Crop Yield Mapping with ARC using only Optical Remote Sensing

Philip E. Lewis[1], Feng Yin[1], Jose Luis Gómez-Dans[2], Thomas Weiß[3,4], Elhadi Adam[5]

[1] Dept. Geog., and NCEO, UCL, Gower Street, London, WC1E 6BT UK - p.lewis@ucl.ac.uk
[2] Dept. Geog., KCL, 40 Bush House (NE Wing, Aldwych, London WC2B 4BG, UK - jose.gomez-dans@kcl.ac.uk
[3] Fraunhofer Inst. Comp. Gr. Res. IGD, J.-Junigus-Straße 11, Rostock, 18059, Germany - thomas.weiss@igd-r.fraunhofer.de
[4] Faculty of Agri. and Env. Sc., Univ. Rostock, Justus-von-Liebig-Weg 6, Rostock, 18059, Germany
[5] Univ. Wits, Geog., Arch. and Env. Stud., E Campus, Wit 2050, Johannesburg, South Africa - elhadi.adam@wits.ac.za

Keywords: Remote Sensing, Radiative Transfer, Big Data, Crop Monitoring, Crop Yield, Sentinel-2 MSI, Biophysical Parameters

Abstract

ARC is a new method to generates time series of a full set of biophysical parameters derived from optical EO. Here, we examine relationships between this ‘full’ set and maize yield. 15 Parameters per pixel are estimated over the US corn belt using ARC, to fully describe the phenology, soil, and crop status over time for typical behaviour. ARC is tested for a new model over an area of irrigated and rain-fed winter crop in South Africa. We find that care must be taken for episodic events, and robust filtering methods should be developed for ARC, but average magnitude and timing is well-expressed. We find that a robust yield model (over time and space) can be created at the county-level for maize using only EO parameters with RMSE of 704-938 kg/ha using a non-linear model, but the results are only slightly poorer if a linear model is used. It compares well to a model that also includes weather data, showing that a model can be driven by optical EO data alone.

1. Introduction

1.1 The Remote Sensing Problem for Crop Monitoring

The remote sensing (RS) problem for crops involves estimating a (desired) set of crop descriptors $\mathcal{D}$ (e.g. crop yield, water use) from measurements $\mathbf{y}$ and other information $\mathcal{W}$. This latter may include crop growth models, meteorological data and/or predictions, and other environmental factors (Desloires et al., 2023). The observations are physical measurements, that themselves depend only on some set of intrinsic (bio)physical parameters, $\mathcal{C}$. So only $\mathcal{C}$ is *directly* accessible from $\mathbf{y}$. The ‘Physical‘ RS problem involves estimating $\mathcal{C}$ from $\mathbf{y}$, with models that relate $\mathcal{D} = m(\mathcal{C}, \mathcal{W})$ separate processes, whether explicit or not. Here, we ask the questions: how can we obtain wider information on $\mathcal{C}$ from optical EO, and how well can estimates of $\mathcal{C}$ alone be used to estimate $\mathcal{D}$ (yield here).

Most approaches to crop yield estimation from Earth Observation (EO) use time-series of vegetation indices (VIs), often with smoothing of low-order function-fitting to give interpolation and/or reduce ‘noise’ (Zhang et al., 2003; Roy and Yan, 2020). Relationships are calibrated between the implied phenology (timing and/or mostly peak VI magnitude or a time-integral of VI) and yield, broadly similar to Becker-Reshef et al. (2010), mostly with some additional environmental data such as temperature and rainfall or soil moisture. Temperature is sometimes used to normalise the phenological time in Growing Degree Days (GDD) (Skakun et al., 2019). Empirical relationships vary from simple linear regression (Bolton and Friedl, 2013; Johnson et al., 2021) to machine learning mappings (Luo et al., 2022), though the advantage of the latter can be limited. A smaller number of studies, e.g. (Skakun et al., 2019) use biophsyical parameters (mainly Leaf Area Index (LAI)) or reflectance data directly rather than VIs to achieve similar or better results for a more physical basis (Baez-Gonzalez et al., 2005; Lambert et al., 2018). Some studies use mechanistic crop growth models explicitly through data assimilation (DA) to model other crop processes in $\mathcal{D}$ and/or calibrate for climate change studies, but tend to be limited to EO information on LAI (Dorigo et al., 2007; Machwitz et al., 2014; Huang et al., 2019; Dokoohaki et al., 2022). Growth models can also be used to calibrate empirical relationships with yield (Lobell et al., 2015). ‘Good’ results in this area can be as high as the root mean square error (RMSE) of 200 kg/ha in yield for a single year calibration, but more typically, around 800-900 kg/ha (Khan et al., 2023), with the coefficient of determination $R^2 \approx 0.75 - 0.80$ (e.g. (Skakun et al., 2019; Luo et al., 2022)), and depend on method and scale.

In the era of Analysis-Ready Data (ARD) (Frantz, 2019), $\mathbf{y}$ is readily available as level-2 surface reflectance. For crop monitoring, we can usefully split $\mathcal{C}$ into $(\mathcal{P}, \mathcal{S})$, with $\mathcal{P}$ and $\mathcal{S}$ the properties of the target canopy and underlying soil, respectively, and define $\mathbf{y} = f((\mathcal{P}, \mathcal{S})) + \varepsilon_i$ with a radiative transfer model (RTM) $f()$, with approximations and generalisations that give random uncertainty ε_i. In the optical domain, for measurements from sensors such as Sentinel-2 (S2) MSI (Drusch et al., 2012) or Planet, RTMs such as PROSAIL (Jacquemoud et al., 2009) are readily available to fulfil this role for crops and are widely used. Recently, more ARD with finer spatial, temporal and spectral sampling is becoming readily available, allowing more comprehensive explorations and links to growth processes and ultimately yield. But how much information on yield is contained within the optical EO data themselves? On the one hand, researchers such as Waldner et al. (2019) or Skakun et al. (2019) show that useful yield mapping may be obtained using only limited EO data (normalised in time to GDD, using external temperature data), but others such as (Desloires et al., 2023) suggest that reliance on (optical) EO data alone will always be inadequate as biophysical parameters or VIs cannot express the full impacts on yield. No previous studies look at using the ‘full’ set of biophysical parameters $\mathcal{C}$, likely because of the perceived difficulty in reliably extracting such information from the observations due to parameter coupling and limited but still relevant sensitivity of some parameters causing confounding effects.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-3-2024-199-2024 | © Author(s) 2024. CC BY 4.0 License.

1.2 Archetypes: A New Hope

Recently, Yin et al. (2024) presented an approach to the interpretation of EO data for crops that inverts time series of observations to an empirical parametric model describing the time-course of $\mathcal{P}$ and call this an ‘archetype’ model. It is developed from a ‘big data’ analysis involving the interpretation of large numbers of pixels through inverse radiative transfer model mappings. An algorithm solving for estimates of a full set of $\mathcal{P}$ over time and $\mathcal{S}$ is called ARC, with code and datasets made available (Yin and Lewis, 2023, 2024). ARC is initialised by a broad *a priori* distribution to model the temporal trajectory of its 15 biophysical parameters. From this, candidate estimates of surface reflectance are generated for known satellite observation times and configurations. The resulting ensemble is matched to the observations from any optical sensor operating in the 400 to 2500 nm regime (characteristics of the underlying PROSAIL RTM (Jacquemoud et al., 2009)).

The *a posteriori* estimation of the 15 parameters gives a ‘full’ description of the scene (*wrt* the optical domain), which can be seen as a statement of $\mathcal{P}$ in the context above. Here, $\mathcal{P}$ ”compresses” or summarises the complete spectral, temporal and angular information about the canopy contained in the observations (on top of the prior). This statement is made within the context of a particular RTM, so can only be valid as far that RTM and its parameterisation allows a ‘full’ description of radiative transfer within the defined domain. But the approach allows an examination of the optical EO and yield relationship with the lens of a small set of physically meaningful parameters that completely define the observations.

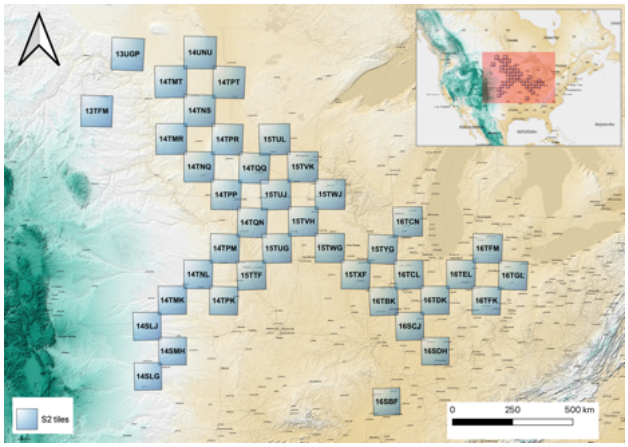

Figure 1. S2 tiles over USA used in training of archetypes

1.3 Aim and Objectives

The aim of the study is to examine the role of the ‘full’ set of ARC crop biophysical parameters in predicting crop yield. We set two objectives to achieve this: (i) give further validation of ARC to cover both irrigated and rain-fed crops; (ii) apply ARC to crop yield mapping and compare the impact of using $\mathcal{C}$ alone or in combination with environmental factors $\mathcal{W}$.

1.3.1 ARC for Winter Wheat, using Planet data: Whilst there is validation of ARC using the maize model at a site in Germany using S2 data in Yin et al. (2024), more areas and sensor configurations need to be tested to understand the applicability and reliability of the approach. Here, we develop a Winter Wheat ARC model using a big data analysis over the USA and validate the model using LAI and C_{ab} time-series data over one irrigated and one rain-fed field in South Africa for validation. We drive ARC with 4 waveband Planet data.

1.3.2 ARC model parameter for maize yield estimation: Since we suppose the ARC parameters to supply a ‘full’ description of the vegetation canopy and its dynamics, mappings between these and crop yield should allow us to better understand the applicability of the ‘full’ information content of optical EO data to crop yield estimation. We know from previous empirical models that maximum and integral LAI (or its surrogate, VI) are related to cereal yield, as does phenological information. These are directly parameters of ARC, alongside descriptions of dry matter, chlorophyll, leaf angle, etc. This second part of the study will examine the relationship between this fuller set of parameters, environmental factors, and crop yield at the county level (where we have reliable statistics over the USA).

2. Data and Method

2.1 Development of Archetype model for Winter Wheat

Yin et al. (2024) applied their maize model to a range of crops and found that it mostly suffices for other crops. But, whilst there is likely some value in using a generic archetype set over a range of crops, a more focused representation can also be developed for specific crops. We apply their approach to develop a Winter Wheat archetype with training data from the USA. The archetypes in ARC are developed for average conditions, so it is of interest to find any issues when dealing with crops with suffer from episodic events. To this end, we gathered a dataset over sites in South Africa for irrigated and rain-fed Winter Wheat crops. We are further interested in exploring if we can drive ARC with lower information content observations, so we applied Planet data in this part of the study. We process S2 reflectance data for the year 2019, masked for Winter Wheat using the NASS Crop Data Layer (CDL) (Boryan et al., 2011) over the USA tiles shown in Figure 1. Inverse emulators using Artificial Neural Nets (ANNs) are trained using the PROSAIL RTM and the parameters in $\mathcal{P}$ to give estimates of each biophysical parameter given in Table 1. The resultant time series are normalised with a double logistic model (4 time-based parameters) to a common temporal framework following (Yin et al., 2024), who show that the mean value for each sample in normalised time gives rise to an optimal first-order scaling model. Figure 2 shows the development of a Winter Wheat archetype. Archetypes are developed as averages from this (results below). A set of 4 phenological parameters $\mathcal{H} = (n_1, n_2, m_1, m_2)$ describe the time-mapping of the archetypes, and we have 8 parameters for $\mathcal{P}$ as the model of the canopy state development.

2.2 ARC Solver

The ARC solver uses a Monte Carlo sampler to generate random values over broad *a priori* distributions of (4) parameters $\mathcal{H}$ and (7) scaling parameters of $\mathcal{P}$ (see Table 1), along with 4 parameters describing a constant soil reflectance function ($\mathcal{S}$ =(soil brightness B, soil shape parameter #1 ϕ, soil shape parameter #2 λ, soil moisture SM_p)), giving a total of 15 parameters to estimate. These are used with a PROSAIL emulator to generate time courses of reflectance for the target sensor (S2 here), as shown in Figure 3. The reflectance trajectories are matched against observations from the target sensor, and the set filtered to give *a posteriori* distributions of all parameters (Figure 3, bottom). With the priors alone, ARC should be capable

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-3-2024-199-2024 | © Author(s) 2024. CC BY 4.0 License.

of simulating *any typical* crop scenario observations at optical wavelengths permissible with PROSAIL over the course of a growing season (see top panels in Figure 3 for S2 wavebands and related indices). We can think of this as the envelope of trajectories in the top panels in Figure 3 that need to encompass the full range of reflectance variation of all *typical* crop scenarios we might encounter in this area. The addition of the information in the observations should narrow this envelope to more refined solution of the physical RS problem for the site under study (bottom panels in Figure 3). How (un)certain that is depends on the information content of the observations, but we can state that the solution uses the ‘full information content’ of the data and prior, in that it cannot affect it more than this. The solution is a estimate of 15 parameters, 7 of which are related to canopy parameters $\mathcal{P}$ that can be verified with ground measurements, 4 of which describe phenology $\mathcal{H}$ and 4 with the optical soil properties $\mathcal{S}$. The approach is not tied to any particular optical sensor, although Yin et al. (2024) drive their study with S2 data and provide validation of most parameters relative to time series ground measurements at a site in Germany.

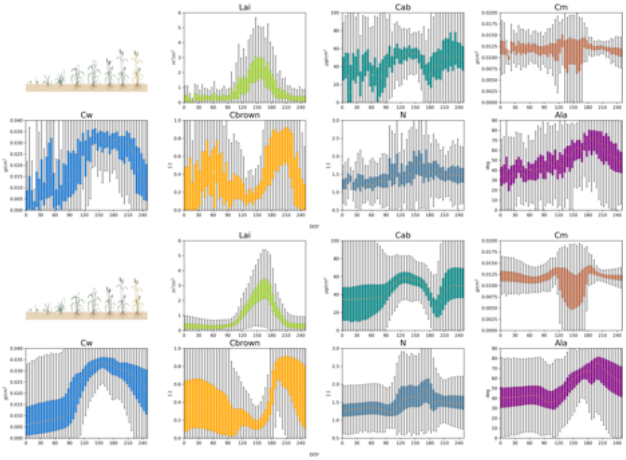

Figure 2. $\mathcal{P}$ from inverse emulators for Winter Wheat. Top: time-series of $\mathcal{P}$, every 5 days, with $25^{th} - 75^{th}$ centiles shown in colour, with black lines showing range. Bottom: same data, but time-normalised to display clearer synchronised patterns.

Symbol in $\mathcal{P}$	Description	Range
N	mesophyll structure coeficient	1-3 (1-3)
C_{ab}	chlorophyll a and b concentration	0-120 (20-80) $\mu g/cm^2$
C_m	leaf dry matter per unit leaf area	0-0.02 (0.001-0.040) g/cm^2
C_w	leaf equivalent water thickness	0.000-0.060 (0.001-0.100) g/cm^2
C_{brown}	brown pigment content	0-1 (0-1)
LAI	leaf area index	0.0-8.0 (0.1-8.0) m^2/m^2
ALA	average leaf inclination angle	0-90 (45-80) $^\circ$

Table 1. Parameters of PROSAIL in $\mathcal{P}$ and related unit and bound information in training (modelling).

2.3 Winter Wheat Archetype

Data for Winter Wheat LAI and C_{ab} were collected for two seasons in South Africa. The field for the 2022 measurements is close to the town of Villiers and was a circular irrigated wheat

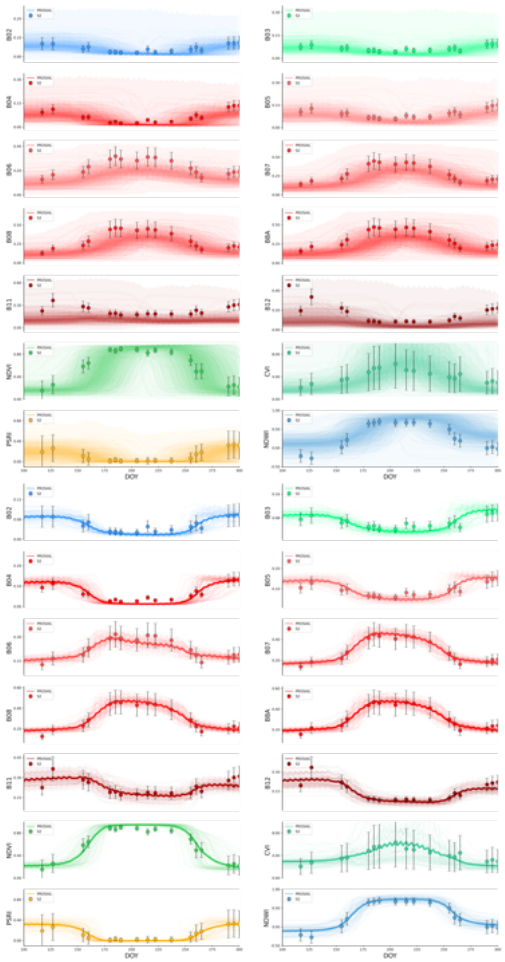

Figure 3. *a priori* (top) and *a posteriori* distributions (bottom) over S2 bands and related indices

field, planted around the start of September 2022. Data for 2023 were collected at near Meets, Bethlehem, South Africa. The study field was rain-fed wheat, planted in early September 2023. Harvesting for the crops spanned from the end of November to early December. LAI and C_{ab} were measured at each of the 11(9) sample times in 2022 (2023). Planet PS2.SD data were collected on the sites for the growing seasons of 2022 and 2023, with 87 clear images for 2023 and 97 for 2022, available as cross-calibrated S2 MSI reflectance, making it straightforward to use the ARC code (calibrated for S2). We used Planet bands 2, 4, 6, 7 and 8 as surrogates for S2 bands 2, 3, 4, 5 and 8a. A 10% uncertainty was assumed for Planet surface reflectance.

Figure 4 shows the red and near infrared (NIR) reflectance over the 2023 sample field, with an estimate of soil moisture for the site. The crop suffered a significant water deficit during the green-up period, which is probably a factor affecting the final LAI (around half the LAI of the irrigated crop). Since the canopy cover for this field is quite low then, the crop reflectance maintains a sensitivity to soil moisture throughout the season, although any more severe drought may also affect the leaf reflectance. We can see this in the NIR Planet data in Figure 4 where some values suddenly drop (below the dashed line shown) when conditions are very dry. We interpret these measurements as ‘affected by low moisture’ and remove them from the main analysis and fitting in ARC. We found that if such anomalous data were included, they could affect the result if they

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-3-2024-199-2024 | © Author(s) 2024. CC BY 4.0 License.

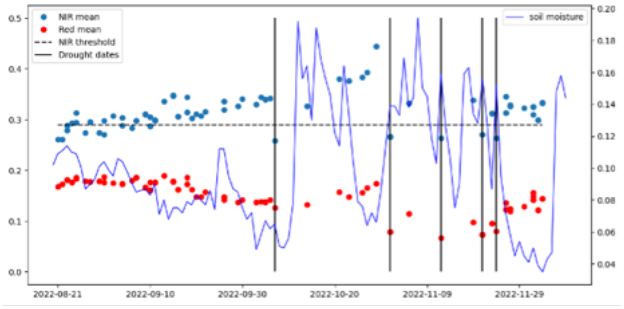

Figure 4. Planet red and NIR reflectance (dots) for 2023, along with soil moisture (blue line)

appeared at critical times in the time series, e.g. in green-up or in times of sparse sampling.

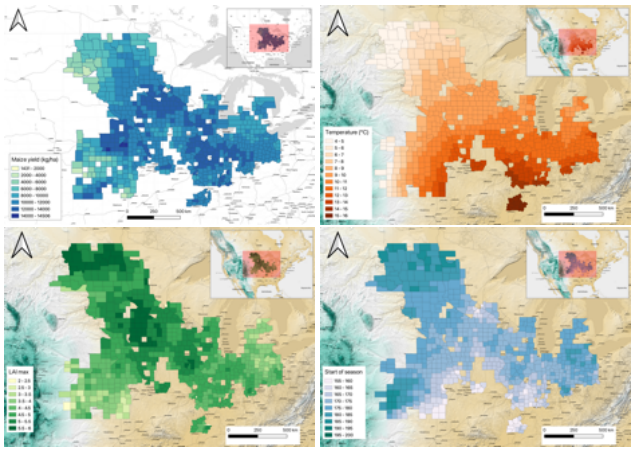

Figure 5. (a) NASS reported yield at the county level 2020, (b) ERA-5 mean air temperature, 2020. The ARC retrieval of (c) LAI max, (d) start of season (SOS) averaged at county level during maize growing season in 2020 over the US Corn Belt.

2.4 USA Maize yield

The county-level yield data for maize in the US from 2017 to 2020 were obtained from NASS. For the subsequent analysis of yield prediction, only counties that are more than 50% covered by the 122 sampled S2 tiles are included, resulting in a total of 2,297 county yield samples.

2.5 ERA-5 weather data

The ECMWF ERA-5-Land reanalysis dataset (CAMS, 2019) was obtained from Google Earth Engine (GEE) (Gorelick et al., 2017) from April to November. The variables acquired from GEE are specified in Table 2 and the GEE operations used to convert the hourly data to the daily data are also listed in the table.

Daily variable name	Hourly ERA variable name	GEE Operation
Daily total solar radiation	Surface solar radiation downwards	Sum
Daily mean dewpoint temperature	Dewpoint temperature at 2 m	Mean
Daily maximum temperature	Temperature at 2 m	Max
Daily minimum temperature	Temperature at 2 m	Min
Daily total precipitation	total precipitation	Sum
Daily mean soil water	Volumetric soil water layer 1	Mean

Table 2. Variables from ERA-5-Land reanalysis dataset used, and the operations used to compute the daily variables.

2.6 Bio-physical parameters over USA

ARC was run for 2017 to 2020 over the S2 tiles covering the USA Corn Belt to give estimates of the 15 model parameters as input to the yield study. Samples were selected for each county for the maize crop as indicated in the NASS CDL.

Examples of the retrieved parameters for 2020 are given in Figure 5. Figure 5(c) shows the variations of the maximum LAI over different counties that broadly correspond to the yield maps shown in Figure 5 (a), with higher LAI values in the central part of the Corn Belt and lower on the west and east sides. However, in Figure 5 (a) we see a lower yield in the northern part of the Corn Belt, even though the LAI is high there. This suggests that other variables are at play here. The start of season parameter shown in Figure 5(d) shows a strong correspondence to the temperature variation over the area (Figure 5(d)).

2.7 County-level yield estimation

We first examine the influence of the various factors in the full set $(\mathcal{C}, \mathcal{W})$ on yield. To provide points of comparison to the modelling with $\mathcal{D} = m(\mathcal{C})$, the ARC parameter set alone, we first explored mapping the crop yield as $\mathcal{D} = m(\mathcal{C}, \mathcal{W})$. We applied two response functions to empirical modelling: (i) linear Ordinary Least-Squares (OLS), applied to normalised values; and (ii) the non-linear ensemble tree-based Extra Trees method using the linear OLS and ExtraTreesRegressor methods inside the scikit-learn (Varoquaux et al., 2015) Python package. A stratified sampling scheme is used to split the training and validation sets between all counties, where county yields are equally divided into 20 categories from low to high. Then 70% of the samples are randomly drawn from each category for the training, and the remaining 30% used for validation. The input to the regressors are variously $\mathcal{C}$ and/or $\mathcal{W}$, depending on the experiment. To examine the performance of the models over time, 6 counties with high yield variance over 4 years are set aside from the sampling and not used in the calibration or main validation.

3. Results

3.1 Winter Wheat Archetype

The upper panel set of Figure 6 shows the archetypes of $\mathcal{P}$ for Winter Wheat from the distributions in Figure 2. We examine both mean and median but find them mostly close so we use the latter as archetype. However, there is some departure in several archetypes at the start of season, which may require further investigation. One feature of the Winter Wheat archetypes is that C_{ab} appears to decrease and then increase toward the end of season, which is different from the maize archetypes developed by Yin et al. (2024). Though present clearly also in Figure 2, this is likely an artifact of the data processing with low LAI but green material appearing in fields at the end of season. This feature is not present in the maize archetype, we notice.

Figure 7 shows the same information as the scatterplots. The regressions for LAI show high values of R (0.88 and 0.91 for 2022 and 2023, respectivley) with relatively small bias and slopes of 1.1 and 1.2 respectively, and low values of standard error (SE) 0.082 in both cases. These are similar to the findings of Yin et al. (2024) when using the maize archetype with S2, and in line with or better than other LAI validation results. Despite this, we can see that the LAI in 2022 (the irrigated crop) is consistently overestimated in the retrievals. The magnitude of

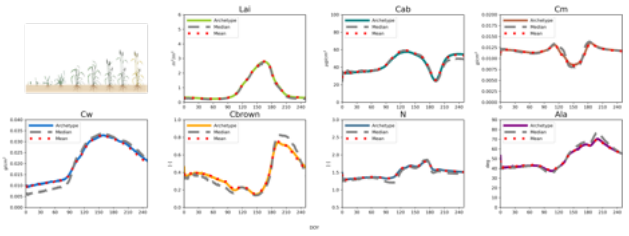

Figure 6. Crop Archetypes developed here for Winter Wheat.

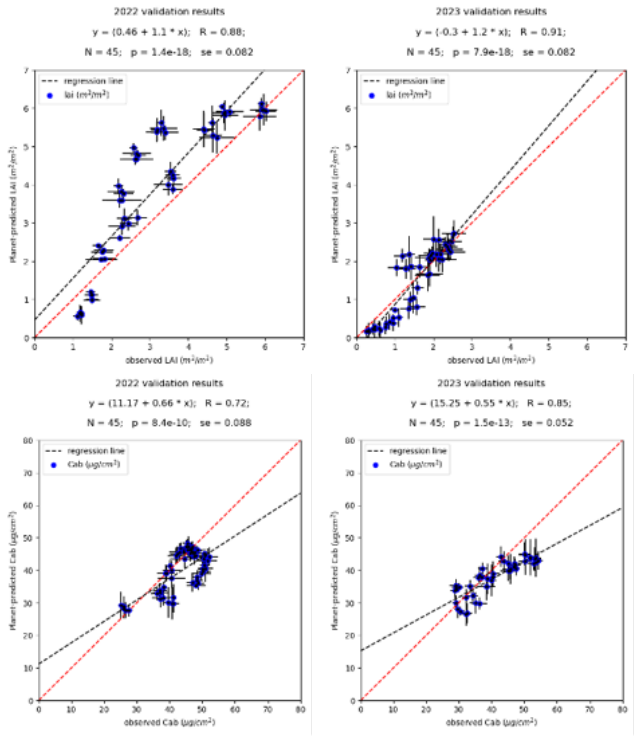

Figure 7. LAI (T) and C_{ab} (B) validation, L: 2022 R: 2023

the retrieved patterns of LAI is overall close to the measured data, for both irrigated and non-irrigated crops, though slightly low in both cases. The pattern of C_{ab} retrieved is slightly low, with a slope of 0.66 and 0.55, respectively, in 2022 and 2023, though the range of values of C_{ab} is rather small in both cases. SE is small in both cases, at 0.088 and 0.052 respectively.

3.2 USA Maize yield

3.2.1 Linear and non-linear yield model with $(\mathcal{C}, \mathcal{W})$: The OLS result is shown in Figure 8 (top), giving an accuracy of 998 kg/ha in RMSE and $R = 0.86$. The OLS coefficients show the significance of the phenology parameters (n_0, n_1), LAI, and solar radiation for the yield prediction. The feature importance and the validation of the ET model against the NASS reported yield are shown in Figure 8. The non-linear regressor provides slightly more accurate estimates of yield with a correlation value of 0.91 and an RMSE value of 812 kg/ha. The most influential parameters for the yield estimation are LAI, temperature, and phenology parameters (n_1, n_0), which is interestingly different from the OLS findings. However, both methods have unsurprisingly highlighted the importance of maximum LAI and phenology parameters.

3.2.2 Yield estimation with $(\mathcal{C})$ or $(\mathcal{W})$: By using weather data $(\mathcal{W})$ only, yield estimation gives R of 0.74 and a RMSE

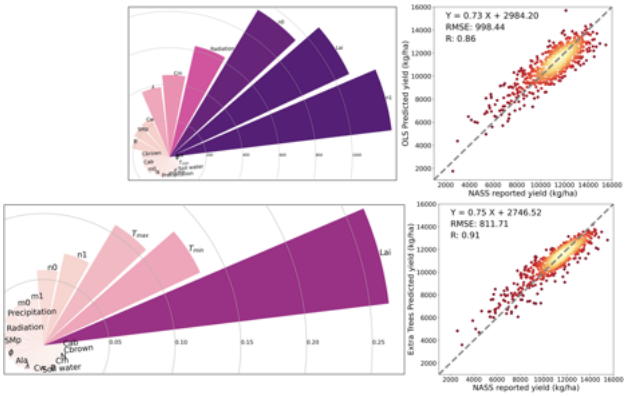

Figure 8. Top: OLS, Bottom: ET, L: feature importance, R: yield validation

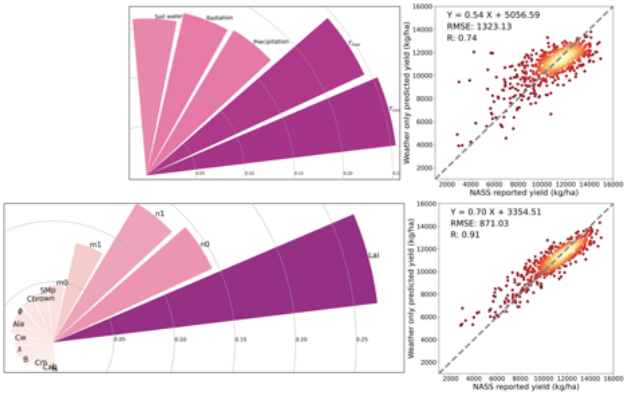

Figure 9. Top: weather $(\mathcal{W})$ only, Bottom: EO $(\mathcal{C})$ only, with ET, L: feature importance, R: yield validation

of 1323 kg/ha (Figure 9, top). This indicates that a coarse description of weather can explain around 54% of maize yield variation over the Corn Belt. This is almost 30% reduction in the explained variance of the yield compared to the EO plus the weather prediction shown in Figure 8. The temperature is identified as having the highest feature importance, which agrees with previous results in Section 2.7. Precipitation ranked 3rd in the feature importance, probably because a large proportion of maize planted within the Corn Belt are non-irrigated (Grassini et al., 2009; Hunt et al., 2020).

Yield estimation with $(\mathcal{C})$ only, the EO-derived variables, is shown in Figure 9. It achieves similar level of accuracy as the results by using the EO and weather combined, where the correlation value is the same, but with slightly higher RMSE (871 kg/ha). The performance of the yield estimates over different years is shown in Figure 10, where there is little difference between those four years, indicating the model can provide consistent yield estimation regardless of the observation year.

The trained linear OLS model and non-linear ET model performance in yield estimation is shown for sample 6 counties in Figure 11, top and bottom, respectively. The OLS yield estimates for these 6 counties, treated independently of the training samples, show good agreement with the reported yield values (see Figure 11 for details). Compared to the linear OLS model, the trained ET model shows an improved statistic with a higher R value of 0.95 and smaller RMSE value of 946 kg/ha. However, there is a larger bias in the yield estimation from the ET model with a slope of 0.78 and an intercept of 2524 compared to 0.83 and 1522 from the OLS model estimation. The trained

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-3-2024-199-2024 | © Author(s) 2024. CC BY 4.0 License.

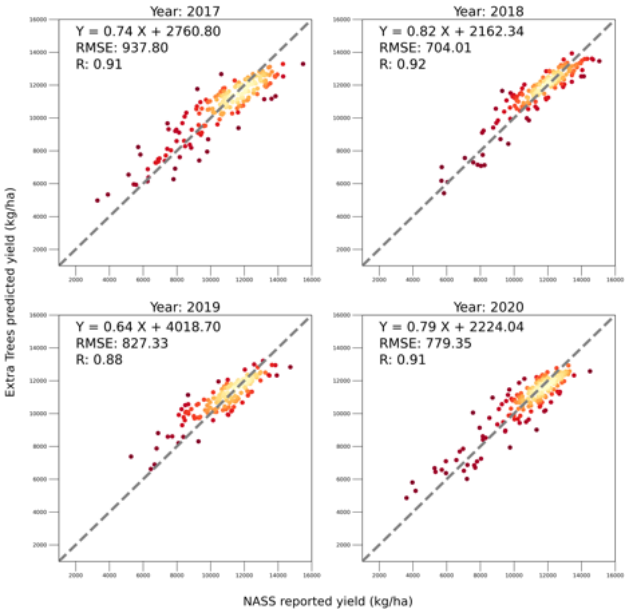

Figure 10. Maize yield estimations by only using EO derived variables with Extra Trees method compared to NASS reported yield for different years.

ET model can also capture the yield variation well over those four years for those six counties.

4. Discussion

4.1 Winter Wheat Archetype testing

In this study, we develop a new ARC model specifically for Winter Wheat and test this using Planet data against ground observations over irrigated and rain-fed conditions. The ARC model deals well with average conditions, and if we filter anomalous samples from episodic events such as very low canopy and/or soil moisture, it is able to describe the overall conditions and trajectories reasonably well in terms of magnitude, but with some inconsistencies in the details at particular points in the season. We have not examined the value of using the Winter Wheat ARC model over using a generic (maize) model as in the previous study, but that would be an interesting adjunct. The poorer performance of model fitting in the early part of the season may be due to assumptions about horizontal homogeneity in the RTM not being valid for a row crop at that time and deserves further investigation. It may be ameliorated by looking closer at both mean and median archetypes. It is also possible that the apparent ‘over-smoothing’ here arises from treatment of the ensemble, given that the ‘shape’ of the measured LAI trajectory seems closer to the raw LAI archetype ‘shape’ in Figure 6. That said, the magnitude of the parameters examined seems well-captured, which is important for the yield monitoring.

The need to filter anomalous samples from areas that suffer episodic impacts is an important learning experience for future development of ARC, and we believe that such filtering should generally be considered to deal with such as this. The very appearance of such data that do not conform to the average conditions of ARC is also interesting and may suggest that they contain additional information content beyond the ARC parameters. Such anomalies may be hard to pick up if we use less dense time series than the Planet data examined here, and

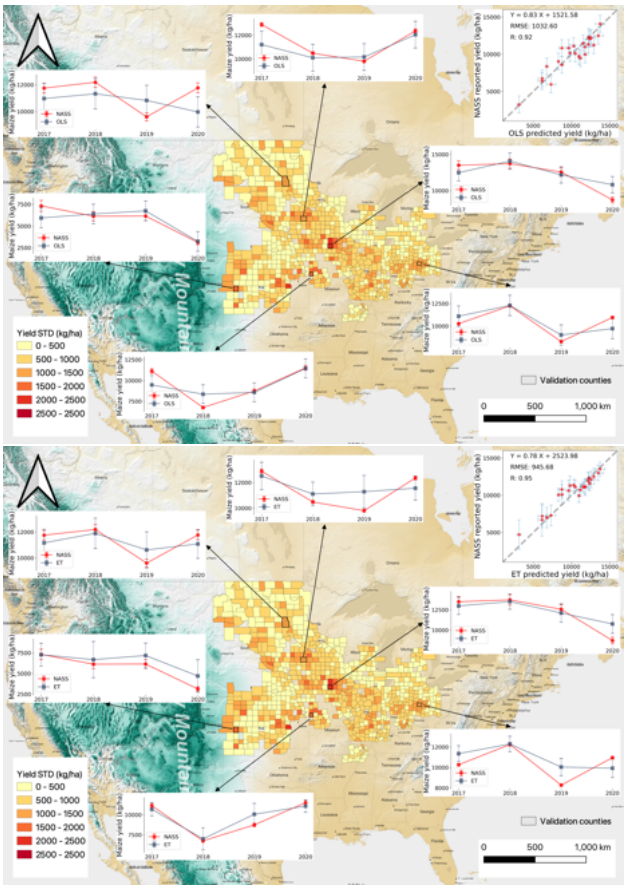

Figure 11. Linear OLS (top) and nonlinear ET (bottom) yield estimation for six counties over Corn Belt from 2017 to 2020 with only the EO information. The standard deviation of yields over four years time is also shown for all the counties over the Corn Belt.

a combination of S2 and Planet observations would be worth exploring.

4.1.1 Yield estimation with $(\mathcal{C})$ or $(\mathcal{W})$: The experiments here mimic other studies that show that non-linear empirical models outperform linear models, though the gain is generally only slight. We are able to form models using canopy and environmental data that act similarly to those from other studies, with temperature, then precipitation, and radiation the ordering of environmental factors and maximum LAI and phenological parameters mostly showing as the most informative EO data. The influence of further parameters in $\mathcal{P}$ seems to be of second-order importance, although their inclusion is shown to add information and help refine the models.

The similar performance of $(\mathcal{W})$-only compared to $(\mathcal{P}, \mathcal{W})$ models indicates that most of the information from the weather variables is already incorporated within the EO signals, i.e. the meteorological inputs $(\mathcal{W})$ had impacted $(\mathcal{C})$ and contains little additional information for yield estimation (but note that the meteorological variables used in the modelling are temporal aggregates over the entire growing season). This agrees broadly with the finding of Waldner et al. (2019) who found that the temporally resolved LAI profile has already reflected the majority of the influence of weather on crop yield. LAI magnitude and phenology parameters (n_0, n_1, m_0, m_1) can alone explain more than 60% of the yield variation. It is interesting to notice that the canopy structure parameter ALA is more important

for yield estimation than the leaf colour parameters C_{ab}, while the leaf colour during the senescence period C_{brown} shows a greater influence on yield estimation.

5. Conclusions

In this study, we have used the ARC model to examine the role a ‘full’ set of crop biophysical parameters in predicting crop yield. We find it straightforward to develop new bespoke ARC models for crops such as Winter Wheat, but the resulting model is quite similar in most ways to a previously generated maize model. We find we can drive the process with more limited information content Planet observations, will have important ramifications for mapping data from smaller fields and in cloudy areas, because of the higher spatial resolution and temporal sampling available. We find that the parameters currently most widely-used (explicitly or implicitly), LAI and phenology are confirmed as the first-order EO parameters. However, second-order parameters such as ALA can also add information. Leaf colour parameters tend to have only low amounts of information with respect to yield. A useful EO-only maize yield estimation model has been calibrated using the ARC parameters only and is seen to operate well over multiple years in the US corn belt.

.

Acknowledgements

FY, PL and JGD thank the following for financial support: NERC NCEO (PR140015); STFC UK-Newton Agritech Programme(533651). PL and FY were also supported by NERC-BBSRC AgZero^{+} (NE/W005050/1). PL, FY and EA were funded from the UCL-Wits University Strategic Partnership and the ESA and African Union project ‘BIG data archetypes for crops from EO’ for fieldwork. FY thanks the NERC CPEO programme, project reference NE/X006328/1 for support. The authors thank NERC for access to the JASMIN computing facility, widely used in this study and Planet for the use of data.

References

Baez-Gonzalez, A. D., Kiniry, J. R., Maas, S. J., Tiscareno, M. L., Macias C, J., Mendoza, J. L., Richardson, C. W., Salinas G, J., Manjarrez, J. R., 2005. Large-area maize yield forecasting using leaf area index based yield model. *Agron. J.*, 97(2), 418–425.

Becker-Reshef, I., Justice, C., Sullivan, M., Vermote, E., Tucker, C., Anyamba, A., Small, J., Pak, E., Masuoka, E., Schmaltz, J. et al., 2010. Monitoring global croplands with coarse resolution earth observations: The Global Agriculture Monitoring (GLAM) project. *Rem. Sens.*, 2(6), 1589–1609.

Bolton, D. K., Friedl, M. A., 2013. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. *Agric. For. Met.*, 173, 74-84.

Boryan, C., Yang, Z., Mueller, R., Craig, M., 2011. Monitoring US agriculture: the US Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer Program. *Geocarto International*, 26(5), 341–358.

CAMS, 2019. ERA5-Land hourly data from 1981 to present. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*, 10.

Desloires, J., Ienco, D., Botrel, A., 2023. Out-of-year corn yield prediction at field-scale using Sentinel-2 satellite imagery and machine learning methods. *Computers and Electronics in Agriculture*, 209, 107807.

Dokoohaki, H., Rai, T., Kivi, M., Lewis, P., Gómez-Dans, J. L., Yin, F., 2022. Linking Remote Sensing with APSIM through Emulation and Bayesian Optimization to Improve Yield Prediction. *Rem. Sens.*, 14(21), 5389.

Dorigo, W., Zurita-Milla, R., de Wit, A., Brazile, J., Singh, R., Schaepman, M., 2007. A review on reflective remote sensing and data assimilation techniques for enhanced agroecosystem modeling. *Int. J. App. EO and Geoinf.*, 9(2), 165 - 193.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., Bargellini, P., 2012. Sentinel-2: ESA’s Optical High-Resolution Mission for GMES Operational Services. *Rem. Sens. Env.*, 120, 25-36.

Frantz, D., 2019. FORCE—Landsat + Sentinel-2 Analysis Ready Data and Beyond. *Rem. Sens.*, 11(9).

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Rem. Sens. Env.*, 202, 18–27.

Grassini, P., Yang, H., Cassman, K. G., 2009. Limits to maize productivity in Western Corn-Belt: A simulation analysis for fully irrigated and rainfed conditions. *Agric. For. Met.*, 149(8), 1254–1265.

Huang, J., Gómez-Dans, J. L., Huang, H., Ma, H., Wu, Q., Lewis, P. E., Liang, S., Chen, Z., Xue, J.-H., Wu, Y., Zhao, F., Wang, J., Xie, X., 2019. Assimilation of remote sensing into crop growth models: Current status and perspectives. *Agric. For. Met.*, 276-277, 107609.

Hunt, E. D., Birge, H. E., Laingen, C., Licht, M. A., McMechan, J., Baule, W., Connor, T., 2020. A perspective on changes across the U.S. Corn Belt. *Env. Res. Let.*, 15(7), 071001.

Jacquemoud, S., Verhoef, W., Baret, F., Bacour, C., Zarco-Tejada, P. J., Asner, G. P., François, C., Ustin, S. L., 2009. PROSPECT+ SAIL models: A review of use for vegetation characterization. *Rem. Sens. Env.*, 113, S56–S66.

Johnson, D. M., Rosales, A., Mueller, R., Reynolds, C., Frantz, R., Anyamba, A., Pak, E., Tucker, C., 2021. USA Crop Yield Estimation with MODIS NDVI: Are Remotely Sensed Models Better than Simple Trend Analyses? *Rem. Sens.*, 13(21).

Khan, S. N., Khan, A. N., Tariq, A., Lu, L., Malik, N. A., Umair, M., Hatamleh, W. A., Zawaideh, F. H., 2023. County-level corn yield prediction using supervised machine learning. *Eur. J. Rem. Sens.*, 56(1), 2253985.

Lambert, M.-J., Traoré, P. C. S., Blaes, X., Baret, P., Defourny, P., 2018. Estimating smallholder crops production at village level from Sentinel-2 time series in Mali’s cotton belt. *Rem. Sens. Env.*, 216, 647–657.

Lobell, D. B., Thau, D., Seifert, C., Engle, E., Little, B., 2015. A scalable satellite-based crop yield mapper. *Rem. Sens. Env.*, 164, 324–333.

Luo, Y., Zhang, Z., Cao, J., Zhang, L., Zhang, J., Han, J., Zhuang, H., Cheng, F., Tao, F., 2022. Accurately mapping global wheat production system using deep learning algorithms. *Int. J. App. EO and Geoinf.*, 110, 102823.

Machwitz, M., Giustarini, L., Bossung, C., Frantz, D., Schlerf, M., Lilienthal, H., Wandera, L., Matgen, P., Hoffmann, L., Udelhoven, T., 2014. Enhanced biomass prediction by assimilating satellite data into a crop growth model. *Env. mod. & soft.*, 62, 437–453.

Roy, D., Yan, L., 2020. Robust Landsat-based crop time series modelling. *Rem. Sens. Env.*, 238, 110810. Time Series Analysis with High Spatial Resolution Imagery.

Skakun, S., Vermote, E., Franch, B., Roger, J.-C., Kussul, N., Ju, J., Masek, J., 2019. Winter Wheat Yield Assessment from Landsat 8 and Sentinel-2 Data: Incorporating Surface Reflectance, Through Phenological Fitting, into Regression Yield Models. *Rem. Sens.*, 11(15).

Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F., Mueller, A., 2015. Scikit-learn. *GetMobile: Mobile Computing and Communications*, 19(1), 29–33.

Waldner, F., Horan, H., Chen, Y., Hochman, Z., 2019. High temporal resolution of leaf area data improves empirical estimation of grain yield. *Scientific Reports*, 9(1), 1–14.

Yin, F., Lewis, P., 2023. Archetypes: Arc software.

Yin, F., Lewis, P., 2024. Eo-africa: Eo africa d3.

Yin, F., Lewis, P. E., Gómez-Dans, J. L., Weiß, T., 2024. Archetypal Crop Trait Dynamics for Enhanced Retrieval of Biophysical Parameters from Sentinel-2 MSI. *Rem. Sens. Env.*, In Review, Oct 2023.

Zhang, X., Friedl, M. A., Schaaf, C. B., Strahler, A. H., Hodges, J. C., Gao, F., Reed, B. C., Huete, A., 2003. Monitoring vegetation phenology using MODIS. *Rem. Sens. Env.*, 84(3), 471-475.