

Using Microsatellite Data to Map the Persistence of Field-level Yield Gaps and Their Drivers in Smallholder Systems

Ambica Paliwal (✉ ambica@umich.edu)

University of Michigan

Balwinder Singh

International Maize and Wheat Improvement Center (CIMMYT)

Shishpal Poonia

International Maize and Wheat Improvement Center (CIMMYT)

Meha Jain

University of Michigan

Article

Keywords: Yield gap, microsatellite data, smallholder farming system, India, wheat, SkySat PlanetScope

Posted Date: May 24th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1654249/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Scientific Reports on July 10th, 2023. See the published version at <https://doi.org/10.1038/s41598-023-37818-2>.

Abstract

One way to meet growing food demand is to increase yields in regions that have large yield gaps, including smallholder systems. To do this, it is important to quantify yield gaps, their persistence, and their drivers at large spatio-temporal scales. Here we use microsatellite data to map field-level yields from 2014 to 2018 in Bihar, India and use these data to assess the magnitude, persistence, and drivers of yield gaps at the landscape scale. We find that overall yield gaps are large (33% of mean yields), but only 17% of this yield gap is persistent across time. We find that sowing date, plot area, and weather are the factors that most explain variation in yield gaps across our study region, with earlier sowing associated with significantly higher yield values. Simulations suggest that if all farmers were able to adopt ideal management strategies, including earlier sowing and more irrigation use, yield gaps could be closed by up to 42%. These results highlight the ability of micro-satellite data to understand yield gaps and their drivers, and can be used to help identify ways to increase production in smallholder systems across the globe.

1. Introduction

Food production may have to increase by up to 70% by mid-century to achieve global food security¹. One way to increase yields is to close existing yield gaps, which are defined as the difference between current yields and the potential yields that could be achieved under ideal management^{2,3}. Closing yield gaps is particularly important in smallholder systems, where yield gaps are large⁴ and food demand is likely to increase the most over the coming decades⁵. Correctly estimating yield gaps and their drivers is challenging in smallholder systems due to a lack of agricultural production data, particularly at the field scale⁶. Conventional methods of collecting field-level yield estimates, such as crop cuts, are time and cost intensive⁷. Satellite data have been shown to be a valuable tool for estimating yields, and quantifying yield gaps and their drivers at the landscape scale⁸⁻¹⁰. Yet, historically available imagery, such as Landsat (30 m) and MODIS (250 m), which have primarily been used to map yield gaps in smallholder systems^{11,12} are likely too coarse in spatial resolution to map individual field-level yields¹³.

Over the last five years, new micro-satellite data have become readily available and have shown promise for mapping yields at the field scale^{6,13}. Despite the availability of such data since 2015, to date microsatellite data have not been used to map yield gaps and their drivers over multiple years. Yet, understanding the persistence of yield gaps through time provides insights into the possible drivers of yield gaps¹⁰. For example, if yield gaps are persistent, this suggests that there may be some consistent infrastructural (e.g., irrigation access) or biophysical (e.g., soil type) factors that largely explain yield gaps in these systems. On the other hand, if yield gaps are not persistent, meaning that different fields have the largest yield gaps in different years, this suggests that there may be some time varying factor, such as weather (e.g., inter-annual rainfall variability) or management (e.g., sowing date), that primarily explain yield gaps¹⁴. Such analyses can provide critical information about what potential interventions may effectively close yield gaps at the field scale in these systems¹⁵.

Here we show that microsatellite data are able to accurately map field-level yields, and quantify yield gaps and their persistence, at regional scales and over multiple years in smallholder systems. We also show that these data can also be combined with ancillary data on management, weather, and biophysical factors to identify which factors are the most important in explaining field-level yield gaps over multiple years. We focus our study in wheat systems in eastern India, where yield gaps are large and field sizes are especially small^{11,16} (< 0.3 ha). We specifically use SkySat and PlanetScope data (~ 3 m resolution) to map field-level yields from 2014 to 2019 and assess the magnitude and persistence of yield gaps and their possible drivers. Using this information, we then assess how much yield gaps can be closed under ideal management conditions, providing invaluable insights for how much yields may be able to increase over the coming years. While our study is focused on wheat systems in eastern India, our approach can be used to quantify yield gaps, their persistence, and their drivers in smallholder systems across the globe.

2. Methods

2.1. Study area

We conducted our study in Arrah district, Bihar, India (25.47°N, 84.52°E), which is in the eastern portion of India's main grain belt, the Indo-Gangetic Plains (IGP) (Fig. 1). Our analyses are focused on an 8 × 16 km² area where we had access to SkySat and PlanetScope imagery over multiple years. The region is dominated by smallholder farms (< 0.3 ha) (Jain et al. 2016) and over 80% of the land area is under agriculture. Most farmers in the region follow a rice-wheat cropping system, with rice planted during the monsoon (kharif) growing season and wheat planted in the winter (rabi) growing season. Our analyses focus on wheat as previous studies have shown that yield gaps for wheat are large in this region^{11,17} and are expected to increase given the negative impacts of warming temperatures on wheat yield^{18,19}. The wheat growing season spans from early November to mid-April, and wheat management varies widely across farms resulting in significant across-farm heterogeneity in yield¹³. For example, wheat sowing dates vary from early November to early January and farmers also vary the number of irrigations applied throughout the growing season (ranging from 1 to 3 irrigations¹³).

2.2. Field data collection and processing

We collected crop cut data from 271 wheat fields from 2014-15 to 2018-19, with the number of fields ranging from 36 to 79 in a given year (details provided in Table S1). It is important to note that different fields were visited in different years, meaning that repeat samples were not collected for the same field through time. To collect crop cut data, the field team visited each of the 271 fields at the time of crop harvest and randomly selected subplots to harvest from each of the farmer's fields (details in Table S1). The team harvested the crop from these subplots, threshed the sample, and weighed the grain on site. We then averaged the yields from each subplot for each field to estimate the average yield per hectare for the full field. In addition, the field team collected five GPS points, one from the field's center and four from the

corners of each field. For years 2014-15 and 2015-16, we also conducted a survey about farmers' management practices in the field for that year. The survey included questions about management factors shown to be important for explaining wheat yields in the previous literature, including sow date and the number of irrigations throughout the season. This work was considered exempt for Human Subjects Research by the University of Michigan Institutional Review Board (HUM00156479, HUM00128955, HUM00120778).

To derive field polygon boundaries from the five GPS coordinates for each field, we used the `rgeos`²⁰ and `sp`²¹ packages in R Project Software Version 4.1.0²². We overlaid all field polygons over high-resolution imagery from Google Earth (<https://www.google.com/earth/>), and adjusted field polygons to match field boundaries that were visible in the high-resolution imagery^{23,24}. We then linked all associated yield and management factors with each polygon, and used the resulting shapefile for all subsequent analyses.

2.3. Satellite yield estimation

We used SkySat (2 m) imagery for the 2014–2015 and 2015–2016 growing seasons, and PlanetScope (3 m) imagery for the 2016–2017 to 2018–2019 growing seasons to estimate wheat yield. The number of images and specific dates used varied across years depending on image availability and cloud cover (details provided in Table S2). We assessed image availability by searching the PlanetScope API (www.planet.com) for all available images for our study site (Fig. 1) from November 1 to April 15 for each year's growing season. We then visually inspected all available images and selected only those images that were cloud free. Since multiple tiles encompassed our study area, we mosaicked tiles into one image that covered the full study area extent using histogram-matching of overlapping areas in ENVI Software. SkySat imagery were provided as top of the atmosphere reflectance, so we corrected imagery to surface reflectance by stretching histograms to match distributions of each band as seen in cloud-free, surface reflectance corrected Landsat 7 and 8 imagery obtained from Google Earth Engine²⁵ (GEE). Specifically, images were matched to cloud-free Landsat scenes from the closest available image date, and if a cloud-free Landsat scene was not available within two days of a given SkySat scene, we used a date-weighted average of the histograms from the two closest Landsat scenes before and after each available SkySat date (more details provided in Jain et al.¹³). All PlanetScope imagery were provided as surface reflectance corrected data, and thus all images were used directly without additional corrections.

SkySat and PlanetScope have blue (450– 515nm, 450– 515nm), green (515– 595nm, 500– 590nm), red (605– 695nm, 590– 670nm), and near infrared (NIR, 740– 900nm, 780– 860nm) bands. Using these bands, we calculated the green chlorophyll vegetation index for each image (GCVI) (Eq. 1) as previous studies have shown that GCVI has a linear relationship with the leaf area index for wheat²⁶.

$$\text{GCVI} = (\text{NIR}/\text{green}) - 1 \quad (1)$$

We then extracted the mean GCVI for each field polygon for each image date in all years, and these mean GCVI values were used to create our yield estimation model. To estimate yield, we used a two-step

approach. First, we predicted yield using random forest linear regressions, where each year's crop cut data were used to train individual random forest models for each year (Eq. 2)

$$\text{Crop cut yield} \sim \beta_0 + \beta_1 \text{GCVI}_1 + \beta_2 \text{GCVI}_2 \dots + \beta_n \text{GCVI}_n + \varepsilon \quad (2)$$

where crop cut yield (kg/ha) is the observed yield estimated using crop cuts for each polygon, GCVI_1 to GCVI_n are the mean GCVI values for each polygon for each image date (n) within a given growing season, and ε represents the residual error. For each year, the coefficients derived from the random forest model were applied to the stacked GCVI raster layer for the respective year. We found that random forest yield estimation resulted in models with high correlation between estimated and observed yields, however, the results did not fall consistently on the one-to-one line, suggesting systematic over or underestimation of yields in each year (Figure S1). Thus, to correct for this systematic bias, we conducted a second step, where we regressed the observed crop cut yields on the random forest estimated yields (Eq. 3).

$$\text{Crop cut yield} \sim \beta_0 + \beta_1 \text{RF estimated yield} + \varepsilon \quad (3)$$

where crop cut yield (kg/ha) is the observed yield estimated using crop cuts for each polygon, RF estimated yield (kg/ha) represents the mean predicted yield for each polygon derived from the random forest model from the first step (Eq. 2), and ε represents the residual error. To calculate mean satellite yield, we took the mean value of all pixels within each field's polygon. We then applied the coefficients from Eq. 3 to the full raster stack to correct predicted random forest yields at the pixel scale across the study site. It is important to note that this correction (Eq. 3) was conducted separately for each of the five years. We validated our satellite yield estimates at the polygon scale by comparing estimated yields after the two-step approach with observed crop cut yields at the field scale. Accuracy was evaluated using R^2 and root mean squared error (RMSE).

2.4. Yield gap estimation and persistence

To estimate yield gaps (YG), we subtracted the mean yield (Y_{mean}) for each polygon from yield potential (Y_p) for each year (y; Eq. 4).

$$\text{YG}_y = Y_{p_y} - Y_{\text{mean}_y} \quad (4)$$

Where YG_y equals the yield gap (kg/ha) for each polygon in year y, Y_{p_y} represents the yield potential (kg/ha) for all polygons in a given year y, and Y_{mean_y} represents the mean yield (kg/ha) for each polygon in a given year y. We defined Y_{p_y} as the 95th percentile yield value found in each year's satellite estimated yield raster for the study region after masking out non-cropland pixels using land cover classes from the Global Land Cover product²⁷. Previous studies have suggested that such empirically estimated Y_p better captures realistic economically-achievable yields that consider infrastructural, management, and economic constraints, which are not well accounted for in modelled estimates of Y_p ¹⁰. We calculated Y_{mean} using the mean satellite estimated yield for each polygon for each year, as crop cut yield values were not available for each field in each year.

We measured yield persistence in two ways. First, we estimated how consistently fields were relatively high or low yielding by conducting a decile analysis developed in Lobell¹⁰. Specifically, all 271 fields were categorized into one of ten deciles based on their yield rank using 2014-15 satellite estimated yields. Keeping each field within its original categorized decile from 2014-15, we plotted boxplots of all yields seen for all fields across all remaining years (2015-16 to 2018-19). If yields are persistent, we would expect there to be little overlap between boxplots across decile values, as the lowest yielding fields would always be low yielding and the highest yielding fields would always be high yielding. However, if there is a large amount of boxplot overlap across deciles, this suggests that there is variability in yield through time. Second, we quantified the percent of yield variation that was persistent through time using methods from Lobell et al.²⁸. Specifically, for the highest yielding decile of fields found in 2014-15, we calculated the fields' average anomaly from the study site mean yield in 2014-15. For these same fields, we then calculated the fields' average anomaly compared to the study site mean for all subsequent years (2015-16 to 2018-19). By comparing yield anomalies from 2014-15 with yield anomalies for all remaining years, we gain an understanding of the amount of yield persistence from 2014-15 across later time periods.

2.5. Drivers of yield gaps and the ability to close yield gaps

To understand which factors most influence yield gaps, we conducted random forest regressions where we regressed yield gap estimates for each year on a suite of management, weather, and biophysical variables that have been suggested to be important drivers of yield gaps in the previous literature (Eq. 5). Specifically, for management variables we considered wheat sowing date and the number of irrigations applied, for weather variables we considered average temperature and total rainfall within each winter season, and for biophysical variables we considered soil nitrogen and soil organic carbon.

$$YG_y \sim \beta_0 + \beta_1 \text{DOS}_y + \beta_2 \text{Irrigation}_y + \beta_3 \text{AvgTemp}_y + \beta_4 \text{Tot_Rain}_y + \beta_5 \text{Nitrogen} + \beta_6 \text{Soil_Org_C} + \beta_7 \text{Plotarea} + \varepsilon \quad (5)$$

where YG_y represents the yield gap (kg/ha) calculated for each year for each field (from Eq. 4), DOS_y represents the sowing date of wheat (days since November 1) for each field in each year, Irrigation_y represents the number of irrigations (ranging from 1–3) applied to each field during the wheat growing season in each year, AvgTemp_y represents the average temperature (°C) for each polygon in each year, Tot_Rain_y represents the total amount of rainfall (mm) for each polygon in each year, Nitrogen represents mean soil nitrogen (cg/kg) for each field across all years, Soil_Org_C (dg/kg) represents mean soil organic carbon (SOC) for each field across all years, Plotarea represents area of the field, and ε represents error. We calculated variable importance for our random forest regression (Eq. 5) by examining the mean decrease in accuracy (%IncMSE) over all out-of-bag cross validated predictions when each variable was permuted.

We obtained sowing date and irrigation information from management surveys that were conducted in 2014-15 and 2015-16. Given that we only had management variables available for these two years, we restricted our analyses (Eq. 5) to only these two years. We calculated average temperature using

temperature data from Terra Climate²⁹; specifically we calculated mean temperature for each month (November to April) for each year (2014-15 to 2018-19) using the mean of monthly maximum and minimum temperature. We calculated total rainfall as the sum of monthly rainfall from November to April using monthly precipitation data from Terra Climate²⁹. Finally, we calculated soil nitrogen and SOC using World Soil Information Service (WoSIS) global raster data³⁰. Weather and soil raster data were extracted as the mean value for each polygon using the raster package³¹ in R Project Software 4.1.0²². More details about each dataset, including their source and resolution are included in Table S3.

Finally, we ran simulations to quantify how much yield gaps could be closed if all farmers adopted optimal management strategies. For this analysis, we focused on the two management variables considered in our analyses (Eq. 5), sowing date and number of irrigations applied. To identify what management values were optimal, we examined the partial dependence plots of sow date and irrigation and identified which values were associated with the smallest yield gaps. Partial dependence plots show the marginal effect of each feature on the predicted outcome from our random forest analysis (Eq. 5). Based on the partial dependence plots, we found that a sowing date of November 12 and three irrigations were associated with the lowest yield gaps. In our scenario analysis, we therefore altered all sowing dates to be November 12 and all irrigations to equal 3, and we predicted what yields would be for each field using our random forest model (Eq. 5). To estimate how much yield gain could be achieved, we quantified the difference between this predicted yield value under optimal management and $Y_{\text{mean } y}$. All analyses were done using the Random Forest³² and partial dependence plot³³ packages in R Project Software 4.1.0²².

3. Results

3.1 Accuracy of satellite estimated yields at the field scale

Overall, we find that micro-satellite data can accurately map yield at the field scale across multiple years. While accuracies varied from year to year, all years resulted in moderate to high accuracies (R^2 values range from 0.68 to 0.85), suggesting good fit with observed yield values (Figure 2A-E). RMSE values were also moderate, ranging from 272 kg/ha to 410 kg/ha across the five years.

3.2 Magnitude and persistence of yield gaps

We find that overall yield gaps are large, with an average value of 985 kg/ha across all polygons and all years. This is equal to 33% of mean yield values. The average yield gap varies from year to year, ranging from 543 to 1579 kg/ha (Figure 3). Considering persistence, we find that yields are somewhat persistent through time and the level of persistence likely varies for low versus high yielding fields. Specifically, we find that there is high overlap in our decile boxplot analysis, particularly for lower decile values (Figure 4A). This suggests that fields that are the lowest yielding in 2014-15 are not consistently low yielding across the timeframe of our study. There is, however, a positive increase in yields across decile values, and less overlap in boxplots for higher decile bins, suggesting that there is some yield persistence across

time, particularly for higher yielding fields. Considering the amount of persistence that occurs for the highest yielding fields, we find that 17% of yield anomalies persist from 2014-15 to later time periods (Figure 4B).

3.3 Drivers of yield gaps and the ability to close yield gaps with ideal management

When considering which biophysical, weather, and management factors drive yield gaps, we find that weather and management variables are the most important factors explaining variation in yield gaps (Figure 5). Our model shows that amongst all variables considered, sowing date is the variable that explains the most variation in yield gap. Weather variables and plot area are also found to be important explanatory factors of yield gap, with average temperature and total rainfall explaining similar amounts of variation (Figure 5). Considering partial dependence plots of the management variables considered in our study, we find that later sowing dates are associated with larger yield gaps (Figure S2A) and November 12th is the sowing date associated with the lowest yield gap. Furthermore, yield gaps decrease as more irrigations are applied, with the lowest yield gap seen with three irrigations (Figure S2B). Inputting these ideal management factors (November 12, three irrigations) into our scenario analysis, we find that yields could be increased on average by 414 kg/ha across all fields, which is 42% of the estimated yield gap in this region (Figure 6). Shifting only the sowing date to November 12 could close the yield gap by 25% and only optimizing irrigation could close the yield gap by 18% (Figure 6)

Discussion And Conclusion

We find that microsatellite data can accurately map field-level yields in smallholder systems, and these data can be used to understand the magnitude of yield gaps, their persistence, and their drivers at the landscape scale. This is exciting given that previous remote sensing studies that have focused on yield gap and persistence analyses in smallholder systems have relied on coarser resolution imagery (e.g., Landsat) that are unable to resolve yields at the field scale¹³.

Using micro-satellite data, we find that yield gaps are large on average (985 kg/ha), though the magnitude of yield gaps vary from year to year (Figure 3). Our yield gap estimates are mostly smaller than those found in previous studies for the region, likely because we used an empirically derived estimate of economic yield potential (Y_p). Previous studies have largely used crop model simulations or yields obtained from on-farm trials to estimate Y_p ³⁴⁻³⁶. Yet such estimates of Y_p are not realistic because they do not account for economic and/or infrastructural constraints in the system that are difficult to alleviate. It is important to note that our yield gap estimates are similar to those found in a previous study¹¹ that used Landsat satellite data to map wheat yield gaps across the Indo-Gangetic Plains (IGP). This is likely because this study used a similar approach to quantify Y_p and associated yield gaps. Considering yield persistence, we find that yields are 17% persistent in this region, and yields seem to be more persistent for fields with higher-than-average yields. This suggests that systemic factors that consistently vary across farms, such as differences in soil quality and farmer skill, will also need to be addressed to close yield gaps in this region, particularly for the highest yielding fields.

When analyzing the drivers of yield gaps, we find that sowing date, plot area, and weather are the factors that most explain variation in yield gap in our study region. Plot area is likely capturing the effect of improved management in larger fields that are typically owned by wealthier farmers^{37,38}. The importance of sowing date has been highlighted by previous studies^{11,39,40}, with farmers who sow earlier experiencing higher yields. This is because wheat is one crop that is particularly negatively impacted by heat stress that occurs at the end of the growing season during the time of grain filling⁴¹. If farmers are able to sow their wheat earlier, allowing the crop to mature prior to heat stress at the end of the growing season, negative yield impacts can be reduced^{42,43}. We find that if farmers are able to sow their wheat earlier, they will be able to close yield gaps by 25% (Figure 6). Furthermore, if farmers are additionally able to increase the amount of irrigation they use to three irrigations, yield gaps can be closed by 42%. While this is a substantial amount of the yield gap, it must be noted that the adoption of these ideal management strategies may not be possible for all farmers. Previous studies have suggested that while farmers are aware of yield gains associated with earlier wheat planting, they are often constrained on when they can plant based on irrigation availability, monsoon rainfall patterns, and decisions made in the prior monsoon growing season^{40,44}.

In conclusion, our study highlights the ability of microsatellite data to map yield gaps, their persistence, and their drivers in smallholder farming systems at the field scale. We show that our yield gap estimates are for the most part smaller than those reported in previous studies, as we are able to better capture economic yield potential using empirically driven estimates. We find that yield gaps are somewhat persistent in the region, and can be closed by up to 42% if farmers are able to adopt ideal management strategies. To our knowledge this is the first study where microsatellite data have been used to assess yield gaps over multiple years in smallholder systems. While our study is focused in one region of India, we believe that microsatellite data can be similarly beneficial for yield gap analysis in other smallholder systems across the globe.

Declarations

Acknowledgements: We would like to thank the field staff and CSISA-CIMMYT for their help in collecting crop cut data for this study. This work was supported by the National Science Foundation SEES Postdoctoral Fellowship (Grant Number 1415436) and the NASA Land Cover and Land Use Change grant (no. NNX17AH97G) awarded to M.J.

Author information: The study was conceived of by M.J. and A.P. Data collection was performed by S.P. All analyses were conducted by A.P. A.P. wrote the first draft of the manuscript with edits and contributions from all other authors.

Data availability statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Competing interests statement: There is no conflict of interest to report.

References

1. Alexandratos, N. Proceedings of a Technical Meeting of Experts, How to feed the world in 2050, FAO, Rome. (2009).
2. Licker, R., Johnston, M., Foley, J. A., Barford, C., Kucharik, C. J., Monfreda, C. & Ramankutty, N. Mind the gap: how do climate and agricultural management explain the 'yield gap' of croplands around the world?: Investigating drivers of global crop yield patterns. *Glob. Ecol. Biogeogr.* **19**, 769–782 (2010).
3. van Ittersum, M. K., Cassman, K. G., Grassini, P., Wolf, J., Tittonell, P. & Hochman, Z. Yield gap analysis with local to global relevance—A review. *Field Crops Res.* **143**, 4–17 (2013).
4. Mueller, N. D., Gerber, J. S., Johnston, M., Ray, D. K., Ramankutty, N. & Foley, J. A. Closing yield gaps through nutrient and water management. *Nature* **490**, 254–257 (2012).
5. van Dijk, M., Morley, T., Rau, M. L. & Saghai, Y. A meta-analysis of projected global food demand and population at risk of hunger for the period 2010–2050. *Nat. Food* **2**, 494–501 (2021).
6. Burke, M. & Lobell, D. B. Satellite-based assessment of yield variation and its determinants in smallholder African systems. *Proc. Natl. Acad. Sci.* **114**, 2189–2194 (2017).
7. Paliwal, A. & Jain, M. The Accuracy of Self-Reported Crop Yield Estimates and Their Ability to Train Remote Sensing Algorithms. *Front. Sustain. Food Syst.* **4**, (2020).
8. Kamir, E., Waldner, F. & Hochman, Z. Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods. *ISPRS J. Photogramm. Remote Sens.* **160**, 124–135 (2020).
9. Lobell, D. B., Ortiz-Monasterio, J. I. & Lee, A. S. Satellite evidence for yield growth opportunities in Northwest India. *Field Crops Res.* **118**, 13–20 (2010).
10. Lobell, D. B. The use of satellite data for crop yield gap analysis. *Field Crops Res.* **143**, 56–64 (2013).
11. Jain, M., Singh, B., Srivastava, A. A. K., Malik, R. K., McDonald, A. J. & Lobell, D. B. Using satellite data to identify the causes of and potential solutions for yield gaps in India's Wheat Belt. *Environ. Res. Lett.* **12**, 094011 (2017).
12. Sussy, Ola, Maria, Niklas, Cecilia, Willis, Håkan, & Djurfeldt. Micro-Spatial Analysis of Maize Yield Gap Variability and Production Factors on Smallholder Farms. *Agriculture* **9**, 219 (2019).
13. Jain, M., Srivastava, A. K., Balwinder-Singh, Joon, R. K., McDonald, A., Royal, K., Lisaius, M. C. & Lobell, D. B. Mapping Smallholder Wheat Yields and Sowing Dates Using Micro-Satellite Data. *Remote Sens.* **8**, (2016).
14. Maestrini, B. & Basso, B. Drivers of within-field spatial and temporal variability of crop yield across the US Midwest. *Sci. Rep.* **8**, 14833 (2018).
15. Deines, J. M., Patel, R., Liang, S.-Z., Dado, W. & Lobell, D. B. A million kernels of truth: Insights into scalable satellite maize yield mapping and yield gap analysis from an extensive ground dataset in the US Corn Belt. *Remote Sens. Environ.* **253**, 112174 (2021).
16. Tripathi, A. & Mishra, A. K. Knowledge and passive adaptation to climate change: An example from Indian farmers. *Clim. Risk Manag.* **16**, 195–207 (2017).

17. Mueller, N. D., Gerber, J. S., Johnston, M., Ray, D. K., Ramankutty, N. & Foley, J. A. Closing yield gaps through nutrient and water management. *Nature* **490**, 254–257 (2012).
18. Rao, B. B., Chowdary, P. S., Sandeep, V. M., Pramod, V. P. & Rao, V. U. M. Spatial analysis of the sensitivity of wheat yields to temperature in India. *Agric. For. Meteorol.* **200**, 192–202 (2015).
19. Lobell, D. B. & Field, C. B. Global scale climate–crop yield relationships and the impacts of recent warming. *Environ. Res. Lett.* **2**, 014002 (2007).
20. Bivand, R. & Rundel, C. RGeos: Interface to Geometry Engine-Open Source ('GEOS'). R package version 0.3–26. (2017).
21. Pebesma, E., Bivand, R., Rowlingson, B., Gomez-Rubio, V., Hijmans, R., Sumner, M., MacQueen, D., Lemon, J., O'Brien, J. & O'Rourke, J. sp: Classes and methods for spatial data. R Package Version **1**, (2018).
22. R Development Core Team. *R: A Language and Environment for Statistical Computing.* (2021). at <<http://www.r-project.org/>>
23. Jain, M., Srivastava, A. K., Balwinder-Singh, Joon, R. K., McDonald, A., Royal, K., Lisaius, M. C. & Lobell, D. B. Mapping Smallholder Wheat Yields and Sowing Dates Using Micro-Satellite Data. *Remote Sens.* **8**, (2016).
24. Jain, M., Balwinder-Singh, Rao, P., Srivastava, A. K., Poonia, S., Blesh, J., Azzari, G., McDonald, A. J. & Lobell, D. B. The impact of agricultural interventions can be doubled by using satellite data. *Nat. Sustain.* **2**, 931–934 (2019).
25. Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D. & Moore, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Big Remote. Sensed Data Tools Appl. Exp.* **202**, 18–27 (2017).
26. Nguy-Robertson, A. L., Peng, Y., Gitelson, A. A., Arkebauer, T. J., Pimstein, A., Herrmann, I., Karnieli, A., Rundquist, D. C. & Bonfil, D. J. Estimating green LAI in four crops: Potential of determining optimal spectral bands for a universal algorithm. *Agric. For. Meteorol.* **192–193**, 140–148 (2014).
27. Gong, P., Liu, H., Zhang, M., Li, C., Wang, J., Huang, H., Clinton, N., Ji, L., Li, W., Bai, Y., Chen, B., Xu, B., Zhu, Z., Yuan, C., Ping Suen, H., Guo, J., Xu, N., Li, W., Zhao, Y., Yang, J., Yu, C., Wang, X., Fu, H., Yu, L., Dronova, I., Hui, F., Cheng, X., Shi, X., Xiao, F., Liu, Q. & Song, L. Stable classification with limited sample: transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017. *Sci. Bull.* **64**, 370–373 (2019).
28. Lobell, D. B., Thau, D., Seifert, C., Engle, E. & Little, B. A scalable satellite-based crop yield mapper. *Remote Sens. Environ.* **164**, 324–333 (2015).
29. Abatzoglou, J. T., Dobrowski, S. Z., Parks, S. A. & Hegewisch, K. C. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015. *Sci. Data* **5**, 170191 (2018).
30. Batjes, N. H., Ribeiro, E. & van Oostrum, A. Standardised soil profile data to support global mapping and modelling (WoSIS snapshot 2019). *Earth Syst. Sci. Data* **12**, 299–320 (2020).

31. Hijmans, Robert J and van Etten, Jacob. *raster: Geographic analysis and modeling with raster data*. (2012). at < <http://CRAN.R-project.org/package=raster>>
32. Liaw, A. & Wiener, M. randomForest: Breiman and Cutler's random forests for classification and regression. R Package Version **4**, 14 (2015).
33. Greenwell, Brandon M. pdp: An R Package for Constructing Partial Dependence Plots. R J. **9**, 421–436 (2017).
34. Sendhil, R., Singh, R., Ramasundaram, P., Kumar, A., Singh, S. & Sharma, I. Yield gap in wheat: Approach, quantification and resetting research priorities in India. J. Wheat Res. **6**, 138–149 (2014).
35. Chenu, K., Porter, J. R., Martre, P., Basso, B., Chapman, S. C., Ewert, F., Bindi, M. & Asseng, S. Contribution of Crop Models to Adaptation in Wheat. Trends Plant Sci. **22**, 472–490 (2017).
36. Jha, G. K., Palanisamy, V., Sen, B. & Kumar, A. Explaining Rice and Wheat Yield Gaps in Eastern Indian States: Insights from Stochastic Frontier Analysis. Agric. Res. (2022). doi:10.1007/s40003-021-00599-z
37. Key, N. Farm size and productivity growth in the United States Corn Belt. Food Policy **84**, 186–195 (2019).
38. Sheng, Y. & Chancellor, W. Exploring the relationship between farm size and productivity: Evidence from the Australian grains industry. Food Policy **84**, 196–204 (2019).
39. Ortiz, R., Sayre, K. D., Govaerts, B., Gupta, R., Subbarao, G. V., Ban, T., Hodson, D., Dixon, J. M., Iván Ortiz-Monasterio, J. & Reynolds, M. Climate change: Can wheat beat the heat? Int. Agric. Res. Clim. Change Focus Trop. Syst. **126**, 46–58 (2008).
40. Newport, D., Lobell, D. B., Balwinder-Singh, Srivastava, A. K., Rao, P., Umashaanker, M., Malik, R. K., McDonald, A. & Jain, M. Factors Constraining Timely Sowing of Wheat as an Adaptation to Climate Change in Eastern India. Weather Clim. Soc. **12**, 515–528 (2020).
41. Akter, N. & Islam, R., M. Heat stress effects and management in wheat. A review. Agron. Sustain. Dev. **37**, 37 (2017).
42. Mullarkey, M. & Jones, P. Isolation and analysis of thermotolerant mutants of wheat. J. Exp. Bot. **51**, 139–146 (2000).
43. Tewolde, H., Fernandez, C. J. & Erickson, C. A. Wheat Cultivars Adapted to Post-Heading High Temperature Stress. J. Agron. Crop Sci. **192**, 111–120 (2006).
44. Ishtiaque, A., Singh, S., Lobell, D., Balwinder-Singh, Fishman, R. & Jain, M. Prior crop season management constrains farmer adaptation to warming temperatures: Evidence from the Indo-Gangetic Plains. Sci. Total Environ. **807**, 151671 (2022).

Figures

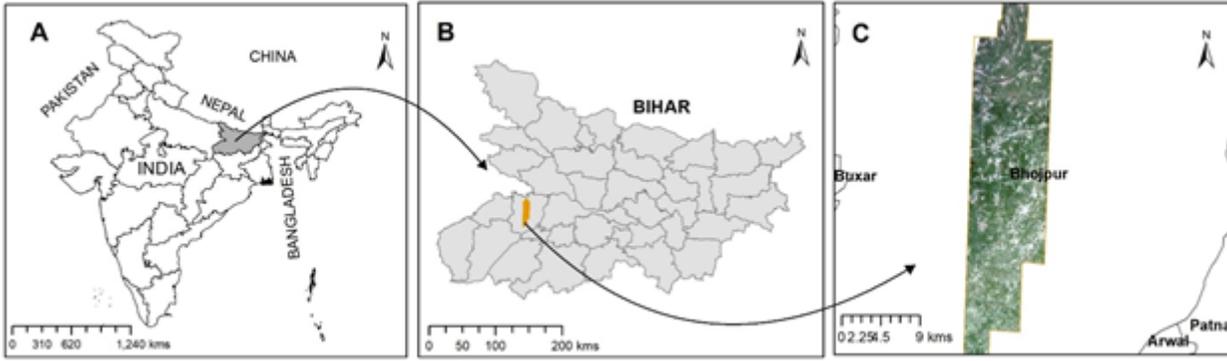


Figure 1

Map of India with Bihar highlighted in gray (Panel A) and the extent of the study region in Arrah district, Bihar highlighted in orange (Panel B). Panel C shows the PlanetScope image from February 15, 2018.

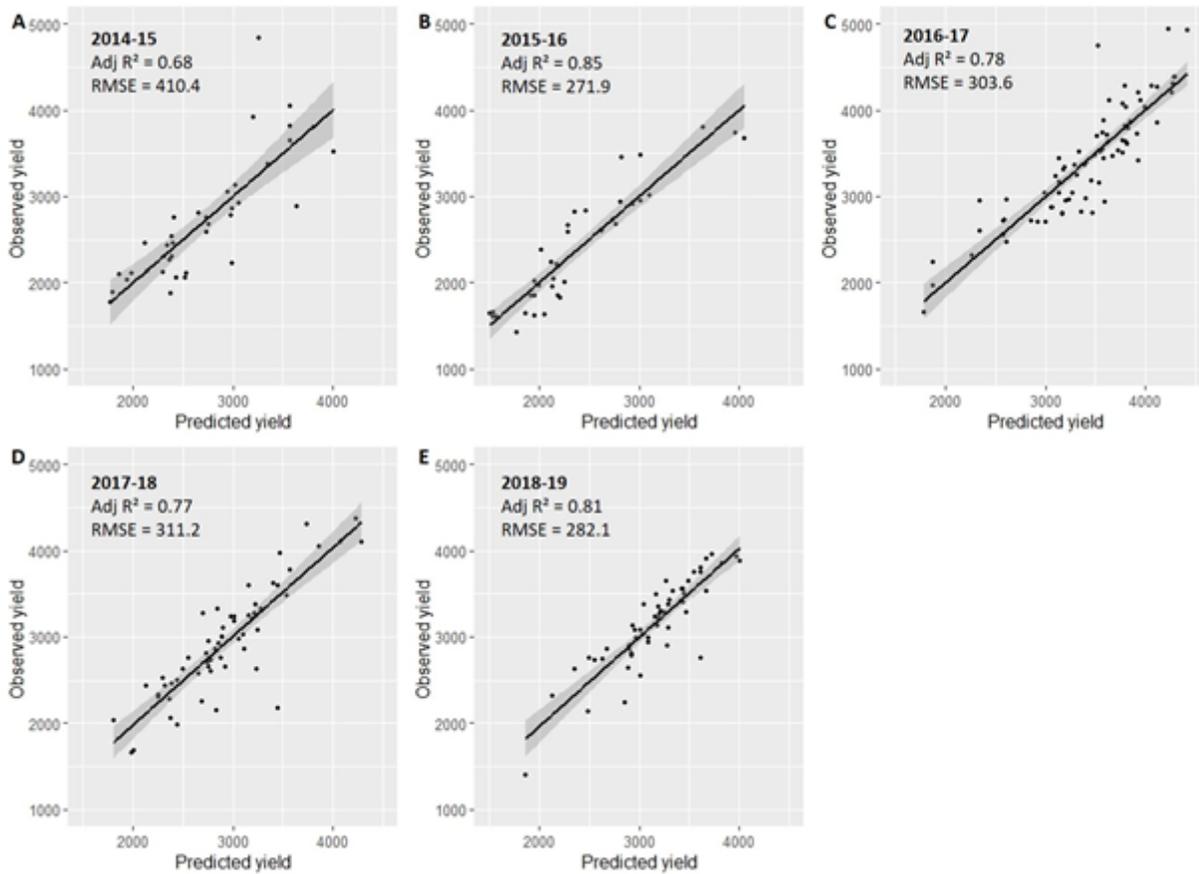


Figure 2

Plot level prediction accuracy for each year 2014-2015 to 2018-2019 (Panels A through E).

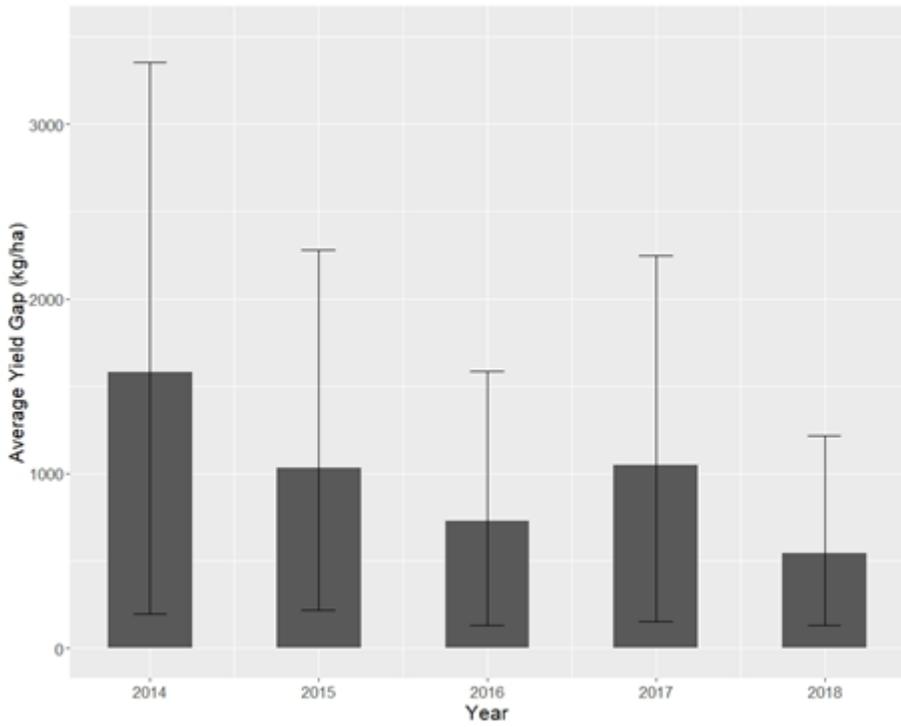


Figure 3

Average yield gap with 95% confidence intervals for each year.

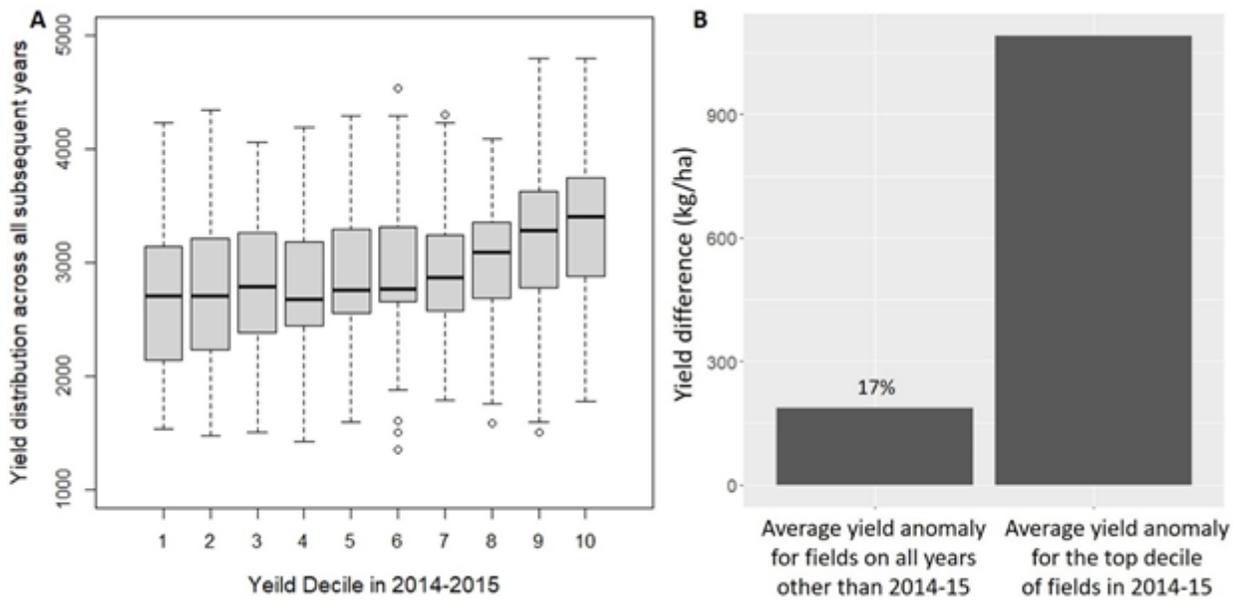


Figure 4

Yield distribution for the ten decile groups defined using 2014-15 yield estimates (Panel A). Average yield anomaly for the top decile of the fields in 2014-15 and average yield anomaly for the fields for all

remaining years (2015-16 to 2018-19, Panel B).

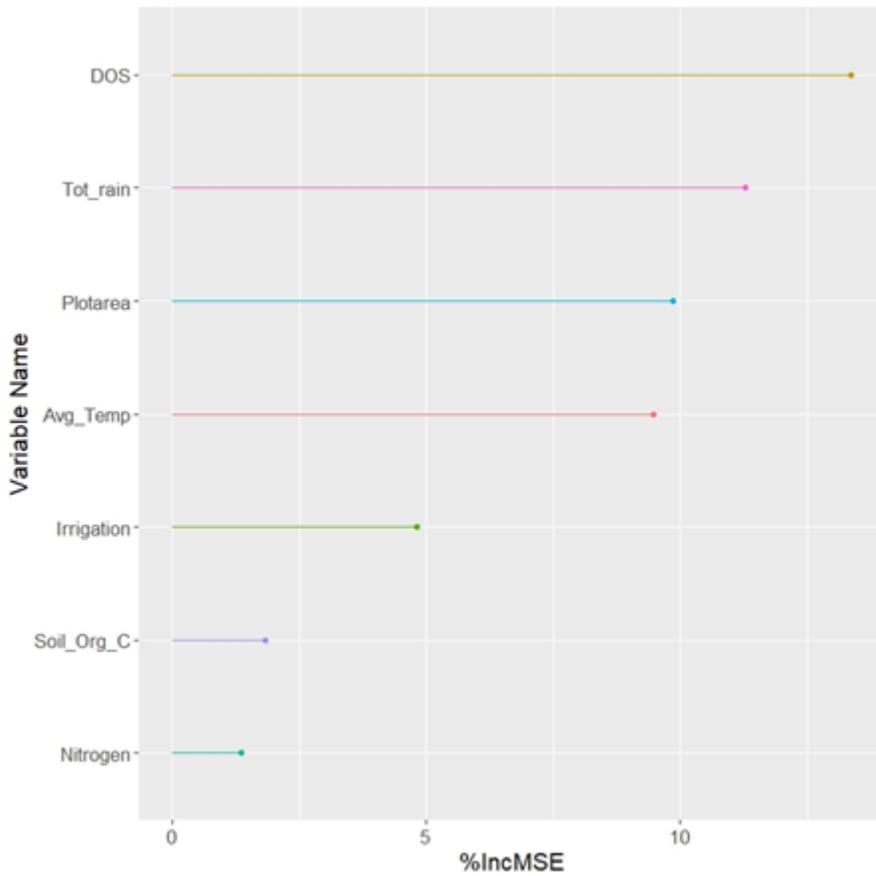


Figure 5

Variable importance plots for the management, weather, and biophysical factors considered to explain yield gaps for the years 2014-15 and 2015-16.

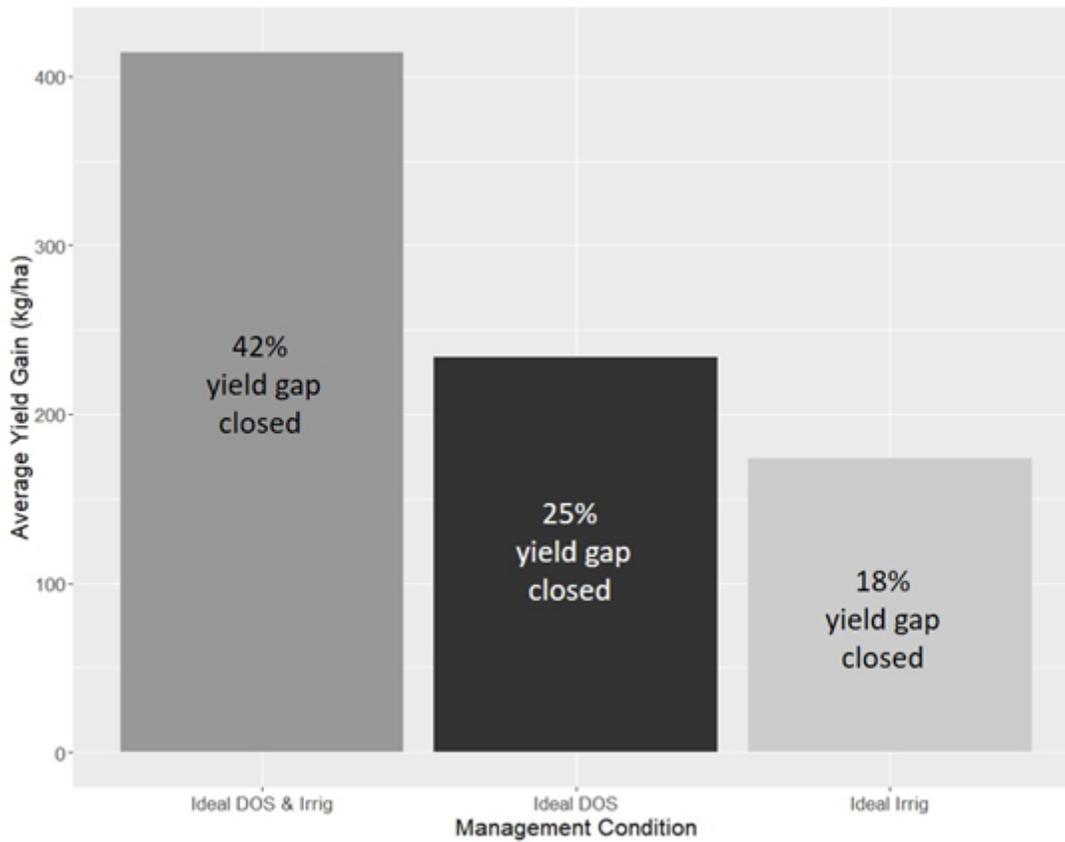


Figure 6

Average yield gain under three ideal management conditions (Ideal DOS represents sowing on November 12, and ideal irrig represents three irrigations).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryinfoPaliwalABSSPMJ.doc](#)