssen, S., Keating, B.A., Munoz-Carpena, R., Porter, C.H., Rosenzweig, C., Wheeler, T.R., 2017: Toward a new generation of agricultural system data, models, and knowledge products: State of agricultural systems science. *Agricultural Systems*, **155**, 269-288.

Kogan, F., 2019: Vegetation health for insuring drought-related yield losses and food security enhancement. In: Remote Sensing for Food Security. Sustainable Development Goals Series. Springer, Cham, 163-173.

Kogan, F., Kussul, N., Adamenko, T., Skakun, S., Kravchenko, O., Kryvobok, O., Shelestov, A., Kolotii, A., Kussul, O., Lavrenyuk, A., 2013: Winter wheat yield forecasting in Ukraine based on Earth observation, meteorologicaldata and biophysical models. *International Journal of Applied Earth Observation and Geoinformation*, **23**, 192-203.

Kogan, F., 1995: Application of vegetation index and brightness temperature for drought detection. *Advances in Space Research*, **15** (11), 91-100.

Lawless, C., Semenov, M.A., 2005: Assessing lead-time for predicting wheat growth using a crop simulation model. *Agric. For. Meteorol.*, **135** (1-4), 302-313.

Lionello, P., Scarascia, L., 2018: The relation between climate change in the Mediterranean region and global warming. *Regional Environmental Change*, **18** (5) .

Martinez, C.J., Baigorria, G.A., Jones, J.W., 2009: Use of climate indices to predict corn yields in southeast USA. I*nternational Journal of Climatology*, **29** (11), 1680-1691.

Peng, Y.H., Hsu, C.S., Huang, P.C., 2016: Developing crop price forecasting service using open data from Taiwan markets, in: TAAI 2015 - Conference on Technologies and Applications of Artificial Intelligence.

Van Klompenburg, T., Kassahun, A., Catal, C., 2020: Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, **177**, 105709.

Vicente-Serrano, S.M., Quiring, S.M., Peña-Gallardo, M., Yuan, S., Domínguez-Castro, F., 2020: A review of environmental droughts: Increased risk under global warming? *Earth-Science Reviews*, **201**, 102953.

Zhang, A., Jia, G., 2013: Monitoring meteorological drought in semiarid regions using multi-sensor microwave remote sensing data. *Remote Sensing of Environment*, **134**, 12-23.