ENVIRONMENTAL RESEARCH
LETTERS

LETTER • OPEN ACCESS

A scalable crop yield estimation framework based on remote sensing of solar-induced chlorophyll fluorescence (SIF)

To cite this article: Oz Kira *et al* 2024 *Environ. Res. Lett.* **19** 044071

ENVIRONMENTAL RESEARCH
LETTERS

LETTER

A scalable crop yield estimation framework based on remote sensing of solar-induced chlorophyll fluorescence (SIF)

Oz Kira[1,2,3,*], **Jiaming Wen**[3], **Jimei Han**[3], **Andrew J McDonald**[3,4], **Christopher B Barrett**[5], **Ariel Ortiz-Bobea**[5], **Yanyan Liu**[6], **Liangzhi You**[6], **Nathaniel D Mueller**[7,8] and **Ying Sun**[3,*]

[1] Department of Civil and Environmental Engineering, Ben-Gurion University of the Negev, Beer-Sheva 8410501, Israel
[2] School of Sustainability and Climate Change, Ben-Gurion University of the Negev, Beer-Sheva 8410501, Israel
[3] School of Integrative Plant Science, Soil and Crop Sciences Section, Cornell University, Ithaca, NY 14853-7801, United States of America
[4] School of Integrative Plant Sciences and Department of Global Development, Cornell University, Ithaca, NY 14853-7801, United States of America
[5] Charles H. Dyson School of Applied Economics and Management and Jeb E. Brooks School of Public Policy, Cornell University, Ithaca, NY 14853-7801, United States of America
[6] Department of Transformation Strategies, International Food Policy Research Institute (IFPRI), 1201 I Street, NW, Washington, DC 20005, United States of America
[7] Department of Ecosystem Science and Sustainability at Colorado State University, Fort Collins, CO 80523, United States of America
[8] Department of Soil and Crop Sciences at Colorado State University, Fort Collins, CO 80523, United States of America
[*] Authors to whom any correspondence should be addressed.

E-mail: ozkira@bgu.ac.il and ys776@cornell.edu

Keywords: solar-induced chlorophyll fluorescence (SIF), crop yield, mechanistic light reactions, agricultural monitoring, satellite remote sensing, machine learning

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
19 September 2023

REVISED
31 January 2024

ACCEPTED FOR PUBLICATION
7 March 2024

PUBLISHED
12 April 2024

Abstract

Projected increases in food demand driven by population growth coupled with heightened agricultural vulnerability to climate change jointly pose severe threats to global food security in the coming decades, especially for developing nations. By providing real-time and low-cost observations, satellite remote sensing has been widely employed to estimate crop yield across various scales. Most such efforts are based on statistical approaches that require large amounts of ground measurements for model training/calibration, which may be challenging to obtain on a large scale in developing countries that are most food-insecure and climate-vulnerable. In this paper, we develop a generalizable framework that is mechanism-guided and practically parsimonious for crop yield estimation. We then apply this framework to estimate crop yield for two crops (corn and wheat) in two contrasting regions, the US Corn Belt US-CB, and India's Indo–Gangetic plain Wheat Belt IGP-WB, respectively. This framework is based on the mechanistic light reactions (MLR) model utilizing remotely sensed solar-induced chlorophyll fluorescence (SIF) as a major input. We compared the performance of MLR to two commonly used machine learning (ML) algorithms: artificial neural network and random forest. We found that MLR-SIF has comparable performance to ML algorithms in US-CB, where abundant and high-quality ground measurements of crop yield are routinely available (for model calibration). In IGP-WB, MLR-SIF significantly outperforms ML algorithms. These results demonstrate the potential advantage of MLR-SIF for yield estimation in developing countries where ground truth data is limited in quantity and quality. In addition, high-resolution and crop-specific satellite SIF is crucial for accurate yield estimation. Therefore, harnessing the mechanism-guided MLR-SIF and rapidly growing satellite SIF measurements (with high resolution and crop-specificity) hold promise to enhance food security in developing countries towards more effective responses to food crises, agricultural policies, and more efficient commodity pricing.

1. Introduction

Projected increases in food demand driven by population growth and heightened agricultural vulnerability to climate change pose severe threats to global food security in coming decades, especially for developing nations (Godfray *et al* 2010, Foley *et al* 2011, Lobell *et al* 2011). Efforts to monitor agricultural productivity in real time are increasingly critical to help forecast short-term disruptions to food supply and inform the development of longer-term strategies to enhance climate resilience. Remote sensing holds great promise for estimating crop yields at large scales and low cost (Lobell *et al* 2015). Significant methodological advances in this regard have been developed in recent decades (e.g. Lobell *et al* 2005, 2015, Guan *et al* 2016, Burke and Lobell 2017, Jin *et al* 2019, Peng *et al* 2020). Among them, one approach employs satellite observations of solar-induced chlorophyll fluorescence (SIF) (Guanter *et al* 2014, Guan *et al* 2016, Peng *et al* 2020), an optical signal emitted by chlorophyll upon light absorption and thus carries direct and mechanistic information about photosynthesis (Papageorgiou and Govindjee 2004, Porcar-Castell *et al* 2014, 2021). This mechanistic advantage of SIF combined with its multiple practical benefits over conventional vegetation indices (VIs), including lower sensitivity to thin cloud interference (Frankenberg *et al* 2012), muted sensitivity to the background soil as non-fluorescing targets (Wang *et al* 2019), and less susceptibility to saturation under high leaf area index saturation, promote SIF as a promising tool for large-scale crop yield prediction. For example, studies have demonstrated that satellite SIF is a stronger predictor of net primary production (NPP) than VIs, e.g. enhanced vegetation index (Guan *et al* 2016). However, other studies argued that SIF, measured at the current spatial and temporal resolution, is not better than VIs for predicting yield (Cai *et al* 2019, Peng *et al* 2020, Sloat *et al* 2021). The possible reason for such contrasting conclusions is that the current satellite SIF contains substantial noise with relatively lower spatial and/or temporal resolutions. Nevertheless, SIF, which contains both structural and functional information about plants, not only correlates well with crop productivity but also may offer an early warning for stress onset to inform management practices (Mohammadi *et al* 2022, Sun *et al* 2023b).

On the analytical side, ML algorithms have become a powerful tool in agricultural monitoring with remote sensing observations as input (Peng *et al* 2018, Cai *et al* 2019, Ghazaryan *et al* 2020, Sishodia *et al* 2020, Gastli *et al* 2021, Khalil and Abdullaev 2021, Paudel *et al* 2022). The main advantage of ML is identifying connections between inputs and outputs that mechanistic formulations cannot easily and/or fully depict. However, the right set of inputs and a well-designed calibration process are required to utilize this tool optimally (Chlingaryan *et al* 2018). ML has significant limitations concerning scalability, extrapolation, and generalization (Morais *et al* 2021), meaning that accurate yield prediction is often restricted to certain regions, periods, crop types, management practices, and environmental conditions, although strategies start to emerge to overcome this limitation (Lobell *et al* 2015, Jain *et al* 2017, Jin *et al* 2017, Yang *et al* 2023, Liu *et al* 2024). Moreover, sufficient and high-quality ground truth data must be available for ML model calibration, which is not always the case, especially in developing countries and landscapes with high spatial heterogeneity (Lobell *et al* 2020). Furthermore, future climatic changes could restrict the predictability of ML-type models calibrated against historical data, as such scenarios have not yet been present (for ML model training).

This study pursues a scalable approach to predict crop yields in time and space at the regional scale utilizing high-resolution satellite SIF and a mechanistic light reaction (MLR) model denoted as MLR-SIF (Han *et al* 2022a). MLR-SIF estimates photosynthesis directly from remotely sensed SIF from the perspective of light reactions. The rationale is that SIF is a direct measure of the actual electron transport rate (ETR) from photosystem II to I, linking the light and carbon reactions of photosynthesis (Gu *et al* 2019). If such a mechanism can be sufficiently represented with minimal parameter calibration, it is hopeful to have a scalable approach, i.e. generalizable in space, time, cropping systems, management practices, etc, to estimate crop photosynthesis and yields. Such a model would have minimal dependence on the quantity and quality of ground truth data for model calibration and, therefore, transferrable to regions and/or environmental regimes areas where/when high-quality yield estimates are unavailable, especially in developing countries where food insecurity and socioeconomic vulnerability are most pressing.

Here we apply MLR-SIF to two contrasting settings: corn in the US corn belt (US-CB), and wheat in India's wheat belt in Indo–Gangetic plain (IGP-WB). US-CB provides an ideal testbed, as it not only produces a significant portion of the global supply of corn, along with other field crops, but also has a wealth of high-quality datasets for model evaluation and a homogeneous landscape (Lobell *et al* 2015, Ortiz-Bobea *et al* 2018). IGP-WB is vital because India is the second-largest global wheat producer, supporting 70% of India's rural households (Erenstein and Thorpe 2011). However, IGP-WB is

characterized by small farms and highly heterogeneous landscapes, making reliable yield prediction from satellite datasets difficult. Also, the magnitude of yields, especially in the eastern IGP-WB (i.e. Bihar and eastern districts of Uttar Pradesh), is lower than in the western portion, challenging a reliable and scalable model calibration (Jain *et al* 2017, McDonald *et al* 2022). The contrasting crop types, landscape characteristics, management practices, and data availability/quality in the US-CB and IGP-WB thus offer an excellent opportunity to examine whether MLR-SIF driven by satellite SIF can improve large-scale yield prediction over ML models and whether the improvement holds across diverse agricultural landscapes and management.

2. Data and methods

This study compared three models for crop yield estimation: (1) the mechanistic MLR-SIF model, (2) artificial neural network (ANN), and (3) random forest (RF). MLR-SIF was applied to estimate photosynthesis (section 2.2), which was subsequently used to calculate crop yield following Lobell *et al* (2015) and Guan *et al* (2016). ANN and RF (section 2.3) are currently among the most commonly used ML models for crop-yield estimation in literature (Chlingaryan *et al* 2018), and therefore were chosen here as the baseline to assess the accuracy and scalability of MLR-SIF.

2.1. Study regions and crop yield data

The county-level crop yield in US-CB came from the USDA National Agricultural Statistics Service. We focused on four states in the heart of US-CB (Indiana, Illinois, Iowa, and Nebraska, totaling 210 counties), because they have corn-specific OCO-2 SIF available from previous work (details below). Yield estimation in US-CB was conducted for five years from 2015 to 2020 (when corn-specific OCO-2 SIF is available) except 2017 (when OCO-2 had an instrument failure in August). The district-level wheat yield in IGP-WB came from the District Level Database (DLD) for India (http://data.icrisat.org/dld/), including 55 districts for the states of Bihar, Uttar Pradesh, and Haryana. Yield estimation in IGP-WB was carried out from 2015 to 2017 (the maximum overlap between OCO-2 SIF and yield data).

2.2. The MLR-SIF yield estimation framework

The MLR-SIF based framework for yield estimation consists of three steps. First, it estimates photosynthesis (or gross primary production GPP) taking observational SIF as input (Gu *et al* 2019, Han *et al* 2022a) (equations (1)–(3)). Next, NPP was computed from GPP (excluding autotrophic respiration) using the NPP/GPP ratio derived from MODIS products (data products described in section 2.4). Finally, crop yield was estimated from NPP and harvest index (HI) along with other crop-type specific parameters obtained from the literature (equation (4)). The last two steps follow Lobell *et al* (2002) and Guan *et al* (2016), while the main novelty of this paper lies in the first step that calculates GPP (μmol CO_2 m^{-2} s^{-1}) from MLR-SIF.

$$\mathrm{GPP} = \begin{cases} \dfrac{C_i - \Gamma^*}{4C_i + 8\Gamma^*} J_a, & C_3 \\ \dfrac{1-x}{3} J_a, & C_4 \end{cases} \quad (1)$$

where x (unitless) is the fraction of total electron transport of mesophyll and bundle sheath allocated to the CO_2 concentrating mechanism (C_4 only), C_i (μmol mol^{-1}) is the intercellular CO_2 concentration (C_3 only), Γ^*(μmol mol^{-1}) is the CO_2 compensation point in the absence of mitochondrial respiration under light (C_3 only), J_a (μmol m^{-2} s^{-1}) is the actual ETR calculated as:

$$J_a = \frac{\Phi_{\mathrm{PSIImax}} \cdot (1 + k_{\mathrm{DF}})}{1 - \Phi_{\mathrm{PSIImax}}} \cdot q_{\mathrm{L}} \cdot \frac{\mathrm{SIF}}{f^{\mathrm{esc}}} \quad (2)$$

where Φ_{PSIImax} (unitless) is the maximum photochemical quantum efficiency of PSII for dark-adapted leaves (Φ_{PSIImax} is assumed as constant as it is highly conservative across plant species under non-stressed conditions) (Gu *et al* 2019), k_{DF} (unitless) is the ratio between the rate constants of constitutive thermal dissipation and fluorescence and can also be reasonably assumed as a constant (Gu *et al* 2019), f^{esc} (unitless) is the canopy escape probability of SIF (equation (5), section 2.4) (Badgley *et al* 2017, Zeng *et al* 2019), and q_{L} (unitless) is the fraction of open PSII reaction centers estimated with a PAR-dependent exponential function (Han *et al* 2022a):

$$q_{\mathrm{L}} = a_{q\mathrm{L}} e^{-b_{q\mathrm{L}} \cdot \mathrm{PAR}} \quad (3)$$

where $a_{q\mathrm{L}}$ and $b_{q\mathrm{L}}$ (unitless) are empirical parameters.

$$\mathrm{Yield} = \frac{\sum \mathrm{NPP} \cdot \mathrm{MRY} \cdot (1 - \mathrm{MC}) \cdot 0.45 \frac{\mathrm{gC}}{\mathrm{gCO_2}}}{\mathrm{HI} \cdot \mathrm{fAB}} \quad (4)$$

where MRY is mass per harvest unit (kg $bushel^{-1}$), MC is the plant's moisture content (unitless), HI is harvest index (unitless), i.e. the ratio of yield mass to aboveground biomass, and fAB is the fraction of aboveground to total biomass (unitless).

There are multiple parameters in this set of equations (table S1); their values were all from the literature except $a_{q\mathrm{L}}$ and $b_{q\mathrm{L}}$ that are used to compute qL in equation (3). This study tested two approaches to determine $a_{q\mathrm{L}}$ and $b_{q\mathrm{L}}$. First, we obtained their values directly from leaf-level measurements (Han *et al* 2022a) and term this approach as uncalibrated MLR-SIF. The second approach calibrated $a_{q\mathrm{L}}$ and $b_{q\mathrm{L}}$

only, using nonlinear least-square optimization, and we term this as calibrated MLR-SIF. To rationale for testing these two approaches was to examine whether/by how much parameter recalibration may help improve the performance of MLR-SIF, a test that can help assess the scalability of MLR-SIF.

2.3. Machine learning algorithms

We utilized MATLAB to perform ANN and RF based yield estimation. Specifically, ANN was created using sigmoidal transfer functions; the number of neurons and hidden layers was optimized by minimizing the mean squared error through the Levenberg–Marquardt algorithm (Kira and Sun 2020). The RF model comprised 100 trees and employed the infinitesimal jackknife method to estimate the model's uncertainty (Wen *et al* 2020). To achieve optimal predictability, both ANN and RF require a substantial crop yield dataset encompassing diverse environmental conditions and management practices, for model calibration. Due to limited yield data (that overlaps with available OCO-2 SIF), especially in IGP-WB, we employed the leave-one-year-out cross-validation strategy for model performance evaluation.

2.4. Crop-type specific OCO-2 SIF and other ancillary datasets

This study leverage advances in high-resolution satellite SIF observations (Yu *et al* 2019) and sub-pixel extraction algorithms (Kira and Sun 2020) to improve crop yield prediction. As the native satellite SIF retrievals are often offered at low spatial resolutions (Wen *et al* 2020, Sun *et al* 2023b), high-resolution SIF products are needed to reduce the percentage of pixel contamination by other crop/vegetation types. Each crop type has its unique SIF emission capacity that changes throughout the growing season as a function of the growth stage and environmental conditions (Kira and Sun 2020). Including other crop/vegetation types in the yield estimation may lead to under/over-estimation. Therefore, this study attempted, for the first time, to use crop-type specific SIF (at 0.05°) for yield prediction in US-CB. To extract corn-specific SIF, we applied the sub-pixel endmember unmixing framework (Kira and Sun 2020) to the spatially contiguous OCO-2 daily mean SIF product at 0.05° and 16-day resolution (Yu *et al* 2019). This spatially contiguous OCO-2 SIF was reconstructed from the native OCO-2 SIF retrievals that have substantial spatial gaps between orbits. This product has been validated with ground and airborne measurements to ensure its quality, and is publicly available at (https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1863). This sub-pixel unmixing framework enables the separation of SIF from corn and soybean in the four major corn production states in US-CB: Indiana, Illinois, Iowa, and Nebraska. The four states have the least contamination of vegetation types other than corn and soybean, and thus are chosen here to demonstrate the importance of crop-type specific SIF for yield estimation, which were all ignored in previous studies that utilized satellite SIF for yield prediction. Nevertheless, the sub-pixel unmixing into crop-type specific values can be extended to other states in future studies to enable crop-type specific yield prediction for broader geographical regions. Note, without otherwise specified, MLR-SIF takes the crop-specific SIF (at 0.05°) as input for US-CB. Unfortunately, in IGP-WB, sub-pixel extraction of wheat-specific SIF was not feasible due to the lack of sufficient pure wheat SIF pixels needed for unmixing. Therefore, SIF at 0.05° (with mixed vegetation types) was employed in IGP-WB. Note, to ensure a fair comparison, all three models tested here, i.e. MLR, ANN, and RF, utilized identical satellite SIF input (16 day) during the growing season, defined as June–September in US-CB and October–January in IGP-WB.

The at-sensor satellite SIF is only a small portion of the total canopy SIF emission (that is directly linked to J_a and thus photosynthesis) escaping out of the vegetation canopy, due to leaf/canopy reabsorption/scattering (Yang and van der Tol 2018). Therefore, a conversion, i.e. escape probability f^{esc}, is necessary to compute the total canopy SIF emission from at-sensor SIF. Here we followed Zeng *et al* (2019):

$$f^{esc} = \frac{\mathrm{NIR_V}}{\mathrm{fPAR}} \approx \frac{\mathrm{NDVI} \cdot \mathrm{NIR}}{\mathrm{fPAR}} \tag{5}$$

where NDVI is the normalized difference vegetation index, and NIR is surface reflectance at near-infrared. NDVI and NIR were derived from MODIS BRDF-corrected surface reflectance (MCD43C4 V6) at 500 m. MODIS pixels labeled with high-quality assurance (QA = 0) for the red and the NIR bands were used here. To ensure consistency with crop-specific SIF in US-CB, f^{esc} was also computed separately for corn, which is possible at 500 m resolution. fPAR came from MOD15A2H V6 at 500 m, with only good-quality pixels (MODLAND_QC = 0) considered.

Other ancillary datasets include PAR (to estimate qL, equation (3)) and the NPP/GPP ratio (to convert SIF-based GPP to NPP). Here we used hourly PAR from the Modern-Era Retrospective analysis for Research and Applications V2 (MERRA-2) at 0.5° × 0.625° (Gelaro *et al* 2017). The NPP/GPP ratio was derived from the MODIS GPP and NPP products (MOD17A2H V6, 8 day, global, 500 m). Only pixels with good quality assurance values (MODLAND_QC = 0) were included. All MODIS products were filtered to include pixels with >70% coverage of the crop of interest.

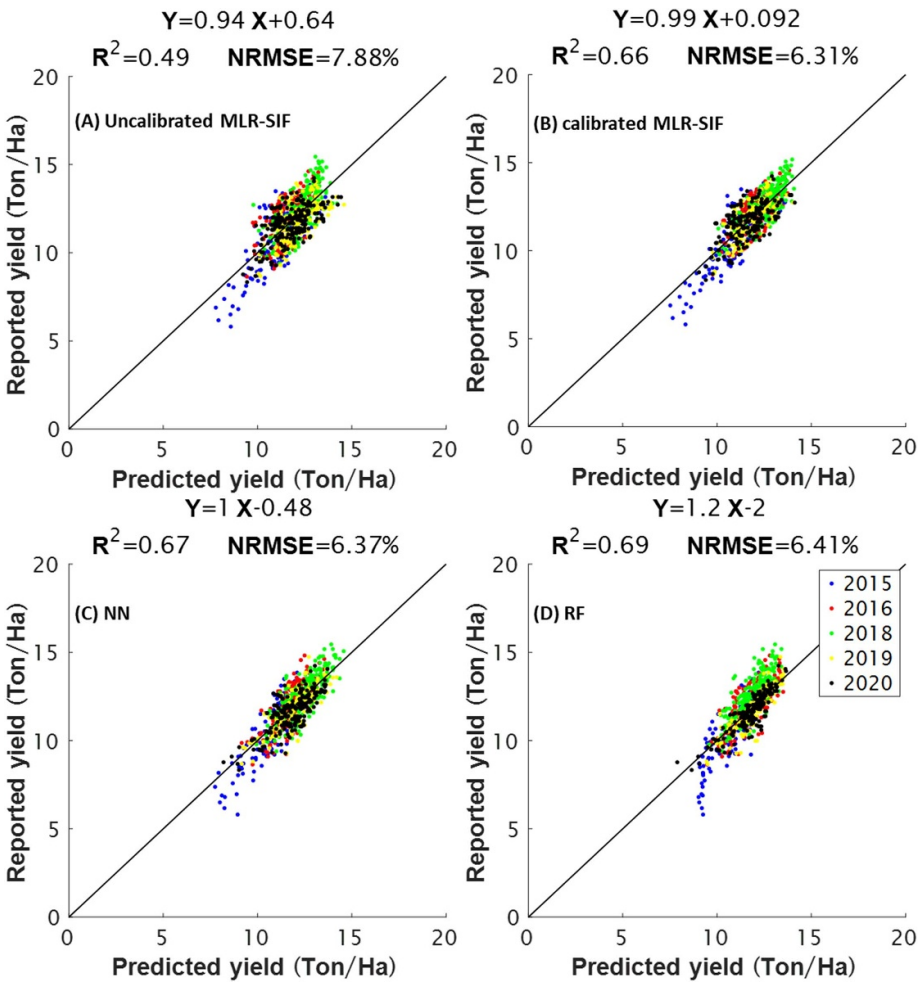

Figure 1. Performance of (A) uncalibrated MLR-SIF, (B) calibrated MLR-SIF, (C) ANN, and (D) RF in estimating the county-level corn yield in US-CB.

3. Results and discussions

3.1. Corn yield estimation in US-CB

MLR-SIF, if uncalibrated, captured 49% of the USDA yield variability in US-CB, with a regression slope of 0.94 (figure 1(A)). This is encouraging, as it requires no model calibration, while ANN and RF require 840 data samples (=210 counties by 4 yr) for model training/calibration. Here ANN and RF (figures 1(C) and (D)) outperformed the uncalibrated MLR-SIF (NRMSE = 6.37%/6.44%, R^2 = 0.67/0.69, respectively), not surprisingly, due to heavy model calibration. However, fine-tuning parameters a_{qL} and b_{qL} significantly improved MLR-SIF's performance for yield estimation (NRMSE = 6.31%, R^2 = 0.66, and slope = 0.99), reaching comparable performance as ANN and RF (figure 1(B)). This calibrated MLR-SIF was used to generate corn yield maps for US-CB (figure 2). MLR-SIF was able to capture the spatiotemporal variability in the USDA-reported yield. Specifically, the interannual variability of reported corn yield was well reproduced by MLR-SIF, e.g. the highest yield in 2018 and the lowest region-wide yield in 2015. The spatial variability of corn yield estimated by MLR-SIF resembled that reported by USDA, i.e. relatively higher yield in Illinois and Iowa than in Indiana and Nebraska. The prediction residual was generally minimal for most counties, and did not exhibit systematic spatial patterns within the study period (figure 2: A3–E3).

3.2. Wheat yield estimation in IGP-WB

The power of the MLR-SIF yield model is manifested more clearly in developing countries, where high-quality yield data is often scarce. In IGP-WB, the uncalibrated MLR-SIF (figure 3(A)) already outperformed (NRMSE = 15.7%, R^2 = 0.51) ANN (figure 3(C): NRMSE = 18.2%, R^2 = 0.41) and RF (figure 3(D): NRMSE = 18.8%, R^2 = 0.39). The calibrated MLR-SIF further improved the accuracy of yield estimation but slightly (NRMSE = 15.2%, R^2 = 0.53). Similar to US-CB, the spatial mapping of MLR-SIF captured the spatiotemporal variability in DLD-reported wheat yield, e.g. it well captured the highest region-wide yield in 2017 and the lowest yield in 2015 (figure 4). In addition, the spatial yield

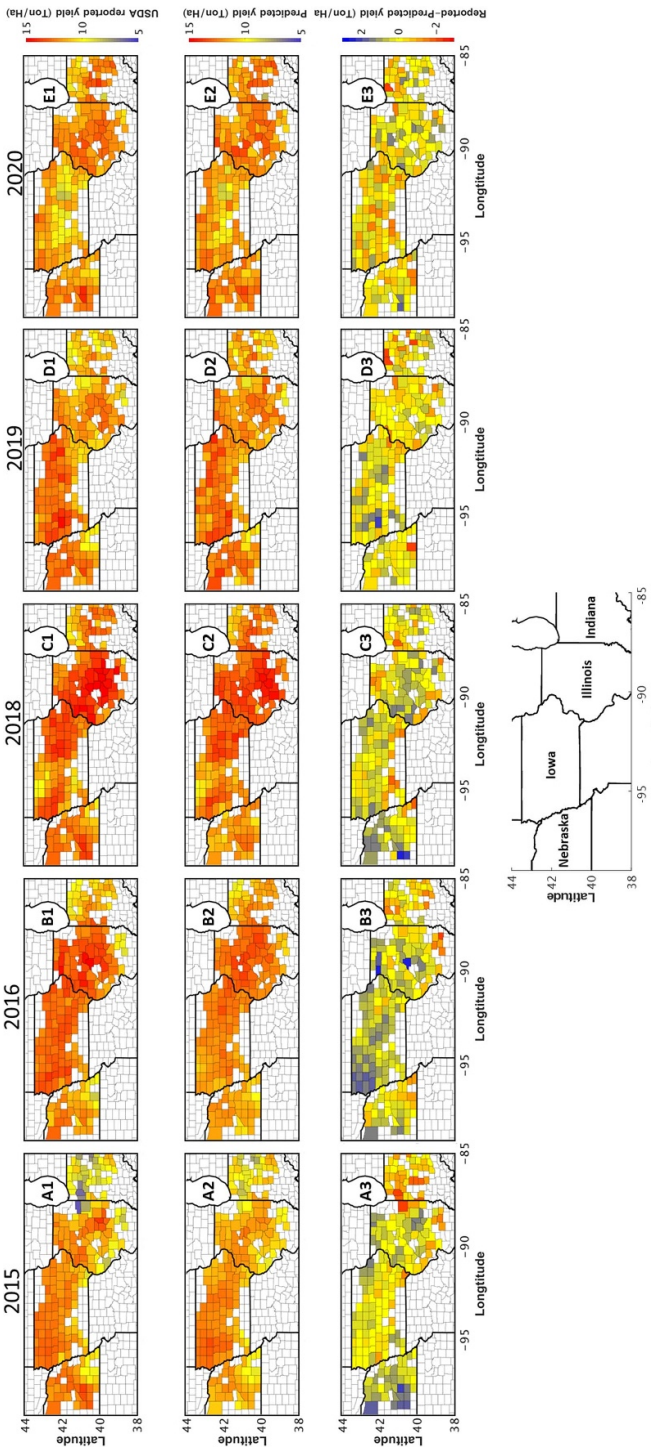

Figure 2. Corn yield maps from USDA reports (A1–E1), MLR-SIF estimates (calibrated) (A2–E2), and their difference (USDA—MLR-SIF) (A3-E3), for 2015, 2016, 2018, 2019, and 2020 respectively.

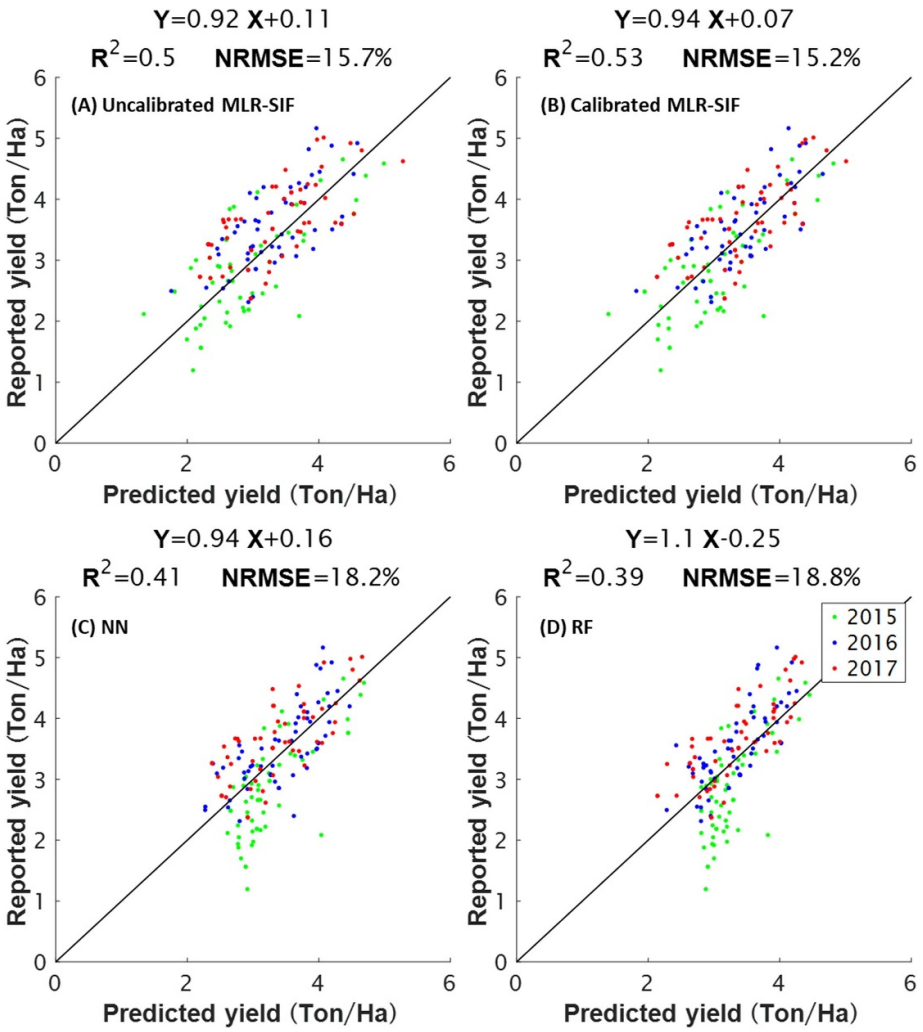

Figure 3. Performance of (A) uncalibrated MLR-SIF, (B) calibrated MLR-SIF, (C) ANN, and (D) RF in estimating the district-level wheat yield in IGP-WB.

gradient from Uttar Pradesh to Bihar (high to low) was well reproduced by MLR-SIF within the study period. The prediction residual was overall minimal, except for a few districts in Haryana.

MLR-SIF's performance in IGP-WB was generally weaker than in US-CB, which is not surprising given the following factors. First, IGP-WB has a much higher cloud cover than US-CB during the growth season. While SIF is relatively insensitive to the interference of thin clouds (Frankenberg *et al* 2012), it is still impacted by thick clouds. Clouds also impact other data input derived from surface reflectance (e.g. land cover types, f^{esc}). Moreover, in IGP-WB, SIF was not purely emitted from wheat, but a mixed signal from all vegetation types within a 0.05 pixel (as explained above). Contamination from other vegetation types could degrade the performance of MLR-SIF. Furthermore, MLR-SIF has different formulations for C_3 from C_4 plants. For wheat (C_3), C_i is required, which was set to be 280 ppm (=0.7xCa, assumed to be 400 ppm at present) for simplicity in this study (a reasonable assumption at the seasonal scale) but is actually dynamic under ambient conditions. While C_4 crops (like corn) are much less impacted by the CO_2 diffusion pathway, due to the coordination of mesophyll and bundle sheath cells to concentrate CO_2 in the vicinity of Rubisco, leaving x (in equations (1)) invariant under environmental variations (von Caemmerer 2000).

3.3. Scalability of MLR-SIF

Unlike ML, MLR-SIF does not require large datasets for model calibration for the same level of prediction accuracy (figure 5). This has important implications for the scalability of yield prediction models when conditions change (e.g. weather/climate, management practices, cultivar types) and data availability, quality, and accessibility are restricted (Weitkamp *et al* 2023). For example, with climate change, shifts in irrigation regimes, and changes in germplasm, models trained on past observations will not necessarily replicate the new prevailing conditions. Indeed, the capability of ML for yield prediction for an 'unobserved' scenario (not present in the training data) is

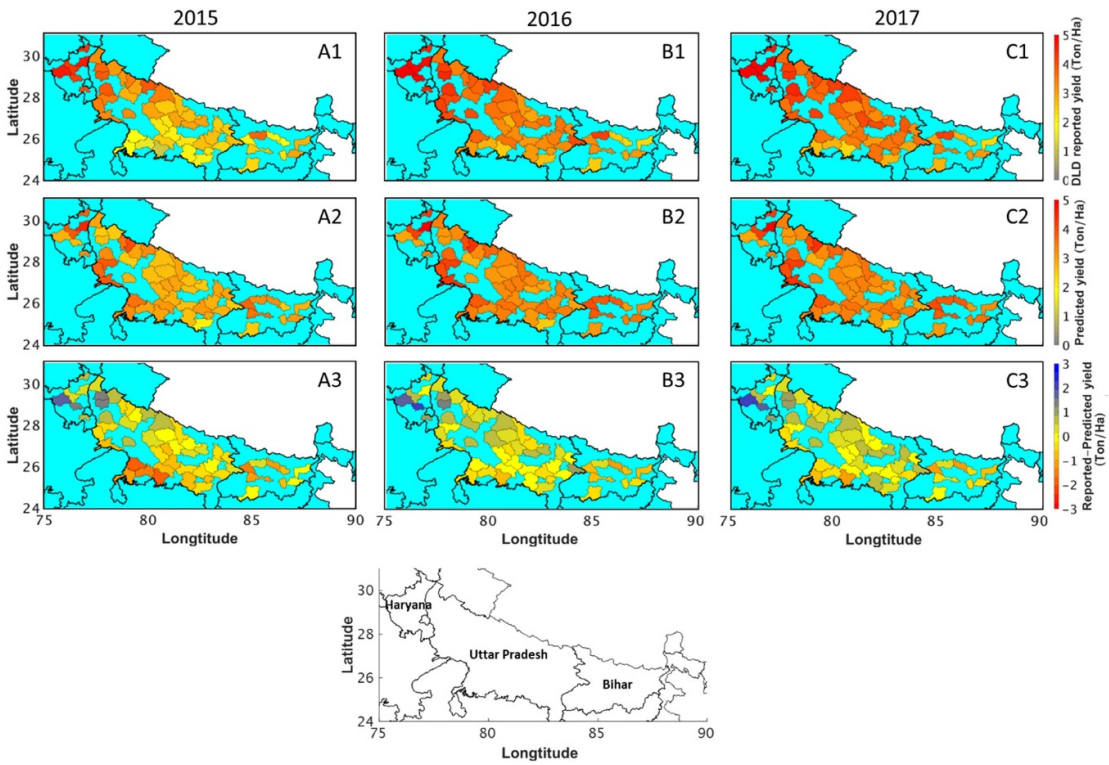

Figure 4. Wheat yield maps from DLD (A1)–(C1), MLR-SIF estimates (A2)–(C2), and their differences (DLD—MLR-SIF, (A3)–(C3)) in IGP-WB for 2015, 2016, and 2017 respectively.

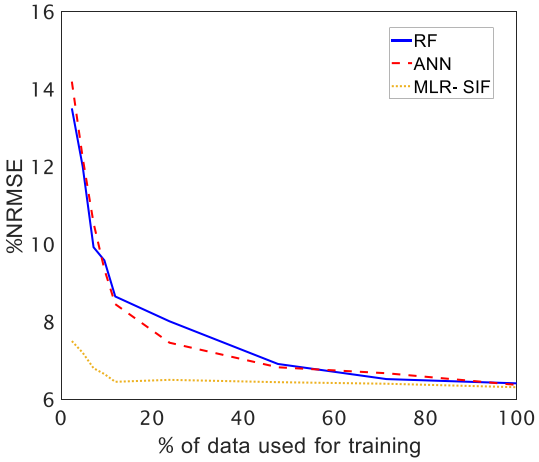

Figure 5. %NRMSE as a function of the training data size, demonstrated by US-CB.

well documented to be limited (due to the implicit "stationarity" assumption). Under such conditions, ML trained with additional or new ground-truth data would be required to re-train a ML model to achieve reasonable performance (Mola-Yudego *et al* 2016). In contrast, MLR-SIF does not require large data for model (re)-calibration, making it more promising for a robust prediction for future changes in both environment and/or management practices. This has already been manifested by our results in two aspects. First, for low-yield years, e.g. 2015 in both US-CB and IGP-WB, ANN and RF considerably overestimated the yield of corn (figures 1(C) and (D)) and wheat (figures 3(C) and (D)), while MLR-SIF was able to reproduce the observed magnitude and variability (figures 1(B) and 3(B)). Second, the performance of ANN and RF significantly dropped in IGP-WB compared to US-CB, likely a consequence of a lower yield variability range (1.2–5.2 Ton/Ha) in the training data than that of US-CB (5.8–15.4 Ton/Ha).

The high scalability of MLR-SIF could also alleviate the need for high-quality ground-truth data for model calibration, which could be a bottleneck in developing countries (Lobell *et al* 2020). For example, wheat yield datasets from DLD may not possess the same degree of quality as the USDA-reported corn yields, which may have also contributed to the degraded performance of ANN and RF in IGP-WB relative to that in US-CB. These results highlight the potential of MLR-SIF for yield prediction using satellite SIF as input, especially when dealing with limited data, both in quality and quantity.

3.4. Caveats of MLR-SIF for crop yield estimation and future work

Impact of spatial resolution and crop-type specific SIF on yield prediction: Previous work only considered the crop-type fraction within a SIF pixel that consists of mixed crop types (and likely other non-crop vegetation) (He *et al* 2020, Peng *et al* 2020), but not the crop-type specific SIF values that can differ significantly due to their differences in phenology, photosynthetic capacity, C_3/C_4 pathways, etc.

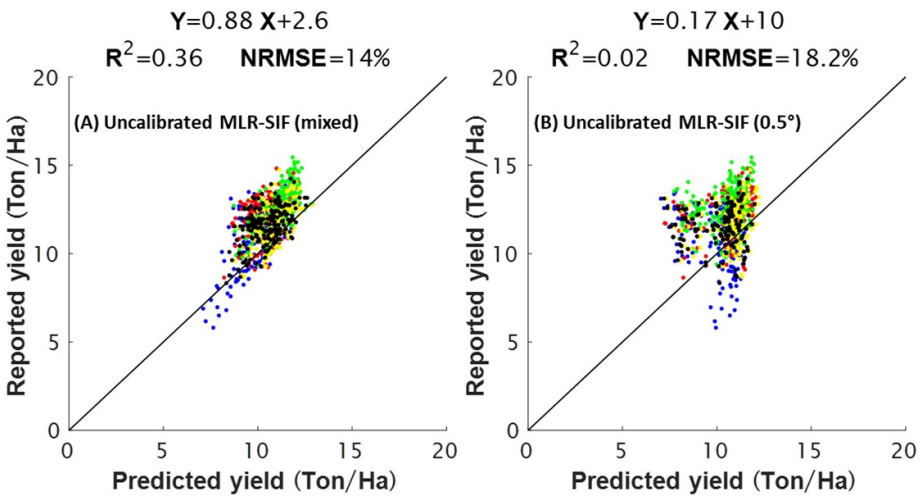

Figure 6. Performance of uncalibrated MLR-SIF for yield prediction in US-CB using SIF (A) from mixed pixels at 0.05°, and (B) at coarser resolution (0.5°).

To quantify such impact on yield prediction, we applied the mixed SIF pixels at 0.05° that are dominated by corn and soybean in US-CB to MLR-SIF (figure 6(A)) and found that prediction performance dropped significantly compared to if corn-specific SIF were used (figure 1(A)). Other studies (e.g. Sloat *et al* 2021) argued that SIF did not possess a comparative advantage for yield prediction over conventional VIs, utilizing coarse-resolution SIF, e.g. 0.5°. By resampling SIF from 0.05° to 0.5°, here we demonstrate that coarse-resolution SIF may considerably obscure the mechanistic advantage of SIF, leading to weak yield predictability (figure 6(B)), as they are mixed with multiple vegetation types, in addition to a higher likelihood of sub-pixel cloud contamination.

Potential uncertainties from parameters required by MLR-SIF: MLR-SIF requires multiple parameters, including Γ^*, C_i, x, $\Phi_{PSIImax}$, a, b, and k_{DF} (table S1). Although leaf-level studies show that some of them are highly convergent across different plant species/biomes without abiotic/biotic stress (Gu *et al* 2019, Han *et al* 2022b), e.g. x, $\Phi_{PSIImax}$, future work should explore the degree to which they vary with environmental variations/stress, especially $\Phi_{PSIImax}$, k_{DF}, a_{qL} and b_{qL} (Sun *et al* 2023a) and the propagated consequences on predicted yields. Moreover, this study utilized a parsimonious model to compute qL, which is a function of PAR only and requires two parameters only. Future studies may explore how qL is affected by other environmental variations, such as temperature (Han *et al* 2022a), and the consequences on yield prediction especially under stress. Additionally, both C_i and Γ^* were assumed constants in calculating GPP of wheat (equation (1)). However, both variables are dynamic. For example, C_i changes with stomatal conductance governed by temperature and vapor pressure deficit (VPD). Γ^* also depends on temperature and the partial pressure of oxygen. Including a dynamic representation of C_i and Γ^* may further improve MLR-SIF's performance for yield prediction. Finally, k_{DF} is needed for J_a estimation; however, it is yet unknown how it varies across biomes (Pfündel 1998, Gu *et al* 2019, Liu *et al* 2022). Such uncertainty is important to quantify given the impact of k_{DF} on yield magnitudes.

There are also potential uncertainties from input datasets required by MLR-SIF. MLR-SIF utilizes multiple satellite products as input; uncertainty in each product can propagate into the final predicted yield. For example, both f^{esc} and NPP/GPP ratio were taken as a regional average because they contribute substantial noise to the yield predictions if using pixel-specific values. More importantly, future improvements in spatial and temporal resolutions and retrieval methods of SIF are crucial to fully unlock the power of satellite SIF for yield estimation or in-season prediction, especially for heterogeneous landscapes. The FLuorescence EXplorer, to be launched in 2025, will have significantly improved spatial resolution (300 m) (Drusch *et al* 2017), which will pave the way for applications of MLR-SIF for small-holder farms in developing countries.

4. Conclusions

This study employed a mechanistic light-reaction-based model driven by satellite SIF, MLR-SIF, to estimate crop yield in US-CB and IGP-WB. We compared MLR-SIF with commonly used ML models for yield prediction (including ANN and RF), and found that ML models lead to high accuracy only when high-quality ground data are available for calibration, while MLR-SIF can perform equally well or better without substantial ground data in both US-CB and IGP-WB. In addition, high-resolution and crop-specific satellite SIF are crucial for accurate yield estimation. This study, for the

first time, demonstrates evidence of the scalability of MLR-SIF for yield prediction in the context of rapidly growing satellite SIF (with increasing resolution and accuracy). Future research is needed to test its global applicability for broader/diverse crop types, agricultural landscapes, climate regimes, and data quality/accessibility restrictions.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary sources listed here). SIF_oco2_005 can be accessed from https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1863. MODIS MCD43A4 can be accessed from http://data.icrisat.org/dld/. https://lpdaac.usgs.gov/products/mcd43a4v061/. MERRA-2 reanalysis can be accessed from https://disc.gsfc.nasa.gov/datasets/. Corn crop yields for the US-CB region can be accessed from https://quickstats.nass.usda.gov/. Wheat crop yields for the IGP-WB region can be accessed from http://data.icrisat.org/dld/.

Acknowledgments

This study is supported by the Cornell Initiative for Digital Agriculture Research Innovation Fund, USDA-NIFA Hatch Fund (1014740), and USAID Feed the Future program (7200AA18CA00014). Y Sun and J Wen also acknowledge support from the NASA MEaSures project.

ORCID iDs

Oz Kira https://orcid.org/0000-0002-1620-2323
Christopher B Barrett https://orcid.org/0000-0001-9139-2721
Ariel Ortiz-Bobea https://orcid.org/0000-0003-4482-6843
Ying Sun https://orcid.org/0000-0002-9819-1241

References

Badgley G, Field C B and Berry J A 2017 Canopy near-infrared reflectance and terrestrial photosynthesis *Sci. Adv.* **3** 1–6

Burke M and Lobell D B 2017 Satellite-based assessment of yield variation and its determinants in smallholder African systems *Proc. Natl Acad. Sci. USA* **114** 2189–94

Cai Y *et al* 2019 Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches *Agric. For. Meteorol.* **274** 144–59

Chlingaryan A, Sukkarieh S and Whelan B 2018 Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: a review *Comput. Electron. Agric.* **151** 61–69

Drusch M *et al* 2017 The fluorescence explorer mission concept-ESA's earth explorer 8 *IEEE Trans. Geosci. Remote Sens.* **55** 1273–84

Erenstein O and Thorpe W 2011 Livelihoods and agro-ecological gradients: a meso-level analysis in the Indo-Gangetic Plains, India *Agric. Syst.* **104** 42–53

Foley J A *et al* 2011 Solutions for a cultivated planet *Nature* **478** 337–42

Frankenberg C and Berry J 2017 *Solar Induced Chlorophyll Fluorescence: Origins, Relation to Photosynthesis and Retrieval* vol 1–9 (Elsevier) (available at: http://linkinghub.elsevier.com/retrieve/pii/B9780124095489106323)

Frankenberg C, O'Dell C, Guanter L and McDuffie J 2012 Remote sensing of near-infrared chlorophyll fluorescence from space in scattering atmospheres: implications for its retrieval and interferences with atmospheric CO_2 retrievals *Atmos. Meas. Tech.* **5** 2081–94

Gastli M S, Nassar L and Karray F 2021 Satellite images and deep learning tools for crop yield prediction and price forecasting *Proc. Int. Joint Conf. on Neural Networks* vol 2021-July (Institute of Electrical and Electronics Engineers Inc.)

Gelaro R *et al* 2017 The modern-era retrospective analysis for research and applications, version 2 (MERRA-2) *J. Clim.* **30** 5419–54

Ghazaryan G, Skakun S, Konig S, Rezaei E E, Siebert S and Dubovyk O 2020 Crop yield estimation using multi-source satellite image series and deep learning *Int. Geoscience and Remote Sensing Symp. (IGARSS)* (Institute of Electrical and Electronics Engineers Inc.) pp 5163–6

Godfray H C J, Beddington J R, Crute I R, Haddad L, Lawrence D, Muir J F, Pretty J, Robinson S, Thomas S M and Toulmin C 2010 Food security: the challenge of feeding 9 billion people *Science* **327** 812–8

Government of India Ministry of Finance Department of Economic Affairs 2023 Economic survey 2022–23

Gu L, Han J, Wood J D, Chang C Y Y and Sun Y 2019 Sun-induced Chl fluorescence and its importance for biophysical modeling of photosynthesis based on light reactions *New Phytol.* **223** 1179–91

Guan K, Berry J A, Zhang Y, Joiner J, Guanter L, Badgley G and Lobell D B 2016 Improving the monitoring of crop productivity using spaceborne solar-induced fluorescence *Glob. Change Biol.* **22** 716–26

Guanter L *et al* 2014 Global and time-resolved monitoring of crop photosynthesis with chlorophyll fluorescence *Proc. Natl Acad. Sci. USA* **111** E1327–33

Han J *et al* 2022a The physiological basis for estimating photosynthesis from Chla fluorescence *New Phytol.* **234** 1206–19

Han J, Gu L, Wen J and Sun Y 2022b Inference of photosynthetic capacity parameters from chlorophyll a fluorescence is affected by redox state of PSII reaction centers *Plant Cell Environ.* **45** 1298–314

He L *et al* 2020 From the ground to space: using solar-induced chlorophyll fluorescence to estimate crop productivity *Geophys. Res. Lett.* **47**

Jain M, Singh B, Srivastava A A K, Malik R K, McDonald A J and Lobell D B 2017 Using satellite data to identify the causes of and potential solutions for yield gaps in India's Wheat Belt *Environ. Res. Lett.* **12** 094011

Jin Z, Azzari G and Lobell D B 2017 Improving the accuracy of satellite-based high-resolution yield estimation: a test of multiple scalable approaches *Agric. For. Meteorol.* **247** 207–20

Jin Z, Azzari G, You C, Di Tommaso S, Aston S, Burke M and Lobell D B 2019 Smallholder maize area and yield mapping at national scales with Google Earth Engine *Remote Sens. Environ.* **228** 115–28

Khalil Z H and Abdullaev S M 2021 Neural network for grain yield predicting based multispectral satellite imagery: comparative study *Proc. Comput. Sci.* **186** 269–78

Kira O and Sun Y 2020 Extraction of sub-pixel C3/C4 emissions of solar-induced chlorophyll fluorescence (SIF) using artificial neural network *ISPRS J. Photogramm. Remote Sens.* **161** 135–46

Liu L *et al* 2024 Knowledge-guided machine learning can improve carbon cycle quantification in agroecosystems *Nat. Commun.* **15** 357

Liu Z, Zhao F, Liu X, Yu Q, Wang Y, Peng X, Cai H and Lu X 2022 Direct estimation of photosynthetic CO_2 assimilation from

solar-induced chlorophyll fluorescence (SIF) *Remote Sens. Environ.* **271** 112893
Lobell D B, Azzari G, Burke M, Gourlay S, Jin Z, Kilic T and Murray S 2020 Eyes in the sky, boots on the ground: assessing satellite- and ground-based approaches to crop yield measurement and analysis *Am. J. Agric. Econ.* **102** 202–19
Lobell D B, Hicke J A, Asner G P, Field C B, Tucker C J and Los S O 2002 Satellite estimates of productivity and light use efficiency in United States agriculture, 1982–98 *Glob. Change Biol.* **8** 722–35
Lobell D B, Ortiz-Monasterio J I, Asner G P, Naylor R L and Falcon W P 2005 Combining field surveys, remote sensing, and regression trees to understand yield variations in an irrigated wheat landscape *Agron. J.* **97** 241–9
Lobell D B, Schlenker W and Costa-Roberts J 2011 Climate trends and global crop production since 1980 *Science* **333** 616–20
Lobell D B, Thau D, Seifert C, Engle E and Little B 2015 A scalable satellite-based crop yield mapper *Remote Sens. Environ.* **164** 324–33
McDonald A J, Keil A, Srivastava A, Craufurd P, Kishore A, Kumar V, Paudel G, Singh S, Singh A K and Sohane R K 2022 Time management governs climate resilience and productivity in the coupled rice–wheat cropping systems of eastern India *Nat. Food* **3** 542–51
Mohammadi K, Jiang Y and Wang G 2022 Flash drought early warning based on the trajectory of solar-induced chlorophyll fluorescence *Proc. Natl Acad. Sci.* **119** e2202767119
Mola-Yudego B, Rahlf J, Astrup R and Dimitriou I 2016 Spatial yield estimates of fast-growing willow plantations for energy based on climatic variables in northern Europe *GCB Bioenergy* **8** 1093–105
Morais T G, Teixeira R F M, Figueiredo M and Domingos T 2021 The use of machine learning methods to estimate aboveground biomass of grasslands: a review *Ecol. Indic.* **130** 108081
Ortiz-Bobea A, Knippenberg E and Chambers R G 2018 Growing climatic sensitivity of U.S. agriculture linked to technological change and regional specialization *Sci. Adv.* **4** eaat4343
Papageorgiou G and Govindjee G 2004 *Chlorophyll A Fluorescence: A Signature of Photosynthesis* vol 19 (Springer)
Paudel D, Boogaard H, de Wit A, van der Velde M, Claverie M, Nisini L, Janssen S, Osinga S and Athanasiadis I N 2022 Machine learning for regional crop yield forecasting in Europe *Field Crops Res.* **276** 108377
Peng B, Guan K, Pan M and Li Y 2018 Benefits of seasonal climate prediction and satellite data for forecasting U.S. maize yield *Geophys. Res. Lett.* **45** 9662–71
Peng B, Guan K, Zhou W, Jiang C, Frankenberg C, Sun Y, He L and Köhler P 2020 Assessing the benefit of satellite-based solar-induced chlorophyll fluorescence in crop yield prediction *Int. J. Appl. Earth Observ. Geoinf.* **90** 102126
Pfündel E 1998 Estimating the contribution of photosystem I to total leaf chlorophyll fluorescence *Photosynth. Res.* **56** 185–95
Porcar-Castell A *et al* 2021 Chlorophyll a fluorescence illuminates a path connecting plant molecular biology to Earth-system science *Nat. Plants* **7** 998–1009
Porcar-Castell A, Tyystjärvi E, Atherton J, Van Der Tol C, Flexas J, Pfündel E E, Moreno J, Frankenberg C and Berry J A 2014 Linking chlorophyll a fluorescence to photosynthesis for remote sensing applications: mechanisms and challenges *J. Exp. Bot.* **65** 4065–95
Sishodia R P, Ray R L and Singh S K 2020 Applications of remote sensing in precision agriculture: a review *Remote Sens.* **12** 1–31
Sloat L L, Lin M, Butler E E, Johnson D, Holbrook N M, Huybers P J, Lee J E and Mueller N D 2021 Evaluating the benefits of chlorophyll fluorescence for in-season crop productivity forecasting *Remote Sens. Environ.* **260** 112478
Sun Y *et al* 2023a From remotely sensed solar-induced chlorophyll fluorescence to ecosystem structure, function, and service: part I—Harnessing theory *Glob. Change Biol.* **29** 2926–52
Sun Y *et al* 2023b From remotely-sensed solar-induced chlorophyll fluorescence to ecosystem structure, function, and service: part II—Harnessing data *Glob. Change Biol.* **29** 2893–925
von Caemmerer S 2000 *Biochemical Models of Leaf Photosynthesis* (CSIRO Publishing) (https://doi.org/10.1071/9780643103405)
Wang C, Beringer J, Hutley L B, Cleverly J, Li J, Liu Q and Sun Y 2019 Phenology dynamics of dryland ecosystems along the North Australian tropical transect revealed by satellite solar-induced chlorophyll fluorescence *Geophys. Res. Lett.* **46** 5294–302
Weitkamp T and Karimi P 2023 Phenology dynamics of dryland ecosystems along the North Australian tropical transect revealed by satellite solar-induced chlorophyll fluorescence *Remote Sens.* **15** 3017
Wen J, Köhler P, Duveiller G, Parazoo N C, Magney T S, Hooker G, Yu L, Chang C Y and Sun Y 2020 A framework for harmonizing multiple satellite instruments to generate a long-term global high spatial-resolution solar-induced chlorophyll fluorescence (SIF) *Remote Sens. Environ.* **239** 111644
Yang P and van der Tol C 2018 Linking canopy scattering of far-red sun-induced chlorophyll fluorescence with reflectance *Remote Sens. Environ.* **209** 456–67
Yang Q, Liu L, Zhou J, Ghosh R, Peng B, Guan K, Tang J, Zhou W, Kumar V and Jin Z 2023 A flexible and efficient knowledge-guided machine learning data assimilation (KGML-DA) framework for agroecosystem prediction in the US Midwest *Remote Sens. Environ.* **299** 113880
Yu L, Wen J, Chang C Y, Frankenberg C and Sun Y 2019 High-resolution global contiguous SIF of OCO-2 *Geophys. Res. Lett.* **46** 1449–58
Zeng Y, Badgley G, Dechant B, Ryu Y, Chen M and Berry J A 2019 A practical approach for estimating the escape ratio of near-infrared solar-induced chlorophyll fluorescence *Remote Sens. Environ.* **232** 111209