Machine learning model accurately predict maize grain yields in conservation agriculture systems in Southern Africa

Francis Muthoni
International Institute of Tropical Agriculture (IITA)
Arusha, Tanzania
f.muthoni@cgiar.org

Christian Thierfelder
International Maize and Wheat Improvement Center (CIMMYT)
Harare, Zimbabwe
c.thierfelder@cgiar.org

Bester Mudereri
International Centre of Insect Physiology and Ecology (ICIPE)
Nairobi, Kenya
mudereri@gmail.com

Julius Manda
International Institute of Tropical Agriculture (IITA)
Arusha, Tanzania
j.manda@cgiar.org

Mateete Bekunda
International Institute of Tropical Agriculture (IITA)
Arusha, Tanzania
m.bekunda@cgiar.org

Irmgard, Hoeschle-Zeledon
International Institute of Tropical Agriculture (IITA)
Ibadan, Nigeria
i.hoeschle-zeledon@cgiar.org

***Abstract*—Adoption of CA in smallholder farmers in Africa is (s)low partly due to poor spatial targeting. Mapping the crop yield from different CA systems across space and time can reveal their spatial recommendation domains. Integration of machine learning (ML) and free remotely sensed big data have opened huge opportunities for data-driven insights into complex problems in agriculture. The objective of this study was to estimate the spatial-temporal variations of maize grain yields from 13-year multi-location on-farm trials implemented across four countries in southern Africa. The agronomic data from the long-term CA trials is used together with gridded biophysical and socio-economic variables. A spatially explicit random forest (RF) algorithm was developed. Spatial variation of yield advantage or loss from CA practices was compared with conventional tillage practices (CP) during seasons with above and below-normal precipitation. The out-of-bag accuracy of the RF model was $R^2 = 0.63$ and RMSE = 1.2 t ha^{-1}. The variable importance analysis showed that the altitude, precipitation, temperature, and soil physical and nutrients conditions variables explained most of the variation in maize grain yield. Maps were generated to identify the locations where CA had a yield advantage over CP during seasons with below and above-average precipitation. The CA showed yield gains of up-to 1 t ha^{-1} during the season with drought compared to CP. In contrast, the CA returned yield losses of similar magnitude during the season with above-normal precipitation, except in Mozambique. The maps on yield advantage will support the spatial targeting of CA to suitable biophysical and socioeconomic contexts. Results demonstrates that multi-source remotely sensed data, coupled with advanced and efficient machine learning algorithms can provides accurate, cost-effective, and timely platform for predicting the optimal locations for the upscaling sustainable agricultural technologies.**

Keywords—agronomic trials, big data, climate variability, conservation agriculture, machine learning, random forest, remote sensing, yield gain

I. INTRODUCTION

Conservation Agriculture (CA) involves the simultaneous application of minimum soil disturbance, crop residue retention, and crop diversification through rotations and intercropping systems [1]. CA improves water infiltration capacity, available soil moisture and gradually increase soil carbon content. Recent publications has shown that CA can increase crop yields compared to conventional tillage (CP) practices [2]. However, adoption of CA by smallholder farmers in Africa is low [3] that is partly attributed to poor spatial targeting [3-5]. To design robust strategists for scaling out CA practices, it is important to develop spatially explicit and evidence-based recommendation domains that show which CA practices are suitable at different agro-ecologies. Tesfaye [4] applied a multi-criteria approach to map the recommendation domains of CA applied that incorporate elements of subjectivity. Recently there have been increase of data from agronomic trials [1, 6, 7]. Moreover, there is increase of big data from earth observation platform coupled with advances in robust spatial analytics and algorithms [8-10] that improves prediction of crop yields over time and space. However, the robust analytics have not been applied to map crop yield from CA systems in Africa. Moreover, robust analytical methods and algorithms have emerged for deriving useful insights from complex data.

The machine learning (ML) algorithms have proofed to be accurate for prediction of crop yields [11-14]. The ML algorithms are flexible in dealing with complex data characterized by high dimensionality, multicollinearity, outliers, and nonlinear relationships [14]. ML coupled with high spatial and temporal resolution earth observation data has shown great potential for spatial-temporal prediction of yields. The ML approaches offer robust algorithms that learn patterns from big data to predict the desired output. Moreover, ML algorithms like RF are flexible to handle unstructured data to unravel complex relationships [14]. However, different ML algorithms have different predictive power depending on the inherent characteristics of input data [11]. Though the current study applied RF for spatial prediction of maize yield, future studies should compare the performance of different ML algorithms.

This paper apply a random Forest (RF) algorithm to estimate the spatial-temporal variation of maize grain yields from 13-year multi-locational on-farm trials over four countries in Southern Africa (Malawi, Zambia, Zimbabwe and Mozambique. The predicted yields are applied to access the areas where CA has yield advantage over CP during seasons with above normal and below normal precipitation regime. Results provides evidence-based information on potential of CA practices to increase resilience to climate change and variability in smallholder farming systems in Southern Africa.

978-1-7281-6561-5/21/$31.00©2021

II. MATERIALS AND METHODS

The study area is covers four countries in Southern Africa and measures over 2 million square kilometers (Fig. 1). The area experience unimodal rainfall regime from October to April. The rainfall regime is experiencing significant increase or decrease although varied over space and time [15].

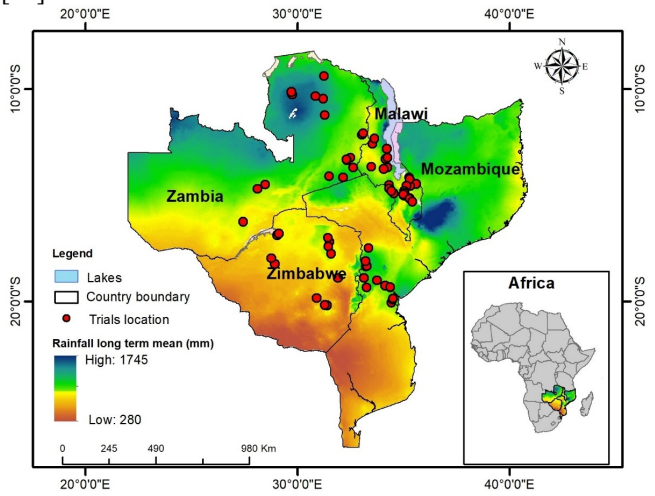

Fig. 1. Map of study area with on-farm trial sites superimposed on long-term (1981 - 2017) mean annual precipitation generated from CHIRPS-v2.

A. Agronomic data

Agronomic data were derived from CIMMYT's long-term on-farm CA trial database established over 13 growing seasons between 2004/2005 and 2016/2017 in the four countries. This comprised georeferenced yields with details of tillage practices, maize varieties, and other complementary agronomic management practices. The database comprised of the five CP and seven CA treatments. A total of 18 maize varieties with greater than 100 observations in the agronomic database were considered in this analysis. Detailed description of the agronomic trials can be found in [1] and [6]. A total of a total of 1137 observations with precise GPS of the farm location.

B. Remote sensing variables

Gridded monthly and seasonally aggregated data on precipitation, actual evapotranspiration (ET), maximum (Tmax), and minimum temperatures (Tmin) was obtained from TerraClimate [16]. The standardized precipitation anomalies were generated as the difference between total precipitation for each growing season and the long-term mean (1981 - 2017) was divided by the standard deviation to derive the seasonal. The ASTER DEM was used to generate the slope layer. Soil physical and chemical characteristics were obtained from SoilGrids250m database [17]. We used the seasonal maximum value composite of MODIS enhanced vegetation index (EVI) with 250m was used. Social economic variables included the cattle density and market access. Length of growing period layer was downloaded from FAO.

III. DATA ANALYSIS

A. Model training and prediction

We applied RF model (Breiman, 2001) from the 'ranger' R package [18] to develop the spatial prediction framework. The RF regression analysis was used to predict the yield using both the continuous (n = 35) and the categorical variables (n = 38). The function 'tuneRF' was initially used to find the optimum parameters to run the random forest algorithm. The optimum number of trees (ntree) obtained through the 'tuneRF' function was 100. The model performance was evaluated using the root mean squared error (RMSE) and the amount of variation explained by the model (R2). The permutation variable importance measure [19] was used to evaluate the importance of features in the RF model. we used the 'pdp' package [20] to produce the partial dependency plots for variables with the highest contribution to the model. The following single model was used to predict maize yield in the four countries in tonnes per hectare (t ha^{-1}). The model was parameterized as follows:

MGY [t ha^{-1}] = f [Tillage treatment + Variety + Soil nutrients + Soil physical + Climate + Terrain + Socioeconomic + EVI+ CA period] [Eq. 1]

Where: MGY is the maize grain yield (t ha^{-1}). The tillage treatments and variety are defined in Tables 1 and 2, and the other continuous variables are as described in Table 3. The 'CA period' represented the number of years since the implementation of CA trials and was divided into two categories i.e., 0 – 5 years and 6 – 13 years. The variables tillage treatment, variety, CA period are categorical variables (Tables 1 - 2).

After training the RF model, it was applied for spatial prediction of maize grain yield for PAN53 maize variety for different CA treatments. Predictions were conducted for two growing seasons with above (2004/2005) and below average (2016/2017) precipitation to evaluate the performance of CA practices under contrasting soil moisture conditions. The multivariate model allows spatial predictions of yield for multiple combinations of agronomic variables, which are impossible if individual models are fitted.

B. Calculating yield advantage of CA over CP treatments

The The spatial yield advantage or loss of CA practices was calculated as the difference in the predicted yield between the CA and CP treatments (Equation 2) because of its ease of interpretation and the relevance for comparing potential gains. The raster representing predicted yield from the CPMS (Table 1) was used as the control and contrasted with the raster outputs from each of the seven CA practices during a dry (2005) and wet (2017) season to examine the mean yield advantage (loss) of CA practices during those respective periods using Eq. 2 below:

Yield advantage = yield (ca) – yield (cpms) [Eq. 2]

Yield advantage is the amount by which the predicted yield from CA is higher or lower than from the CP. Yield $_{(ca)}$ and yield $_{(cpms)}$ is the predicted yield from CA and CPMS systems, respectively. Positive or negative values indicate that CA systems have a yield advantage or loss over CP. The results were plotted against each other at a site, variety, and season to determine which treatment and varieties had a greater yield advantage or loss. The primary assumption is that when the yield of the two treatments is equal, there is no change in yield (yield advantage = 0). In contrast, a positive yield value signifies the superiority of the CA method over

the CP. Similarly, a negative value shows that the CP produces better yields at that location and time compared to the CA practice.

IV. RESULTS

A. Validation of remote sensing estimates with gauge data

The RF regression model predicted the maize yield with a relatively high accuracy of R^2 = 0.63 and out-of-bag RMSE = ±1.2 t ha^{-1} (Fig. 2). The top five most important predictors for maize yield were the elevation, February precipitation, ET, total precipitation, and precipitation anomaly (Fig. 3).

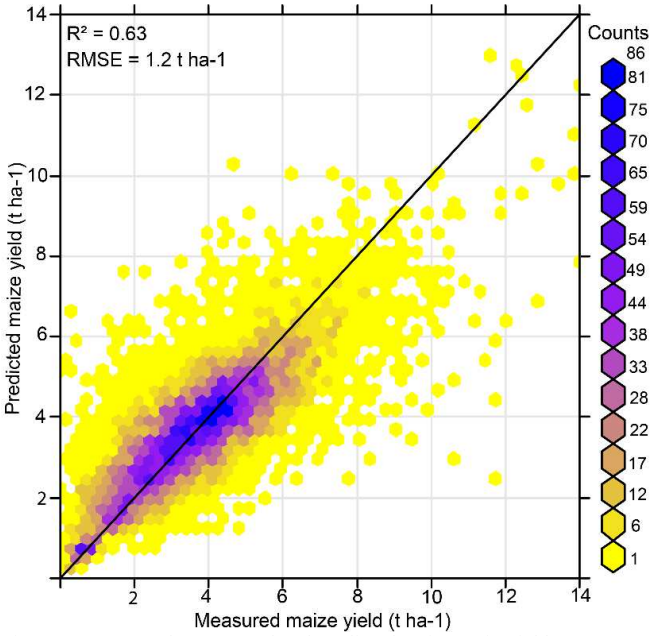

Fig. 2. Agreement the measured and predicted maize grain yields

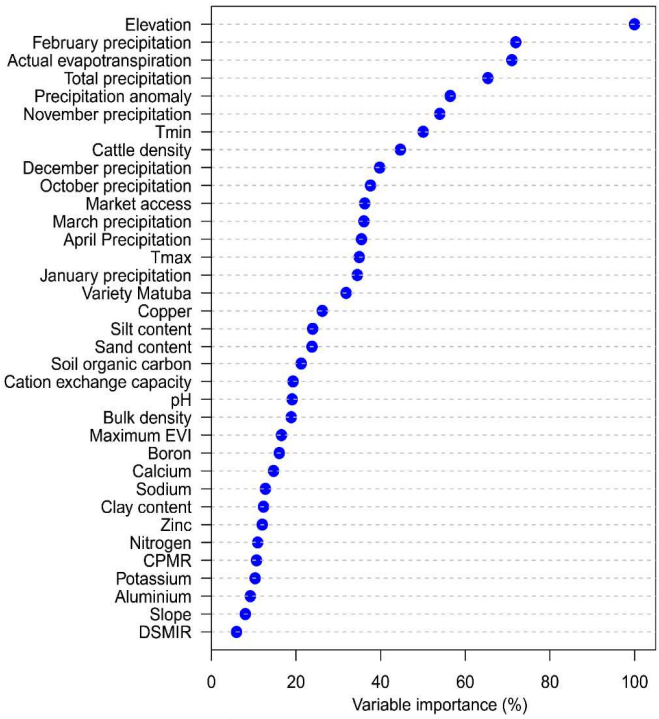

Fig. 3. Long-term spatial temporal trends of CHIRPS-v2 rainfall for six West Africa countries for 37 years period (1981 – 2017)

B. Spatial prediction of maize yields

Fig. 4 and Fig. 5 shows the predicted maize yield maps for the PAN53 variety with CPMS and DSMIR treatments for dry (2004/2005) and wet (2016/2017) growing seasons, respectively. Yields were higher during wet (2016/2017) compared to dry (2004/2005) seasons except in some parts of Northeast Mozambique.

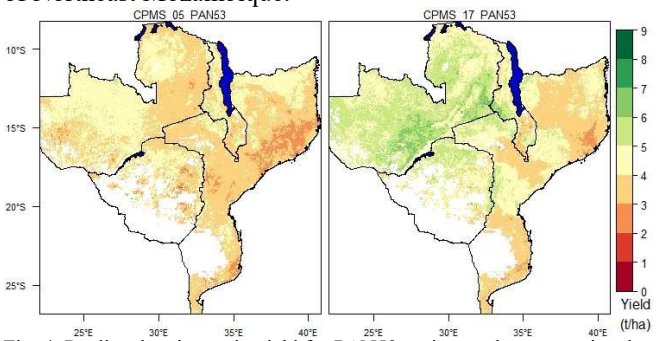

Fig. 4. Predicted maize grain yield for PAN53 variety under conventional practice with sole maize (CPMS) during 2004/2005 (a) and 2016/2017 (b) growing seasons that had below and above normal precipitation, respectively

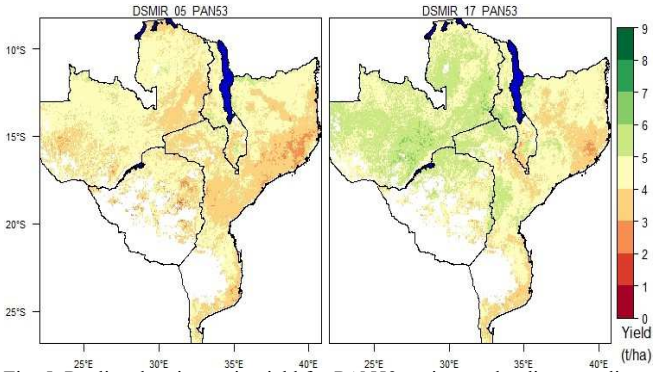

Fig. 5. Predicted maize grain yield for PAN53 variety under direct seeding with maize intercropping and rotation (DSMIR) conservation agriculture practice (CA) during 2004/2005 (a) and 2016/2017 (b) growing seasons.

C. Spatial prediction of maize yields

Fig. 6 shows the yield advantage (loss) for the PAN53 variety in the DSMIR CA systems compared to the CPMS during dry (2004/2005) and wet (2016/2017) seasons. During the dry season, the DSMIR CA systems returned a yield advantage (0.1 – 1 t ha^{-1}) compared to CPMS across the entire region (Fig. 6). The DSMIR system showed the highest yield advantage (0.5 – 1 t ha^{-1}) in Mozambique, north-eastern Zambia, and central Malawi. In contrast, in the wet season, all CA systems returned a yield loss compared to CPMS in a large swath of Zambia, Zimbabwe, and western Malawi (Fig. 6b). Results clearly show a yield loss when CA practices are applied during a season with above-average precipitation in the three countries. This pattern was replicated across other CA treatments (data not shown)

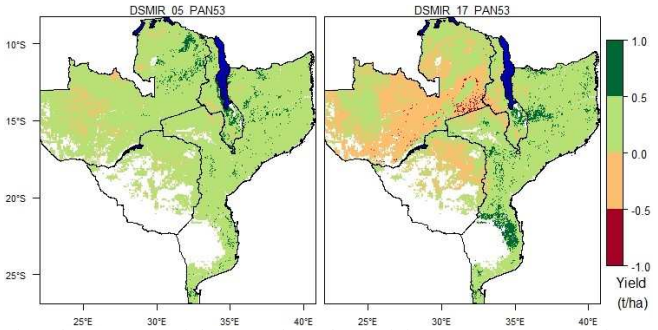

Fig 6. long-term spatial temporal trends of minimum temperature (Tmin) for six West Africa countries

V. DISCUSSIONS

CHIRPS-v2 This paper utilized big data from agronomic trials, remote sensing platforms, and RF algorithm to predict the spatial variation of maize grain yield under CP and CA practices in four southern African countries. The effectiveness of CA over CP systems is examined during the 2004/05 and 2016/17 growing seasons that experienced below and above-normal precipitation, respectively. We generated maps to identify the locations where CA systems had a yield advantage over the CPMS systems during seasons with drought and above-average precipitation. The CA systems showed a yield advantage of 0 – 1 t ha-1 during the drought season (2005) compared to CPMS systems. In contrast, during the season with above-normal precipitation, the CA systems returned yield losses (0 to -1 t ha-1) over a larger area in southern Africa, except in Mozambique. This information is essential for targeting the CA practices to the site where it provides maximum yield advantage as an adaptation measure to droughts.

The order of variable importance showed that maize yield in this region is heavily dependent on precipitation and temperature. These results concur with the studies related to the growth and production of maize to precipitation [7, 21, 22]. Precipitation (especially for February, November, and December) distinctly comes as the most important covariate overall, including its anomaly in the different years. The timing of precipitation are critical since the other growth and maturity stages of maize in this region have varied moisture requirement i.e. precipitation in November – December coincides with the germination, while in February it coincides with silking and anthesis growth stages that require sufficient precipitation [23].

This study demonstrated the utility of ML approaches for generating information from big and complex agronomic and remote sensing data to inform evidence-based targeting of CA technologies in southern Africa. The maps are decision support tools that are important for guiding investment as well as extension and development agencies. These maps are helpful in identifying the best-bet locations for extrapolating CA systems and complementary technologies such as improved maize varieties that are suited for different agro-ecologies. Results show that CA practices enhances resilience to droughts in large area in Southern Africa.

ACKNOWLEDGMENT

This study was funded by USAID through grant numbers: AID-BFS-G-11-00002 under the Feed the Future initiative that supported the Africa RISING program. Long-term on-farm trials were further supported by the donors of the MAIZE CGIAR Research Program (www.maize.org). Other funders that previously supported the establishment of on-farm trials, namely Bundesministerium für wirtschaftliche Zumsammenarbeit (BMZ/GIZ Grants No 2003.7860.4 – 002.00 and 13.22.44.5-002.00), International Fund for Agriculture Development (IFAD, Grants No 898 and 1309), the USAID- funded Feed the Future project Sustainable Intensification in Maize-Legume Systems in Eastern Province of Zambia (SIMLEZA, Grant No EEM-G-00-04-00013) are gratefully acknowledged. The funders had no role in the production of manuscript and vies expressed are solely those of authors.

REFERENCES

[1] C. Thierfelder, R. Matemba-Mutasa, and L. Rusinamhodzi, "Yield response of maize (Zea mays L.) to conservation agriculture cropping system in Southern Africa," *Soil and Tillage Research,* vol. 146, pp. 230-242, 2015.

[2] P. R. Steward, A. J. Dougill, C. Thierfelder, C. M. Pittelkow, L. C. Stringer, M. Kudzala*, et al.*, "The adaptive capacity of maize-based conservation agriculture systems to climate stress in tropical and subtropical environments: A meta-regression of yields," *Agriculture, Ecosystems & Environment,* vol. 251, pp. 194-202, 2018.

[3] F. Baudron, C. Thierfelder, I. Nyagumbo, and B. Gérard, "Where to Target Conservation Agriculture for African Smallholders? How to Overcome Challenges Associated with its Implementation? Experience from Eastern and Southern Africa," *Environments,* vol. 2, pp. 338-357, 2015.

[4] K. Tesfaye, M. Jaleta, P. Jena, and M. Mutenje, "Identifying Potential Recommendation Domains for Conservation Agriculture in Ethiopia, Kenya, and Malawi," *Environmental Management,* vol. 55, pp. 330-346, 2015/02/01 2015.

[5] P. Brandt, M. Kvakić, K. Butterbach-Bahl, and M. C. Rufino, "How to target climate-smart agriculture? Concept and application of the consensus-driven decision support framework "targetCSA"," *Agricultural Systems,* vol. 151, pp. 234-245, 2017/02/01/ 2017.

[6] C. Thierfelder, S. Cheesman, and L. Rusinamhodzi, "A comparative analysis of conservation agriculture systems: Benefits and challenges of rotations and intercropping in Zimbabwe," *Field Crops Research,* vol. 137, pp. 237-250, 2012.

[7] D. B. Lobell, M. Bänziger, C. Magorokosho, and B. Vivek, "Nonlinear heat effects on African maize as evidenced by historical yield trials,"

Nature Climate Change, vol. 1, p. 42, 03/13/online 2011.

[8] C. Nakalembe, I. Becker-Reshef, R. Bonifacio, G. Hu, M. L. Humber, C. J. Justice*, et al.*, "A review of satellite-based global agricultural monitoring systems available for Africa," *Global Food Security,* vol. 29, p. 100543, 2021/06/01/ 2021.

[9] M. Herrero-Huerta, P. Rodriguez-Gonzalvez, and K. M. Rainey, "Yield prediction by machine learning from UAS-based multi-sensor data fusion in soybean," *Plant Methods,* vol. 16, p. 78, 2020/06/01 2020.

[10] G. Hyman, D. Hodson, and P. Jones, "Spatial analysis to support geographic targeting of genetypes to environments," *Frontiers in Physiology,* vol. 4, 2013-March-18 2013.

[11] W. Mupangwa, L. Chipindu, I. Nyagumbo, S. Mkuhlani, and G. Sisito, "Evaluating machine learning algorithms for predicting maize yield under conservation agriculture in Eastern and Southern Africa," *SN Applied Sciences,* vol. 2, p. 952, 2020/04/24 2020.

[12] N. Vergopolan, S. Xiong, L. Estes, N. Wanders, N. W. Chaney, E. F. Wood*, et al.*, "Field-scale soil moisture bridges the spatial-scale gap between drought monitoring and agricultural yields," *Hydrol. Earth Syst. Sci.,* vol. 25, pp. 1827-1847, 2021.

[13] W. T. Dado, J. M. Deines, R. Patel, S.-Z. Liang, and D. B. Lobell, "High-Resolution Soybean Yield Mapping Across the US Midwest Using Subfield Harvester Data," *Remote Sensing,* vol. 12, 2020.

[14] S. Delerce, H. Dorado, A. Grillon, M. C. Rebolledo, S. D. Prager, V. H. Patino*, et al.*, "Assessing Weather-Yield Relationships in Rice at Local Scale Using Data Mining Approaches," *PLoS One,* vol. 11, p. e0161620, 2016.

[15] F. K. Muthoni, V. O. Odongo, J. Ochieng, E. M. Mugalavai, S. K. Mourice, I. Hoesche-Zeledon*, et al.*, "Long-term spatial-temporal trends and variability of rainfall over Eastern and Southern Africa," *Theoretical and Applied Climatology,* vol. 137, pp. 1869-1882, 2019/08/01 2019.

[16] J. T. Abatzoglou, S. Z. Dobrowski, S. A. Parks, and K. C. Hegewisch, "TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015," *Scientific Data,* vol. 5, p. 170191, 01/09/online 2018.

[17] T. Hengl, J. Mendes de Jesus, G. B. M. Heuvelink, M. Ruiperez Gonzalez, M. Kilibarda, A. Blagotić*, et al.*, "SoilGrids250m: Global gridded soil information based on machine learning," *PLOS ONE,* vol. 12, p. e0169748, 2017.

[18] M. N. Wright and A. Ziegler, "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R," *Journal of Statistical Software; Vol 1, Issue 1 (2017),* 2017.

[19] S. Janitza, E. Celik, and A.-L. Boulesteix, "A computationally fast variable importance test for random forests for high-dimensional data," *Advances in Data Analysis and Classification,* 2016.

[20] B. M. Greenwell, "pdp: An R package for constructing partial dependence plots," *R Journal,* vol. 9, pp. 421 – 436, 2017.

[21] J. E. Cairns, J. Hellin, K. Sonder, J. L. Araus, J. F. MacRobert, C. Thierfelder*, et al.*, "Adapting maize production to climate change in sub-Saharan Africa," *Food Security,* vol. 5, pp. 345-360, 2013.

[22] A. R. Ngwira, J. B. Aune, and C. Thierfelder, "DSSAT modelling of conservation agriculture maize response to climate change in Malawi," *Soil and Tillage Research,* vol. 143, pp. 85-94, 2014.

[23] I. Nyagumbo, S. Mkuhlani, W. Mupangwa, and D. Rodriguez, "Planting date and yield benefits from conservation agriculture practices across Southern Africa," *Agricultural Systems,* vol. 150, pp. 21-33, 2017.