Using satellite data to identify the causes of and potential solutions for yield gaps in India’s Wheat Belt

M Jain1,2,4, Balwinder Singh1, A A K Srivastava1, R K Malik3, A J McDonald3 and D B Lobell2

1 School for Environment and Sustainability, University of Michigan, Ann Arbor, MI 48109, United States of America
2 Department of Earth System Science and Center on Food Security and the Environment, Stanford University, Stanford, CA 94305, United States of America
3 International Maize and Wheat Improvement Center (CIMMYT) New Delhi 110012, India
4 Author to whom any correspondence should be addressed.
E-mail: mehajain@umich.edu

Keywords: yield gap, food security, India, wheat, smallholder farms, remote sensing

Supplementary material for this article is available online

Abstract
Food security will be increasingly challenged by climate change, natural resource degradation, and population growth. Wheat yields, in particular, have already stagnated in many regions and will be further affected by warming temperatures. Despite these challenges, wheat yields can be increased by improving management practices in regions with existing yield gaps. To identify the magnitude and causes of current yield gaps in India, one of the largest wheat producers globally, we produced 30 meter resolution yield maps from 2001 to 2015 across the Indo-Gangetic Plains (IGP), the nation’s main wheat belt. Yield maps were derived using a new method that translates satellite vegetation indices to yield estimates using crop model simulations, bypassing the need for ground calibration data. This is one of the first attempts to apply this method to a smallholder agriculture system, where ground calibration data are rarely available. We find that yields can be increased by 11% on average and up to 32% in the eastern IGP by improving management to current best practices within a given district. Additionally, if current best practices from the highest-yielding state of Punjab are implemented in the eastern IGP, yields could increase by almost 110%. Considering the factors that most influence yields, later sow dates and warmer temperatures are most associated with low yields across the IGP. This suggests that strategies to reduce the negative effects of heat stress, like earlier sowing and planting heat-tolerant wheat varieties, are critical to increasing wheat yields in this globally-important agricultural region.

Introduction
Food security will be increasingly challenged over the coming decades (Godfray et al 2010). One way to enhance agricultural yields is to identify the extent and causes of existing gaps between attainable and currently achieved yields, and which management or biophysical factors are most associated with yield gaps (e.g. Zhao et al 2016). This can help provide actionable information on which management factors should be promoted to enhance food security. Conducting research on yield gaps is especially critical for India given that climate change impacts on agriculture are predicted to be especially large (Lobell et al 2008, Mall et al 2006) and food demand is increasing due to a burgeoning population (Lutz and KC 2010). Wheat yields, in particular, have stagnated (Ray et al 2012), and studies predict that warming temperatures could further reduce yields by up to 30% by mid-century (Ortiz et al 2008, Lobell et al 2008). Finding ways to enhance or maintain wheat yields is crucial to food security in India, where wheat provides 20% of household calories (Shiferaw et al 2013), to global food security, as India is the second largest wheat producer worldwide (FAO 2016), and to livelihoods, as over 70% of India’s rural population relies on agriculture as a primary source of income (Erenstein and Thorpe 2011).

© 2017 The Author(s). Published by IOP Publishing Ltd
Most previous studies that have examined the factors associated with yields in India have done so using district-level census data (e.g. Fishman 2012, Taraz 2017). Yet, given the large heterogeneity in both biophysical conditions and crop management across smallholder farms within districts (Jain et al 2015), coarse-scale census statistics may obscure important patterns that occur at finer spatial scales. While fine-resolution yield data can more accurately identify existing yield gaps and their associated drivers, these data do not exist for much of the developing world, including India. Satellite-based remote sensing offers one avenue for estimating yields, but is challenging in smallholder systems for several reasons. First, a majority of studies have relied on Moderate Resolution Imaging Spectroradiometer (MODIS) satellite data because its high temporal frequency provides imagery during important times in the crop growth cycle (e.g. peak greenness; Doraiswamy et al 2005) and can be used to quantify cumulative seasonal greenness (e.g. Duncan et al 2014). However, given that a majority of farms in India are smallholder (< 2 ha), the coarse spatial resolution of MODIS (250 meter) likely leads to inaccurate yield estimates due to sub-pixel heterogeneity (Jain et al 2013). Second, previous studies that have mapped yield typically do so by developing a relationship between satellite vegetation indices (VIs) and field-level yield data (e.g. Shanahan et al 2001, Quarmby et al 1993). However, these field data often do not exist in smallholder systems (Carletto et al 2013), making it difficult to translate satellite measures of greenness into yield. We overcome these challenges by employing Landsat satellite data and a new method termed the Scalable Crop Yield Mapper (SCYM; Lobell et al 2015) to estimate wheat yields at 30 meter resolution from 2001–2015 across the IGP. Instead of relying on ground-based yield training data, SCYM uses crop model simulations to create training data that can translate VIs to yield. Furthermore, these simulations allow for the use of imagery from any date within the crop’s growth cycle, resulting in more complete data coverage since Landsat imagery from different dates can be mosaicked across pixels. We use these high-resolution estimates of wheat yields to: (i) estimate spatial heterogeneity in yields across the IGP at 30 meter resolution, (ii) assess the management and biophysical variables that explain the largest amount of variance in yield, (iii) identify regions that are persistently low yielding, and (iv) quantify the extent to which yields can be increased through improved management. The results from these analyses provide actionable information that can be used to help design policies and extension initiatives, prioritize hotspots for intervention, and estimate the potential for closing yield gaps, which are crucial for bolstering food security in this globally significant agricultural region.

**Methods**

To translate satellite data into yields, we used the Scalable Crop Yield Mapper method (Lobell et al 2015). This method uses crop model simulations to simulate realistic field-level yield data, which are then used to train linear regressions that translate observed satellite vegetation indices to yield (figure S1 available at stacks.iop.org/ERL/12/094011/mmedia). To obtain realistic field-level yield data, we ran a suite of crop model simulations using the APSIM wheat crop model (Holzworth et al 2014, figure S1(a)). We parameterized the crop models (table S1) using a range of realistic management scenarios based on previous literature (Balwinder-Singh et al 2011, 2015) as well as information gathered from household surveys conducted in several districts across the IGP. Specifically we varied sow date, irrigation use, and the amount of fertilizer applied, and kept all other management factors constant, including wheat variety (PBW343, which is the dominant variety grown by farmers across the IGP; Joshi et al 2007). Furthermore, we used daily rainfall from APHRODITE (the Asian Precipitation Highly Resolved Observational Data Integration Towards Evaluation of the Water Resources), daily minimum and maximum temperature from NASA POWER (Prediction of Worldwide Energy Resource), and radiation data from NASA POWER for six randomly selected sites across the IGP (table S1). We ran simulations for each year from 1984 to 2007, ending our simulations in 2007 because NASA POWER used different data sources for radiation after 2007 and the data developers do not recommend using POWER weather data across this time period. We converted daily simulated wheat LAI from the crop models to the Green Chlorophyll Vegetation Index ($GCVI = (NIR/Green) - 1$) using equation (1), which was empirically derived from sub-field estimates (Nguy-Robertson et al 2014).

$$GCVI = \sqrt{\frac{\text{LAI} + 0.20654}{5.7}} - 0.1491 \quad (1)$$

We then developed linear regressions equation (2) that estimated simulated yield based on simulated logged $GCVI$ on a given date within an early and late-season time window ($GCVI_{t}$ represents $GCVI$ during the early growing season time window and $GCVI_{t^2}$ represents $GCVI$ during late season time window) and temperature (temp) experienced at the end of the growing season.

$$Yield = \log(GCVI_{t}) + \log(GCVI_{t^2}) + \text{temp} + \text{temp} \times \log(GCVI_{t}) + \text{temp} \times \log(GCVI_{t^2}) \quad (2)$$

To develop equation (2), we conducted internal model validation to identify the linear regression that resulted in the best model fit and highest yield prediction.
accuracies using the simulated data. We focused on two date windows, one early in the crop’s growth cycle (December 20 to January 20) and one during the time of peak crop growth (February 15 to March 15), as internal model validation suggests imagery from these two windows will result in high prediction accuracies (figure S2). We logged vegetation indices as unlogged values resulted in narrow predicted yield ranges, particularly in the eastern IGP. We also considered mean minimum end of season temperature (February 1–March 31) as previous research has shown terminal heat can have large impacts on wheat yields (Ortiz et al. 2008, Gourdji et al. 2012, Farooq et al. 2011, Reynolds et al. 1994). We used minimum (and not mean or maximum) temperature because internal model validation suggested all three weather variables resulted in similar predictive abilities and previous studies (Jain et al. 2016) have shown that MODIS LST, the satellite weather variable used in this study (figure S1(c)), is more closely correlated with minimum air temperature (when considering nighttime LST) than maximum air temperature (when considering daytime LST; Zhu et al. 2013). Internal model validation also suggested using a hinge function to model temperature (where temperature ≤ 15°C = 0 and temperature > 15°C = temperature − 15°C), given that the negative effects of temperature on yield became significantly greater starting at 15°C. It is important to note that we ran all possible date combinations between the first and second date windows for equation (2) to derive date-specific regression coefficients. This allows us to apply date-specific regression coefficients on a pixel-by-pixel basis depending on which dates of imagery are available and cloud free.

We processed GCVI from Landsat 5, 7, and 8 surface reflectance products and minimum temperature from the Oxford MODIS nighttime temperature product (Weiss et al. 2014) within Google Earth Engine (figure S1(b)). To limit analyses to agricultural pixels, we applied two masks. The first mask selected the maximum Normalized Difference Vegetation Index (NDVI) value for all Landsat imagery over all winter seasons from 2001 to 2015, and masked out any pixels that had an NDVI of 0.4 or lower. This mask reliably removed pixels that were non-vegetated, including urban areas or bare soil. Second, we created an annual cropped area mask using 16 day EVI MODIS data. Specifically we spline smoothed the MODIS time series from 2001 to 2015 using methods from Jain et al. (2013) and masked out any pixels where there was not a peak in vegetation crop growth for each winter season from 2001 to 2015. Previous work has shown that this method accurately detects seasonal cropped area in India (Jain et al. 2013). Next, given that there is a fair amount of cloud cover over the study region, we used pixel compositing techniques within Google Earth Engine to select cloud-free Landsat pixels within each image date window for each year. Specifically, for each date window, we selected the pixel that was both cloud free and had the highest GCVI during the time period of interest. If imagery were unavailable during the first time window (which occurred frequently due to higher cloud cover early in the season), we extended the first season window from January 20 to February 10. We applied the regression coefficients determined from our simulated yield regressions to Landsat GCVI and MODIS minimum temperatures within Google Earth Engine for each year (figure S1(c)), and then calculated mean yield over the course of the study period (2001–2015) to capture long-term spatial yield differences across the study region. Furthermore, it was not possible to produce annual yield estimates due to unavailability of Landsat satellite imagery in some locations and years. We validated 15 year mean satellite-based yield estimates with 13 year mean government census yield statistics reported at the district level from 2001–2013 and with household self-reported yield data collected in 12 districts across all four states in the IGP during the 2009–2010 wheat growing season.

To calculate sow date, we used the inflection point method identified as an effective way to map sow date in previous studies (Sakamoto et al. 2005). Specifically, we used the 16 day spline-smoothed MODIS EVI product from 2001 to 2015 described above (Jain et al. 2013), and identified the 2 week period in which an inflection point occurred between the end of October and the end of January for each year, representing the earliest and latest dates wheat was sown in the survey data. We validated sow date estimates with social survey sow date data across the IGP for 2009–2010, and were able to produce a similar distribution of sow dates (figure S3). We then aggregated these annual estimates to mean estimates from 2001–2015 to match the temporal resolution of the yield data product.

To better understand the factors associated with spatial yield variation, we conducted Random Forest analyses that examined the relative importance of a range of biophysical, socio-economic, and management variables in explaining village-level yield variation. To calculate variable importance, this analysis created 500 different decision trees using 500 random subsamples of the data and computed the prediction error on the out-of-bag portion of the data. It then does this analysis again after permuting each predictor variable. The difference between the full model and permuted model are averaged over all trees and normalized by the standard error of the differences. To calculate variance explained, this method considers the average variance explained in the out of bag sample across all trees. Random Forest accounts for potential non-linear relationships between the independent and dependent variables and also interactions between independent variables. All image processing was done in Google Earth Engine, and all model development and statistical analyses were done in R Project Software.
Results

Landsat satellite data and SCYM accurately map smallholder wheat yields

We find that the satellite-based yield estimates are accurate and capture spatial differences in yield, with correlation values ranging from 0.54 to 0.80 when validated with district-level census data (figure 1). Since there were some biases in satellite estimates that varied by state (e.g. lower predicted yields in Punjab and Haryana and higher predicted yields in Uttar Pradesh and Bihar) we applied a linear correction by state which resulted in the best fit line between remote sensing estimates and census data falling on the one-to-one line (following methods in Lobell and Azzari 2017). Using household survey data as validation, satellite yield estimates were also highly correlated (r = 0.92) at the district scale (figure S4) and were able to capture the distribution of sub-district yield variation (figure S5), suggesting that our yield estimates are accurate at the sub-district scale. It is important to note that this household survey dataset was not geo-located and therefore could not be used for field-level validation. Considering both validation datasets, satellite-derived yield estimates performed better in the western IGP in the states of Punjab and Haryana than in the eastern IGP. This discrepancy can be attributed to five factors: (i) government yield statistics may be less accurate and comprehensive in the east, as evidenced by less complete data coverage (figure S6); (ii) imagery is often not available in the east due to cloud cover and haze during the time periods that result in the highest yield prediction accuracies from the SCYM model (figure S2), (iii) imperfect screening for clouds and haze mean that the images we use in the east are potentially more affected by these factors; (iv) the average landholding size in the east is smaller (0.39 to 0.75 ha) than the west (2.25 to 3.77 ha; Mehrotra 2014), contributing to more mixed pixels in the east, and (v) the eastern IGP has a greater prevalence of non-wheat crops, which may complicate interpretation of Landsat data, although restricting the analysis to villages where wheat is reported to be planted according to census data did not significantly alter yield estimates (figure S7). Despite these limitations, the satellite estimates are
able to accurately capture the broad spatial variation in yields, even in the east where absolute yield estimates may be less accurate. Furthermore, it is likely that noise within the satellite estimates will not lead to inaccurate inferences about the factors associated with yield variation; they will only lead to conservative estimates of how much variance in yield these factors can explain assuming that the noise within the satellite yield estimates are not correlated with other predictor variables. To better understand the spatial variation in persistent yield differences, we focus on mean yield from 2001–2015 for all subsequent analyses, which gives an indication of the average yield of a given region across multiple years.

There is large within-district spatial heterogeneity in wheat yields across the IGP
As expected, we find that yields are much higher in the western IGP, particularly in the states of Punjab and Haryana. Yields can reach up to 6 tons ha\(^{-1}\) in the west, whereas yields are typically around 2 tons ha\(^{-1}\) in the east (figure 1). Importantly, there is also significant fine-scale, within-district heterogeneity in yields, including regions where there is a sharp discontinuity in yields directly across state borders, and these fine scale patterns can only be elucidated with high-spatial resolution yield maps. For example, the border of Haryana and Uttar Pradesh shows a sharp contrast in yields, with higher yields in Haryana and drastically lower yields directly across the border in Uttar Pradesh (figure 2(a)). In addition, while Bihar on average has the lowest yields in the IGP, there are small regions within the state that have yields as high as those typically found in Haryana and Punjab (figure 2(b)).

Remote sensing data elucidate regional differences and the importance of fine-scale factors for yield
To better understand which factors are associated with low versus high yielding regions, we conducted two

---

Figure 2. Results from uncalibrated yield estimates at finer spatial scales. Panel 2(a) highlights a sharp difference in yield across the Haryana and Uttar Pradesh state border. Panel 2(b) shows a region in Rohtas district in the state of Bihar that has yields as high as those typically found in Punjab and Haryana.
sets of analyses. First, we plotted the associations between a suite of weather, socio-economic, biophysical, and infrastructural variables (table S2; figure 3) and mean yield estimates at the village scale ($n = 160,014$), the smallest geographic level in India. While the direction of the relationship between each explanatory factor and yield is largely consistent across states, the magnitude of the effect varies from state to state. Considering infrastructural variables, irrigation is positively associated with yield in all regions, with villages that are closer to canals (figure 4(a)) and have increased area irrigated from any source (figure 4(b)) having higher yields, though this relationship is much stronger in the western IGP than the east. Biophysical factors also have strong associations with yield in all states (figures 4(c)–(e)), with higher monsoon rainfall, warmer temperatures, and higher elevations associated with reduced yields. Finally considering management factors, later sow dates at the village scale are associated with reduced yields (figure 4(f)).

In the second set of analyses, we examined the importance of each of these infrastructural, biophysical, and management variables in explaining yield variation using Random Forest analyses at the village scale (figures 5(a)–(d)). We are able to explain a large amount of variance—approximately 80% in the western IGP and 50% in the eastern IGP. A large proportion of this explained variance is due to biophysical factors, including temperature, rainfall, and elevation, with temperature coming out to be the most or second most important variable in each state. Interestingly, variables that vary at fine scales, such as sow date, soil type, and distance to canals, are also found to be important explanatory factors, with sow date being the most or second most important variable in each state. We also conducted Random Forest analyses with the remote sensing data and census data aggregated at the district scale, which allowed us to understand what information is lost when using typically-used census statistics. We find that biophysical variables, including minimum temperature, are similarly some of the most important factors, but the importance of factors that vary at fine scales, like sow date, are largely muted (table S3). By comparing the results of the remote sensing versus census data analyses, we can also further validate the accuracy of the remote sensing product. We find that both analyses explain approximately 90% of the variance in yield (figures 5(e) and (f)), and this similar predictive power suggests that the remote sensing data are as capable of capturing spatial differences in mean yield as traditionally used census datasets. To ensure that temperature was not driving the results, particularly since temperature was used within our SCYM yield training model, we ran all analyses without temperature and find that the results do not drastically change (figure S8).

Yields in the eastern IGP can likely be enhanced by improved management
To obtain yield gap estimates that reflect attainable yield given realistic constraints, we empirically define potential yield as the 95th percentile yield within a given district, and divide this value by the district's mean yield to estimate current yield gaps (figure 6(b); Lobell 2013). We find that the average yield gap across the IGP is 17% considering production-weighted averages. However, yield gaps vary widely across the
IGP with the largest gaps (up to 56%) occurring in the eastern IGP. Since biophysical factors, which are more difficult to manipulate through management, may be partially driving yield gap estimates, this analysis may overestimate the size of the yield gap that can actually be improved. Therefore, we also quantified the amount of variance explained by biophysical factors in each district (figure 6(c)), and then divided the residual yield gap from this analysis by district mean yield to estimate yield gaps driven by non-biophysical factors (figure 6(d)). The results suggest that yields may be improved on average by 11% and up to 32% in some areas when accounting for variation caused by known biophysical factors, with the largest proportion of high yield gaps (greater than 20%) occurring in Bihar.

It is likely that yields could be further improved if policies or technologies were introduced from other regions that changed current profit margins and/or system constraints. To capture this potential for longer-term yield improvement, we conducted a second analysis in which we defined potential yield across the IGP as the 95th percentile yield obtained in Punjab, the highest yielding state (figure 7). Since Punjab has higher yields partially due to its weather, particularly cooler temperatures (figure 4), we used crop model simulations to estimate the yield gap that was driven by differences in weather that cannot be improved by altering management. We find that using this approach, the potential for improving yields is larger, with production-weighted yield improvements reaching almost 110% in the eastern IGP (figure 7). It is important to note that yield gap estimates are influenced by census data calibration given that mean district-level yields differed between the remote sensing and census datasets, particularly in the eastern IGP (figure 1). Without calibration, satellite data suggest that yields in the eastern IGP may increase by almost 36% when improving yields to the 95th% yield in Punjab (figure 7).
This study examined the extent and causes of yield gaps in India’s main wheat belt, the IGP, and the results of this study have significant implications for both local and global food security. The results suggest that there is a large amount of heterogeneity in wheat yields across the IGP, with the largest yield gaps occurring in the eastern IGP. We estimate that yields can be increased by approximately 11% considering production-weighted averages and up to 32% in the eastern IGP if management is improved to current regional, within-district best practices. However, there is potential for longer-term yield gains of up to 110% if management in the eastern IGP could be made more similar to that in Punjab, the highest yielding state.

When examining which factors are associated with yield differences, we find that sow date and temperature (A) have significant impacts on yield variation. The figure below shows the relative importance of each variable considered in the model, with the most important variable having the largest % increase in mean squared error (MSE) between the full model and one where that variable is permuted. Average variance explained across all trees in the out of bag sample is recorded for both the full model (VE) as well as a model that only considers biophysical factors (VEb).
are the largest explanatory factors across all states. Warmer temperatures negatively impact wheat, particularly terminal heat stress that occurs at the end of the growing season (Asseng et al. 2014, Farooq et al. 2011, Reynolds et al. 1994). This is concerning given that current warming has already reduced wheat yields by 5% in northern India, and future warming will further hamper production (Lobell et al. 2011). Sow date is an important management factor that can mediate the effects of temperature, since negative effects are reduced if the crop matures prior to experiencing terminal heat stress. Based on this information, the largest yield gains could be achieved if farmers adopted strategies to reduce the negative impacts of heat stress. Considering existing strategies, farmers may sow wheat earlier, which allows the crop to mature prior to experiencing terminal heat

![Figure 6](image1.png)

**Figure 6.** Maps highlighting within district yield heterogeneity across the IGP based on SCYM satellite estimates. Panel (a) highlights mean yield differences across the study region. Panel (b) shows the difference (i.e. gap) between the mean and 95th percentile of yield within each district, with the latter used to indicate economically attainable yields. Panel (c) plots the percent of variance explained by biophysical factors using random forest analyses for each district. Panel (d) shows the difference between the mean and 95th percentile residual from this biophysical analysis, expressed as a percentage of the mean yield for that district. This value represents the amount of yield variation that is unexplained by biophysical factors, and thus may represent factors that can be improved through management, like sowing date and irrigation use.

![Figure 7](image2.png)

**Figure 7.** Yield gaps by state when potential yields are defined as the 95th percentile of yields in Punjab (pink) and the amount of yield gap that can be attributed to differences in weather (blue). The effect of weather was estimated by quantifying the influence of regional weather patterns on wheat yields using APSIM crop model simulations.
stress (Ortiz et al. 2008, Erenstein and Laxmi 2008). This is particularly important for the eastern IGP, where delayed planting of wheat is more common and temperatures are warmer than in the west (Ladha et al. 2003, Mathys and McDonald 2013). Future work is needed to better understand the factors constraining early sowing, but likely causes are delayed sowing of the previous monsoon crop due to delayed monsoon rains and long-duration rice varieties, as well as limited access to timely irrigation. In the longer term, more heat-tolerant wheat varieties that are being developed may be an additional way to enhance yields in the face of warming temperatures (Joshi et al. 2007).

Our results also show that irrigation variables are important factors in explaining yield variation within the IGP. Villages that have a greater proportion of area under irrigation or are closer to canals have higher yields. This suggests that a large proportion of farms across the IGP remain water limited, even in the heavily-irrigated states of Punjab and Haryana. Furthermore, one likely reason empirically-derived potential yields are significantly lower in Bihar than in Punjab is because of differences in irrigation access. In the western IGP, most farmers have access to a well-developed system of tubewells and canals, and are also provided with highly subsidized electricity for pumping irrigation (Scott and Sharma 2009, Shah et al. 2009). In the eastern IGP, farmers primarily rely on less developed canal networks and use unsubsidized and expensive diesel pumps to draw irrigation (Kishore 2004, Shah et al. 2009). It is therefore plausible that yields in Bihar could be raised higher than regional optimal levels if policies and technologies were introduced that enhanced irrigation access and reduced pumping costs similar to that in Punjab. However, while increasing irrigation use may improve yields, it may not be an optimal strategy given that current irrigation use is unsustainable in much of the IGP, particularly in the western states of Punjab and Haryana, and is leading to rampant groundwater depletion (Shah 2009, Rodell et al. 2009). Therefore, recommendations to increase irrigation access should not only consider the potential benefits to yields, but also the sustainability of the strategy and its impacts on natural resources.

When comparing the results from this study to previous studies, we find that temperature and irrigation access have similarly been found to be large contributors to yield variability and yield gaps globally (Mueller et al. 2012, Ray et al. 2015). Yet, by using high-resolution satellite data, this study provides information that is often masked in coarser-scale analyses; we are able to highlight the importance of factors that vary at fine-spatial resolutions, such as sow date. In addition, in comparison to previous studies, we estimate slightly smaller yields gaps in the western IGP and similar yield gaps in the eastern IGP. Previous studies have suggested that wheat yield gaps are approximately 50% of mean yields in the western IGP, and over 100% of mean yields in the east (Aggarwal and Kalra 1994, Pathak et al. 2003). There are several potential reasons for this difference in yield gap estimates. First, previous studies may overestimate yield gaps as they often define potential yields using ideal management scenarios within crop model simulations, which may not be realistic of real-world conditions. For example, studies have shown that ideal management at controlled experimental plots within the IGP often reach only 80% of these climatic potential yields (e.g. Timsina et al. 2008). Furthermore, these previous estimates do not account for profit margins or system constraints that farmers consider when making real-world management decisions (Lobell et al. 2009). For example, farmers may not apply optimal amounts of irrigation either because they do not have access to enough irrigation or because the cost of pumping makes applying additional irrigations unprofitable. Second, our yield gap measures are likely conservative compared to yield gaps that are estimated for an individual year. By taking the mean yield over fifteen years, we inherently reduce the magnitude of yield gap since previous studies have shown that the same fields do not fall within the 95th percentile every year. Using estimates from previous studies, mean year yield gaps are typically 80% of annual yield gaps (Farmaha et al. 2016, Zhao et al. 2016).

Our results highlight the benefits of high-resolution yield maps, as they elucidate productivity and their associated factors that are obscured by coarser level census statistics. This is especially crucial for smallholder systems, where field or sub-district level data are often non-existent (Carletto et al. 2013). For example, in India's wheat belt, there are no high-resolution datasets (e.g. social survey data) that offer sub-district yield information across large spatial and temporal scales. The SCYM method overcomes data limitation in these systems by producing high-resolution yield estimates from satellite data, even when crop management is heterogeneous and training data are unavailable. These satellite-produced yields maps can serve as an important tool for understanding existing yield gaps and the potential factors that may enhance yields over the coming decades. Specifically, they allow one to better detect the effect of factors that vary at fine spatial resolutions, like sow date and soil type (figure 5), which are otherwise muted in coarser-scale analyses. These data also elucidate discontinuous variation across space (figure 2) that may be used to correlate and possibly causally identify factors that are associated with improved yields. Finally, these data are helpful for understanding location-specific yield gaps as they allow for the selection of realistic optimal yields that are appropriate for each district (figure 6). In summary, producing high-resolution yield estimates can help provide actionable information by identifying regions that are consistently low-yielding,
which factors are associated with low yields, and potential interventions that may be promoted to enhance yields and food security over the coming decades.

Acknowledgments

We would like to thank the field staff of the Cereal Systems Initiative for South Asia (www.csisa.org) for collecting and sharing field-level yield data used for validation, as well as George Azzari for helpful code to implement SCYM and yield visualizations in Google Earth Engine. This work was funded by an NSF SEES Postdoctoral Fellowship (Award Number 1415436) awarded to M Jain.

References

Aggarwal P K and Kalra N 1994 Analyzing the limitations set by climatic factors, genotype, and water and nitrogen availability on productivity of wheat II. climatically potential yields and management strategies Field Crops. Res. 38 93–109

Asseng S et al 2014 Rising temperatures reduce global wheat production Nat. Clim. Change 4 143–7

Balwinder Singh, Gaydon D S, Humphreys E and Eberbach P L 2011 The effects of mulch and irrigation management on yield of rice, wheat and groundnut in semi-arid climate Agric. Syst. 104 42–53

Doraiswamy P C, Sinclair T R, Hollinger S, Akhmedov B, Stern A and Prueger J 2005 Application of MODIS derived parameters for regional crop yield assessment Remote Sens. Environ. 97 192–202

Duncan J M A, Dash J and Atkinson P M 2014 Elucidating the impact of temperature variability and extremes on cereal crops through remote sensing Glob. Change Biol. 20 1541–51

Erenstein O and Lasmi V 2008 Zero tillage impacts in India’s rice–wheat systems: a review Soil Tillage Res. 100 1–14

Erenstein O and Thorpe W 2011 Livelihoods and agro-ecological gradients: a meso-level analysis in the indo-gangetic plains, India Agr. Syst. 104 42–53

Farmaha B S, Lobell D B, Boone K E, Cassman K G, Yang H S and Grasmann P 2016 Contribution of persistent factors to yield gaps in high-yield irrigated maize Field Crops. Res. 186 124–32

Farooq M, Bramley H, Palta J A and Siddique K H M 2011 Heat stress in wheat during reproductive and grain-filling phases Curr. Rev. Plant. Sci. 10 491–507

Fishman R M 2012 Climate change, rainfall variability, and adaptation through irrigation: evidence from Indian agriculture Working Paper

Food and Agriculture Organization of the United Nations 2016 FAOSTAT Database on Agriculture (httpfaostat.fao.org) (www.fao.org/faostat)

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 Tolunay C 2010 Food security: the challenge of feeding 9 billion people Science 327 812–8

Gourdji S M, Mathews K L, Reynolds M, Cossa J and Lobell D B 2012 An assessment of wheat yield sensitivity and breeding gains in hot environments Proc. R. Soc. B Biol. Sci. 280 20122190

Holzworth D P et al 2014 APSIM—Evolution towards a new generation of agricultural systems simulation Environ. Modell. Softw. 62 327–50

Jain M, Mondal P, DeFries R S, Small C and Galford G L 2013 Mapping cropping intensity of smallholder farms: a comparison of methods using multiple sensors Remote Sens. Environ. 134 210–3

Jain M, Naeem S, Orlove B, Modi V and DeFries R S 2015 Understanding the causes and consequences of differential decision-making in adaptation research: adapting to a delayed monsoon onset in Gujarat, India Glob. Environ. Change 31 98–109

Jain M, Srivastava A K, Singh Balwinder, Joon R K, McDonald A, Royal K, Lsisais M C and Lobell D B 2016 Mapping smallholder wheat yields and sowing dates using micro-satellite data Remote Sens. 8 860 (http://www.mdpi.com/2072-4292/8/10/860)

Joshi A K, Mishra B, Chatzath R, Ortiz Ferrara G and Singh R P 2007 Wheat improvement in India: present status, emerging challenges and future prospects Euphytica 157 431–46

Kishore A 2004 Understanding agrarian impasse in Bihar Econ. Polit. Weekly 39 3484–91

Ladha J K, Hill J E, Duxbury J M, Gupta R K and Buresh R J 2003 Improving the Productivity and Sustainability of Rice– Wheat Systems: Issues and Impacts (Madison, WI: American Society of Agronomy, Inc., Crop Science Society of America, Inc., Soil Science Society of America, Inc.)

Lobell D B 2013 The use of satellite data for crop yield gap analysis Field Crops. Res. 143 56–64

Lobell D B and Azzari G 2017 Satellite detection of rising maize yield heterogeneity in the US midwest Environ. Res. Lett. 12 014014

Lobell D B, Burke M B, Tebaldi C, Mastrandrea M D, Fakon W P and Naylor R L 2008 Prioritizing climate change adaptation needs for food security in 2030 Science 319 607–10

Lobell D B, Cassman K G and Field C B 2009 Crop yield gaps: their importance, magnitudes, and causes Annu. Rev. Environ. Resour. 34 179–204

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, Thu D, Seifert C,_engle E and Little B 2015 A scalable satellite-based crop yield mapper Remote Sens. Environ. 164 324–33

Lutz W and Samir K C 2010 Dimensions of global population projections: what do we know about future population trends and structures Philos. Trans. R. Soc. B Biol. Sci. 365 2379–91

Mall R K, Singh R, Gupta A, Srinivasan G and Rathore L S 2006 Impact of climate change on Indian agriculture: a review Clim. Change 78 445–78

Mathys C and McDonald A 2013 Cereal Systems Initiative for South Asia (CSISA). Annual Report

Menhoura N 2014 Agricultural Land Holdings Pattern in India Report (Mumbai: Department of Economic Analysis and Research, NABARD)

Mueller N, Gerber J S, Johnston M, Ray D K, Ramankutty N and Foley J A 2012 Closing yield gaps through nutrient and water management Nature 490 254–7

NguyenRobertson A L, Peng Y, Gitelson A A, Arkebauer T J, Pimstein A, Herrmann I, Karmeli A, Rumquist D C and Bonfil D J 2014 Estimating green LAI in four crops: potential of determining optimal spectral bands for a universal algorithm Agr. Forest Meteorol. 192–193 140–8

Ortiz R, Sayre K D, Govaerts B, Gupta R, Subbarao G V, Yan T, Hodson D, Dixon J M, Ivan Ortiz-Monasterio J and Reynolds M 2008 Climate change: can wheat beat the heat? Agric. Ecosyst. Environ. 126 46–58
Pathak H et al 2003 Trends of climatic potential and on-farm yields of rice and wheat in the indo-gangetic plains Field Crops. Res. 80 223–34
Quarmby N A, Milnes M, Hindle T L and Silleos N 1993 The use of multi-temporal NDVI measurements from AVHRR data for crop yield estimation and prediction Int. J. Remote. Sens. 14 199–210
Ray D K, Ramankutty N, Mueller N D, West P C and Foley J A 2012 Recent patterns of crop yield growth and stagnation Nat. Commun. 3 1293
Ray D K, Gerber J S, MacDonald G K and West P C 2013 Climate variation explains a third of global crop yield variability Nat. Commun. 6 5989
Reynolds M P, Balota M, Delgado M, Amani I and Fischer R A 1994 Physiological and morphological traits associated with spring wheat yield under hot, irrigated conditions Aust. J. Plant. Physiol. 21 717–30
Rodell M, Velicogna I and Famiglietti J S 2009 Satellite-based estimates of groundwater depletion in India Nature 460 999–1002
Sakamoto T, Yokozawa M, Toritani H, Shibayam M, Ishitsuka N and Ohno H 2005 A crop phenology detection method using time-series MODIS data Remote Sens. Environ. 96 366–74
Scott C A and Sharma B 2009 Energy supply and the expansion of groundwater irrigation in the indus-ganges Basin Int. J. River Basin Manage. 7 119–24
Shah T 2009 Climate change and groundwater: India’s opportunities for mitigation and adaptation Environ. Res. Lett. 4 035005–14
Shah T, Hassan M U, Khattak M Z, Banerjee P S, Singh O P and Rehman S U 2009 Is irrigation water free? a reality check in the indo-gangetic Basin World Dev. 37 422–34
Shanahan J F, Schepers J S, Francis D D, Varvel G E, Wilhelm W W, Tringe J M, Schlemmer M R and Major D J 2001 Use of remote-sensing imagery to estimate corn grain yield Agron. J. 93 583–9
Shiferaw B, Smale M, Braun H-J, Duveiller E, Reynolds M and Muricho G 2013 Crops that feed the world 10. past successes and future challenges to the role played by wheat in global food security Food Sec. 5 291–317
Taraz V 2017 Adaptation to climate change: historical evidence from the Indian monsoon Environ. Dev. Econ. (https://doi.org/10.1017/S1355770X17000195)
Timsina J, Godwin D, Humphreys E, Yadavinder Singh, Bijay Singh, Kukal S S and Smith D 2008 Evaluation of options for increasing yield and water productivity of wheat in Punjab, India using the DSSAT-CSM-CERES-Wheat model Agr. Water Manage. 95 1099–110
Weiss D J, Atkinson P M, Bhatt S, Mappin B, Hay S I and Gething P W 2014 An effective approach for gap-filling continental scale remotely sensed time-series ISPRS J. Photogramm. Remote Sens. 98 106–18
Zhao Y, Chen X and Lobell D B 2016 An approach to understanding persistent yield variation—a case study in North China plain Eur. J. Agron. 77 10–9
Zhu W, Lu A and Jia S 2013 Estimation of daily maximum and minimum air temperature using MODIS land surface temperature products Remote Sens. Environ. 138 62–73