Title
The Value of Light: Crop Response to Optical Scattering and Generalizable Earth Observation

Author
Proctor, Jonathan Neel

Publication Date
2019

Peer reviewed|Thesis/dissertation

The Value of Light:
Crop Response to Optical Scattering and Generalizable Earth Observation

by

Jonathan Neel Proctor

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Agricultural and Resource Economics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Solomon M. Hsiang, Co-chair
Professor Maximilian Auffhammer, Co-chair
Professor Meredith Fowlie
Professor Benjamin Recht

Summer 2019

The Value of Light:
Crop Response to Optical Scattering and Generalizable Earth Observation

Abstract

The Value of Light:
Crop Response to Optical Scattering and Generalizable Earth Observation

by

Jonathan Neel Proctor

Doctor of Philosophy in Agricultural and Resource Economics

University of California, Berkeley

Professor Solomon M. Hsiang, Co-chair

Professor Maximilian Auffhammer, Co-chair

How does human manipulation of the quantity, directionality and spectral distribution of sunlight affect global agricultural productivity? And how might we build a global observation system to provide measurements to answer this and other key questions in environmental science, economics and policy? This thesis quantifies the impact that atmospheric scattering from volcanic sulfate aerosols and clouds has on global crop yields. In turn, this work informs how anthropogenic influences on the global optical environment – from geoengineering to air pollution to climate change – impact global food production and food security. This thesis also develops a system that leverages satellite imagery and machine learning to measure many social and environmental variables with high skill, low cost and no alteration of method. The hope is that the generalizablity, low cost, and simplicity of this system will democratize remote sensing, and accelerate the pace of research into our Earth's socio-environmental systems. Broadly, I hope this thesis contributes to environmental policy and improves the wellbeing of life on Earth. Chapter 1 provides the broad-scale motivation for my work.

Chapter 2 studies the agricultural impacts of Solar radiation management (SRM), which is increasingly considered as an option for managing global temperatures. Yet the economic impacts of ameliorating climatic changes by scattering sunlight back to space remain largely unknown. Though SRM may increase crop yields by reducing heat stress, its impacts from concomitant changes in available sunlight have never been empirically estimated. Here we use the volcanic eruptions that inspired modern SRM proposals as natural experiments to provide the first estimates of how the stratospheric sulfate aerosols (SS) created by the eruptions of El Chichón and Mt. Pinatubo altered the quantity and quality of global sunlight, how those changes in sunlight impacted global crop yields, and the total effect that SS may have on yields in an SRM scenario when the climatic and sunlight effects are jointly considered. We find that the sunlight-mediated impact of SS on yields is negative for both C4 (maize) and C3 (soy, rice, wheat) crops. Applying our yield model to a geoengineering scenario using SS-based SRM from 2050-2069, we find that SRM damages due to scattering sunlight are roughly

equal in magnitude to SRM benefits from cooling. This suggests that SRM – if deployed using SS similar to those emitted by the volcanic eruptions it seeks to mimic – would attenuate little of the global agricultural damages from climate change on net. Our approach could be extended to study SRM impacts on other global systems, such as human health or ecosystem function.

Chapter 3 explores how anthropogenic emissions of air pollutants and greenhouse gases alter the amount, distribution and properties of cloud cover and, in turn, agricultural productivity. Changing cloudiness may impact crop productivity by altering temperature, precipitation and sunlight. While the impacts of temperature and precipitation on crop productivity are relatively well understood, the impacts of cloud optical scattering have never been empirically estimated and remain poorly constrained because of the potentially offsetting effects of changes in total and scattered sunlight. Here, I leverage remotely-sensed cloud observations and subnational crop yield data to provide the first empirical estimates of the sunlight-mediated efffect of cloud optical scattering on maize and soy yields in the United States, Europe, Brazil, and China. I find a consistent concave response of yields to cloud optical thickness across crops and regions. Changing ten days in the growing season from clear to the optimal cloud thickness increases maize and soy yields by 4.0% and 4.4%, respectively; further increasing cloud thickness to the 95th growing season percentile decreases maize and soy yields by 3.4% and 3.5%. Mechanistically, I find that the concavity in the cloud response is driven by concavity in the response to total sunlight as well as – in some regions – benefits from increased diffuse light. Applying these empirical estimates to earth system model simulations, I find that changes in sunlight, due to anthropogenic air pollution-induced changes in clouds, are suppressing maize and soy yields by as much as 5% in heavily polluted areas of India and China by increasing the frequency of days with extremely high cloud optical depths. This costs Chinese maize farmers roughly US$1 billion a year. Changes in sunlight due to changes in clouds from a quadrupling of CO_2 relative to pre-industrial tend to decrease global maize yields and redistribute soy yields. The methodology developed in this chapter could be extended study the impact of changes to the global optical environment on other global-scale economic outcomes.

Chapter 4 develops a system combining satellite imagery with machine learning (SIML) to observe many variables simultaneously. Current case-by-case solutions require custom systems, extensive expert knowledge, access to imagery, and major computational resources in order to estimate a single variable (a *task*) using regional or global imagery. Here, we develop a general solution to constructing global observations via SIML, where a single method for transforming satellite imagery is sufficiently descriptive that it should be able to predict nearly any ground-level variables that are recoverable through inspection of a satellite image, including previously unstudied tasks. Our approach is task-independent, allowing centralized computation of features to be executed only once ever per image, then distributed and applied to potentially unlimited future tasks by users who require neither domain expertise nor access to underlying imagery. We demonstrate this generalizability across tasks by constructing high resolution ($\sim$1km$\times$1km) estimates for forest cover, population density, elevation, nighttime lights, household income, total road length, and housing prices across the entire US

using exclusively daytime images that are processed only once and in advance. Our system outperforms spatial extrapolation of ground-truth data, especially over large distances, and matches or exceeds performance of a state-of-the-art deep convolutional neural network that is much more costly to implement. Our approach requires only that users download a tabular data set, merge it to geolocated labels, and implement a single regression on a personal computer. We demonstrate that our design scales globally with no alterations and naturally achieves super-resolution, where estimates are more spatially granular than the original labels used for training. Generalization enables democratization of SIML, potentially increasing the pace of planet-scale observation and research, accelerating our understanding of global processes and enabling progress towards tackling planetary challenges.

To my family, whose love for me and the world supports and inspires all that I do.

Contents

Acknowledgments

Thank you to my family, mentors, friends and peers whose intellectual and physical shoulders I have leaned, cried, laughed and sat upon. You bring light to life and made this dissertation possible.

I cannot give enough thanks to my mentor, colleague and friend Solomon Hsiang. I appreciate your dedication to the purpose, quality and impact of my work. The countless hours you invest in teaching your students to pose questions, uncover answers and communicate findings are a supreme service. I'll do my best to pay it forward.

I am filled with gratitude from the many other faculty who have guided, supported and inspired my work. I thank Max Auffhammer for sharing his interdisciplinary understanding and joyful encouragement over five years and two departments. I thank Ben Recht for being my guide and interpreter when exploring the intersection of machine learning and environmental economics. And I thank Meredith Fowlie for providing an insightful economic perspective on my work. I thank Jen Burney for her joyful support, intelligent advice, and help navigating the interdisciplinary maze of academia. I thank Marshall Burke and Wolfram Schlenker for their thoughtful insights and guidance. I thank Dennis Baldocchi for many fascinating conversations about biometerology, and Bill Collins for his kindness and his insights on cloud scattering. I thank Roz Naylor for seeding and nourishing my interest in global food security and for being a generous and skilled mentor to me and the community.

Thank you to the Global Policy Lab for being my intellectual home. In particular, I thank Tamma Carleton and Ian Bolliger for helping understand, navigate and balance my work and my life. I also very much appreciated, enjoyed and learned from many conversations with Andy Hultgren, Jonathan Kadish, and Hannah Druckenmiller. I thank Esther Rolf and Vaishaal Shankar for helping us bridge machine learning and economics.

Thank you to ARE and to the 2014 cohort for making the first two grueling years of classes a blast. And thank you to ESPM for being my first intellectual home at Berkeley.

I deeply appreciate my family of friends who give my life meaning and joy. You have taught me so much, and I look forward to our adventures to come.

And thank you to my family, whose decades of love and support live in everything I do.

Chapter 1

Introduction

The motivation for this dissertation stems from a belief that government policies can promote human wellbeing by nurturing a healthy economy and environment. To design effective environmental policies, however, we need a quantitative understanding of how socio-environmental systems function. For example, Pigouvian taxes on pollutants like carbon dioxide require an understanding of the marginal social cost of such emissions to (theoretically) maximize aggregate wellbeing. Further, including distributional impacts into policy design requires an even deeper understanding of our Earth's systems.

While rigorous empirical analyses to inform such global environmental policies has historically tended to be prohibitively costly, recent advances in data acquisition, storage, and processing technologies facilitate sophisticated analyses of socio-environmental systems. By combining global measurements of climate, volcanogenic aerosol, and yield data, for example, the first chapter of this thesis estimates the impact of a potential geoengineering deployment on global agricultural productivity. Such estimates have the potential to inform environmental policies regulating the research or deployment of geoengineering. Similarly, my second chapter estimates the impacts of cloud optical scattering on yields and improves our understanding of how air pollution and climate change shape global agricultural productivity. In turn, my hope is that this informs calculations of the marginal damage of anthropogenic air pollution and greenhouse gas emissions to global wellbeing, and in turn, environmental policy.

A key constraint to these types of analyses, however, is that researchers can only study what is measured. Much of the empirical climate impacts literature, for example, focuses on data-rich topics and locations such as agriculture or health in the United States and Europe. Variables such as crop yields, mortality, or energy use have been recorded in detail for decades because of their importance, yet many other key facets of wellbeing, such as biodiversity, human population density, or built capital have received less research and attention because of their lack of structured and precise measurement.

Remote sensing, and specifically recent advances combining satellite imagery and machine learning, may enable substantially improved measurement of key social and environmental variables. The final chapter of my dissertation proposes and demonstrates the effectiveness of a user-friendly, generalizable and cheap system to measure environmental and social variables at the global scale. Such measurements have the potential to accelerate our understanding of socio-environmental systems and thus engender effective policy.

Chapter 2

Estimating global agricultural impacts of geoengineering using volcanic eruptions

Solar radiation management (SRM) is increasingly considered an option for managing global temperatures [23, 74], yet the economic impacts of ameliorating climatic changes by scattering sunlight back to space remain largely unknown [62]. Though SRM may increase crop yields by reducing heat stress [82], its impacts from concomitant changes in available sunlight have never been empirically estimated. Here we use the volcanic eruptions that inspired modern SRM proposals as natural experiments to provide the first estimates of how the stratospheric sulfate aerosols (SS) created by the eruptions of El Chichón and Mt. Pinatubo altered the quantity and quality of global sunlight, how those changes in sunlight impacted global crop yields, and the total effect that SS may have on yields in an SRM scenario when the climatic and sunlight effects are jointly considered. We find that the sunlight-mediated impact of SS on yields is negative for both C4 (maize) and C3 (soy, rice, wheat) crops. Applying our yield model to a geoengineering scenario using SS-based SRM from 2050-2069, we find that SRM damages due to scattering sunlight are roughly equal in magnitude to SRM benefits from cooling. This suggests that SRM – if deployed using SS similar to those emitted by the volcanic eruptions it seeks to mimic – would attenuate little of the global agricultural damages from climate change on net. Our approach could be extended to study SRM impacts on other global systems, such as human health or ecosystem function.

This chapter is joint work with Solomon Hsiang, Jennifer Burney, Marshall Burke, and Wolfram Schlenker. It was published in the journal *Nature* in 2018 and is available at https://doi.org/10.1038/s41586-018-0417-3

Geoengineering, the purposeful alteration of the climate to offset changes induced by greenhouse gas emissions, is a proposed but still poorly understood approach to limit future warming [76]. One of the most widely suggested geoengineering strategies is "solar radiation management" (SRM). SRM proposals typically involve spraying precursors to sulfate aerosols into the stratosphere to produce particles that cool the earth by reflecting sunlight back into space [94]. The closest natural analogs to these SRM proposals are major volcanic eruptions [93]. Eruptions of El Chichón (1982, Mexico) and Mt. Pinatubo (1991, the Philippines) injected 7 and 20 Mt of sulfur dioxide into the atmosphere, respectively, which was oxidized to form stratospheric sulfate aerosols (SS) [92]. These particles propagated throughout the tropics over several weeks and spread latitudinally over the following months, increasing the opacity of the stratosphere – as measured by optical depth – more than an order of magnitude above baseline levels for multiple years (Fig. 2.1 a-c,e).

The eruptions of El Chichón and Pinatubo had substantial impacts on the global optical environment and climate. We analyze daily data from 859 insolation stations (N= 3,311,553; Fig. 2.1d) [129] paired with stratospheric aerosol optical depth (SAOD) [101] and cloud fraction data under all-sky conditions. We find that the Pinatubo eruption (global avg. +0.15 SAOD) reduced direct sunlight 21%, increased diffuse sunlight 20%, and reduced total sunlight 2.5% (Fig. 2.1f, Extended Data Table 2.1, Supplementary Information II). These global all-sky results generalize previous clear-sky estimates at individual stations [25] (Supplementary Information II.1). Globally, this reduction in insolation led to cooling of ∼0.5C [92] and redistribution and net reduction of precipitation [115], effects that were partially offset by a concurrent El Niño event (Fig. 2.2). Based on these observations, it has been suggested that SRM cooling could mitigate agricultural damages from global warming [82]. The net effect of SRM, however, remains uncertain due to possible unintended consequences from SS-induced changes. Here we empirically estimate how alteration of sunlight by SS may directly affect agricultural yields, after accounting for impacts mediated by temperature, precipitation and clouds.

The sign of SRM's "insolation effect" on agriculture is theoretically ambiguous [35, 95, 36, 37]. Scattering light decreases total available sunlight, which tends to decrease photosynthesis, but also increases the fraction of light that is diffuse, which can increase photosynthesis by redistributing light from sun-saturated canopy leaves to shaded leaves below [36, 96]. It is unknown whether damages from decreasing total light or benefits from increasing diffuse light dominate in the production of crop yield. The sign of this insolation effect will depend primarily on two factors: the forward-scattering properties of the aerosol and the relative benefit of diffuse light for the growth of edible yield (Supplementary Information III.5). The latter may depend on canopy geometry, photosynthetic pathway (e.g. C3 or C4), and ambient conditions [35, 7]. Previous studies of unmanaged ecosystems tend to find that scattering increases biomass growth [66, 36], though not always [7], and importantly, edible yield production may not directly correlate with biomass growth. Studies of agricultural systems tend to estimate negative impacts of tropospheric aerosol scattering [37, 35] and positive effects of solar brightening [114] on yields. Simulations of potential SRM impacts focus on cooling

and precipitation effects [131] and suggest global yields may increase due to cooling [82], although these analyses do not account for the full effect of scattering. This is the first study to estimate and account for the net effects of SS radiative scattering on yields, thereby testing whether the benefits of SS scattering demonstrated in unmanaged ecosystems [36, 66] also apply to agricultural production, as is often hypothesized [82, 95]. This analysis is the first global empirical study of the insolation effect on crops as well as the first study to leverage a quasi-experimental design to estimate the total impact of SRM on any economic sector.

The theoretically ideal experiment would measure the total effect of SRM on yields using many identical Earths, half treated with SS. In practice, we approximate this experiment with one Earth during sequential periods of high and low SS exposure, exogeneously determined by volcanic eruptions. We identify the insolation effect of SS on yields (Extended Data Fig. 2.1) [27] by comparing countries to themselves over time with changing SS treatment–measured in SAOD composited from satellite and other observations (Fig. 2.1e) [101]–while controlling flexibly for potentially confounding climate variables including temperature, precipitation, cloud fraction, and the El Niño-Southern Oscillation (ENSO) (Supplementary Information III.3). Our multivariate fixed-effects panel estimation strategy (Supplementary Information Eqn. 2.16) accounts for unobserved time-invariant factors, such as soil type or historical propensity for civil unrest, as well as country-specific time-trending variables, such as access to fertilizers or trends in damaging tropospheric ozone [46]. Our primary analysis focuses on the Pinatubo eruption because the concentration and distribution of resulting SS were substantially more accurately measured than earlier eruptions [113]. We validate the model by verifying that the estimated responses of crop yields to temperature and precipitation are consistent with previous studies [103] (Extended Data Fig. 2.2).

We find that the changes in sunlight from SS reduce both C4 (maize; p ¡ 0.01, N=2,501) and C3 (soy, rice, wheat; p ¡ 0.05, N=4,828) yields 48% and 28%, respectively, per unit SAOD (Fig. 3a Model 1). This implies that the global average scattering from Pinatubo (+0.15 SAOD) would reduce C4 yields 9.3% and C3 yields 4.8% (Fig. 3b), although some of this loss was likely offset by SS-induced cooling, making it difficult to observe directly. In contrast, process models [66] and empirical analyses of unmanaged-ecosystem biomass growth [36] tend to estimate a positive insolation effect, suggesting that either the diffuse fertilization effect is weaker for crops than ecosystems or scattering light alters the relative production of biomass and edible yield.

Our finding that SS scattering from Pinatubo negatively impacted yields is robust to removing temperature, precipitation, ENSO, and cloud controls (Fig. 3a Models 2-5), estimating the effect separately for each crop, accounting for the zenith angle of incoming sunlight, using two alternative datasets of SS SAOD, dropping observations from countries where the major eruptions occurred, and adding surface CO_2 as a control (Extended Data Table 2.2). We examine the impact of future, current, and past SS on current yields, finding that only contemporaneous SS exposure matters (Fig. 2.3d). We estimate the yield-insolation response flexibly, and fail to reject that the response is linear over the support of our data (Extended Data Fig. 2.3).

Extending the analysis back in time increases sample size but also measurement

error due to weaknesses in the historical observational system. The estimated insolation effect for both C3 and C4 crops becomes smaller and remains significant for C4 crops as we sequentially include data from the eruptions of El Chichón (1982) (Fig. 3a Model 6) and Agung (1963) (Extended Data Table 2.2 Col. 9). This pattern is consistent with both systematic "attenuation bias" from the mis-measurement of SAOD before the satellite era [128] and differences in the radiative properties of the SS generated by Pinatubo and El Chichón, discussed below.

Two results support that our analysis captures a sunlight-mediated effect. First, the response of C3 crops is less negative than that of C4 crops (p¡.01). C3 crops benefit from scattering more than C4 crops because the C3 photosynthetic rate saturates at lower light levels [35]. Second, per unit of SAOD, aerosols from El Chichón are both more forward scattering (Extended Data Tables 2.1-2.3) and less damaging to yields (Fig. 2.3a Models 7-8) than those of Pinatubo. This pattern is consistent with diffuse fertilization increasing edible yield. It also suggests that aerosol radiative properties might explain some heterogeneity in the estimated insolation effect across these eruptions. This heterogeneity substantially affects reconstructed yield losses from SS scattering (Fig 2.3c). We are, however, unable to determine whether differences in SS radiative properties or measurement errors (inducing attenuation bias) across eruptions are responsible for differences in their estimated insolation effects (Supplementary Information III.6).

To calculate the total effect of SS on yields for a future SRM scenario, we apply our empirical results (Fig. 2.3a Model 1) to output from an earth system model and compare future yields under two scenarios: (1) climate change under Representative Concentration Pathway 4.5, a modest mitigation pathway, and (2) the same, but with sulfur dioxide injection to balance all additional anthropogenic forcing after 2020 [70].

Over cropped areas in this simulation (2050-2069), the SRM treatment (avg. 0.084 SAOD) decreases maize growing season average temperatures 0.88 C, reduces precipitation 0.26 mm/month, and increases cloud fraction by 0.0081 relative to the control (Extended Data Fig. 2.4). In turn, average maize yields increase 6.3% due to this cooling (Fig. 2.4a), decrease 5.3% due to SRM-induced dimming (Fig. 2.4b), and change <0.2% due to altered precipitation and clouds (Fig. 2.4c-d). We sum these partial effects, repeating the analysis for soy, rice, and wheat (Extended Data Fig. 2.5). We find that SRM treatment, relative to the control, has no statistically discernible effect on yields once optical effects are accounted for (p¿0.1 for all crops; Fig. 2.4e, Extended Data Fig. 2.6). Failing to account for the insolation effect, as was done in the only prior global estimate [82], substantially overestimates the benefits of SRM to agriculture.

Our analysis finds that volcanogenic SS have statistically significant and economically substantial insolation-mediated costs that are roughly equal in magnitude to their benefits from cooling. This suggests that anthropogenic SS used in SRM may not be able to substantially lessen the risks that climate change poses to global agricultural yields and food security (Extended Data Fig. 2.7).

Our finding that SS from El Chichón were more forward scattering and less damaging than SS from Pinatubo indicates that optimizing the radiative properties of particles used in SRM might mitigate insolation-mediated damages. Although, we cannot rule out that this difference was due instead to poor observation of SS from El Chichón.

Farmer-level adaptations, such as switching to varieties more resistant to dimming, could theoretically mitigate SRM's insolation-mediated damage. However, given that farmer-level adaptations to extreme heat have been modest [14], it is not clear that adaptation to dimming will be easier.

Our quasi-experimental results are consistent with the sunlight-mediated impact of tropospheric aerosols [37] and emissions of their precursors [13] on Indian wheat and rice yields, further supporting that we capture a sunlight-mediated response. Still, it is possible that other factors, such as increased UV exposure from stratospheric ozone destruction, could explain part of the estimated effect. Notably, changes in tropospheric ozone concentrations due to Pinatubo are thought to be negative [110], which would increase yields, suggesting our results might underestimate the SS insolation effect.

Acknowledgements We thank M. Anderson, M. Auffhammer, D. Baldocchi, K. Caldeira, C. Field, A. Goldstein, D. Keith, P. Huybers, R. Kopp, D. Lobell, K. Ricke, J. Sallee, and seminar participants at Berkeley, Chicago, Columbia, Cornell, Harvard, Johns Hopkins, and Stanford universities, the Massachusetts Institute of Technology and the Allied Social Science Association Annual Meeting for useful comments. We thank I. Bolliger for his contributions to the project and all the members of the Global Policy Lab for their valuable feedback. We thank Larry Thomason for generously sharing SAOD data used in Figure 1a-c. This material is based upon work supported by the National Science Foundation Grant No. CNH-L 1715557 and the National Science Foundation Graduate Research Fellowship under Grant No. DGE 1752814.

Author contributions S.H. conceived the study; J.P., S.H., J.B., M.B. and W.S designed the study; J.P. collected and analyzed the data with contributions from J.B.; J.P., S.H., J.B., M.B. and W.S interpreted results; J.P. and S.H. wrote the paper.

Author information Reprints and permissions information is available at `www.nature.com/reprints`. The authors declare no competing interests. Correspondence and requests for materials should be addressed to proctor@berkeley.edu.

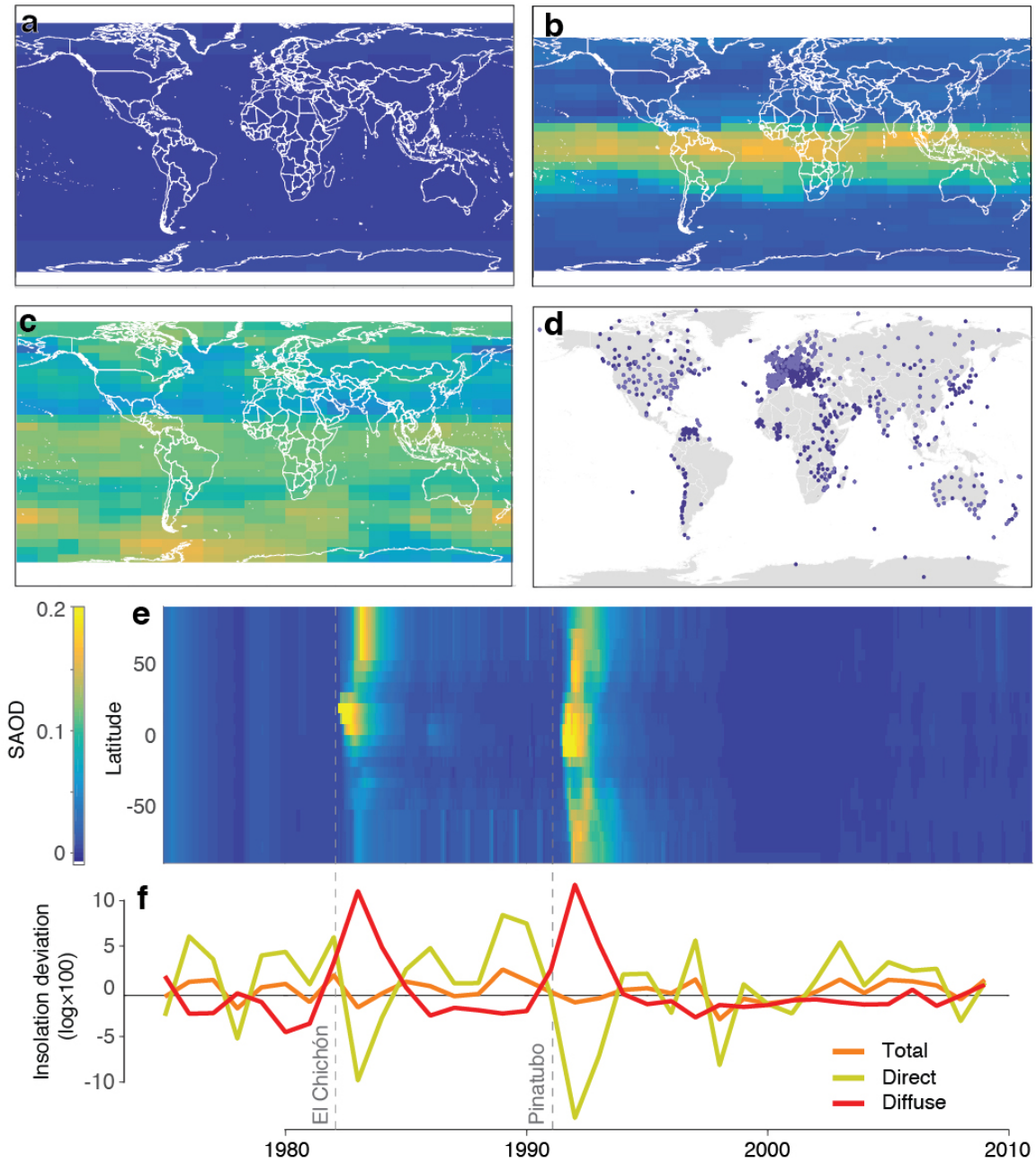


Figure 2.1: Large volcanic eruptions alter the global optical environment. Stratospheric aerosol optical depth (SAOD) (1000nm) **a,** before the Pinatubo eruption (March 1991); **b,** two months after the eruption (August 1991); and **c,** the next year after the aerosol cloud has spread (March 1992). **d,** Surface insolation observing stations used in our analysis of the effect of SAOD on insolation; light blue stations additionally measure diffuse light. **e,** SAOD (550nm) from 1975-2010 [101]. **f,** Annual average daily total (orange), direct (yellow) and diffuse (red) sunlight across all stations; measurements were demeaned by station-by-day-of-year before averaging to remove seasonal effects as well as differences in geography and observational protocols.

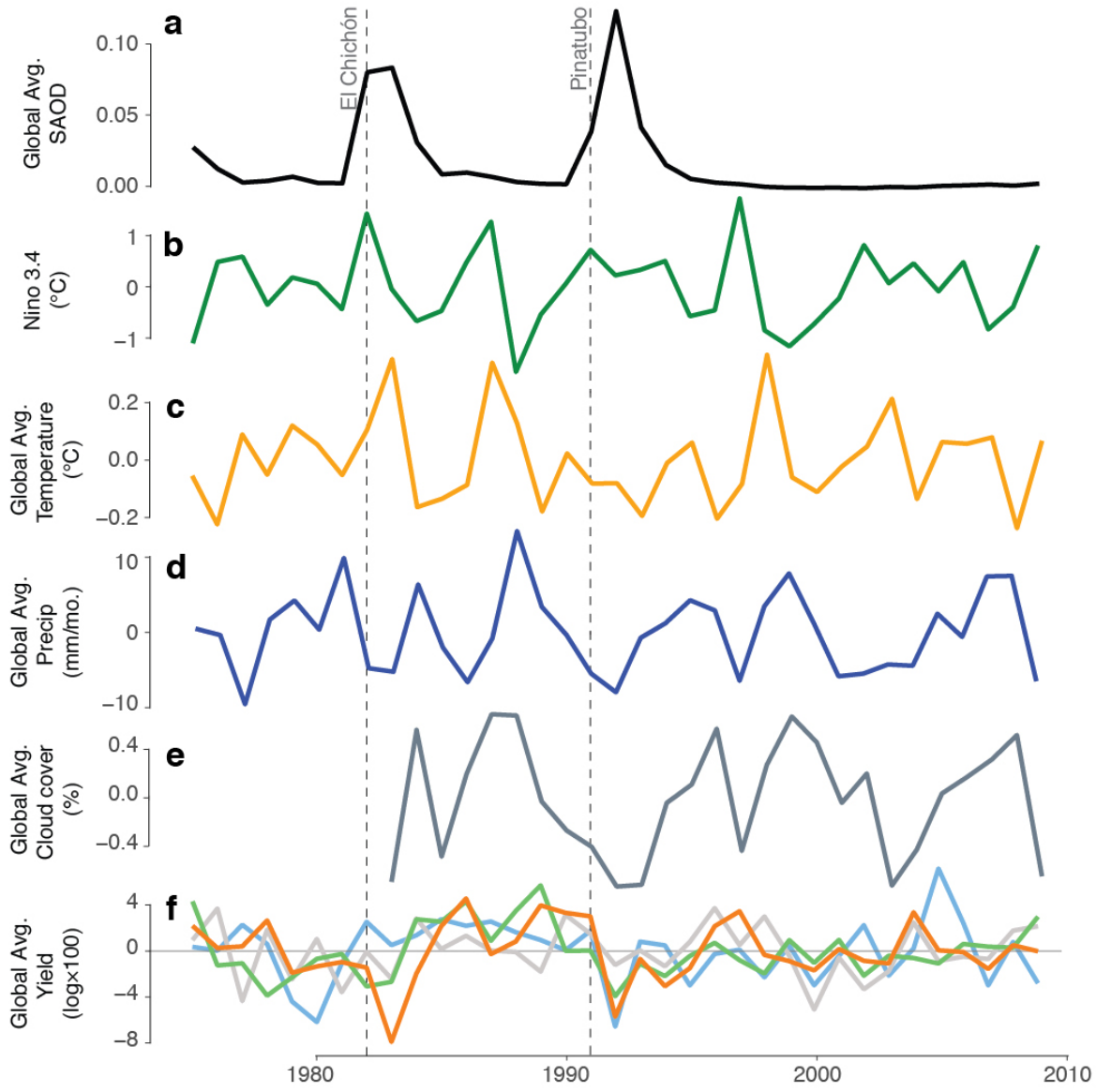


Figure 2.2: Global summary statistics of key model variables. a, Stratospheric aerosol optical depth increases for years following the eruptions of El Chichón (March-April 1982) and Pinatubo (June 1991) (dotted lines). **b,** the ENSO 3.4 index, **c,** surface air temperature, **d,** precipitation, and **e,** cloud fraction during the same period. **f,** Yields of maize (orange), wheat (gray), soy (blue), rice (green) decline following the eruptions. Climate and yield values are growing season averages, de-trended by country-specific quadratic time trends and averaged over countries in the sample. SAOD data is similarly processed but not de-trended.

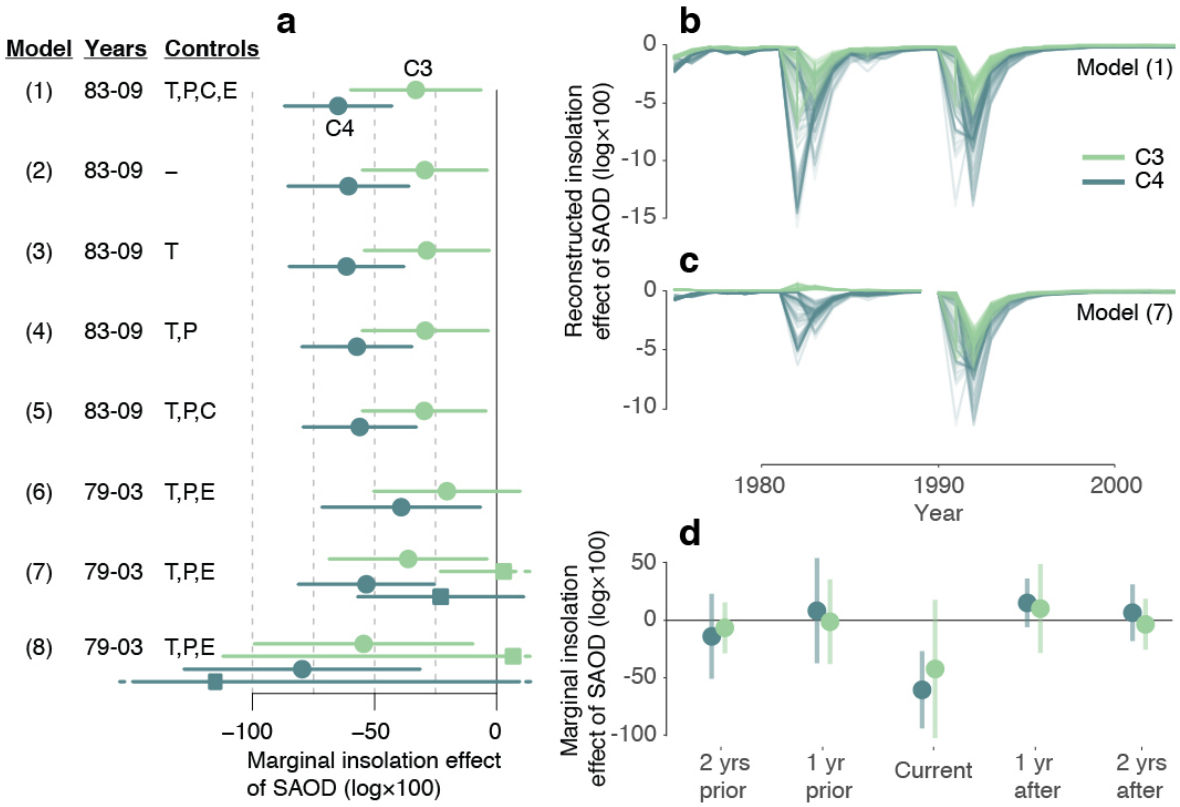


Figure 2.3: Empirical estimates of the insolation effect of SS on crop yield. **a,** The estimated effect of increasing SS optical depth by one unit on C4 (blue) and C3 (green) yields due to changes in sunlight (Model 1, Supplementary Information Eqn. 2.16, and Extended Data Table 2). Models 2-5 drop and then sequentially add temperature (T), precipitation (P), cloud (C) and ENSO (E) controls. Models 7-8 estimate effects separately for Pinatubo (year≥1990, circles) and Chichón (year<1990, squares); Model 8 uses a different SAOD dataset (SPARC). **b,** Reconstructions of the SS insolation effect using Model 1. Each line represents a single country over time. **c,** Same as b, but using Model 7. **d,** Simultaneously estimated insolation effects 2 years prior to and 2 years following the current growing season. See Supplementary Information III.2.3, III.2.2, III.4. In **a,d** whiskers represent 95% confidence intervals.

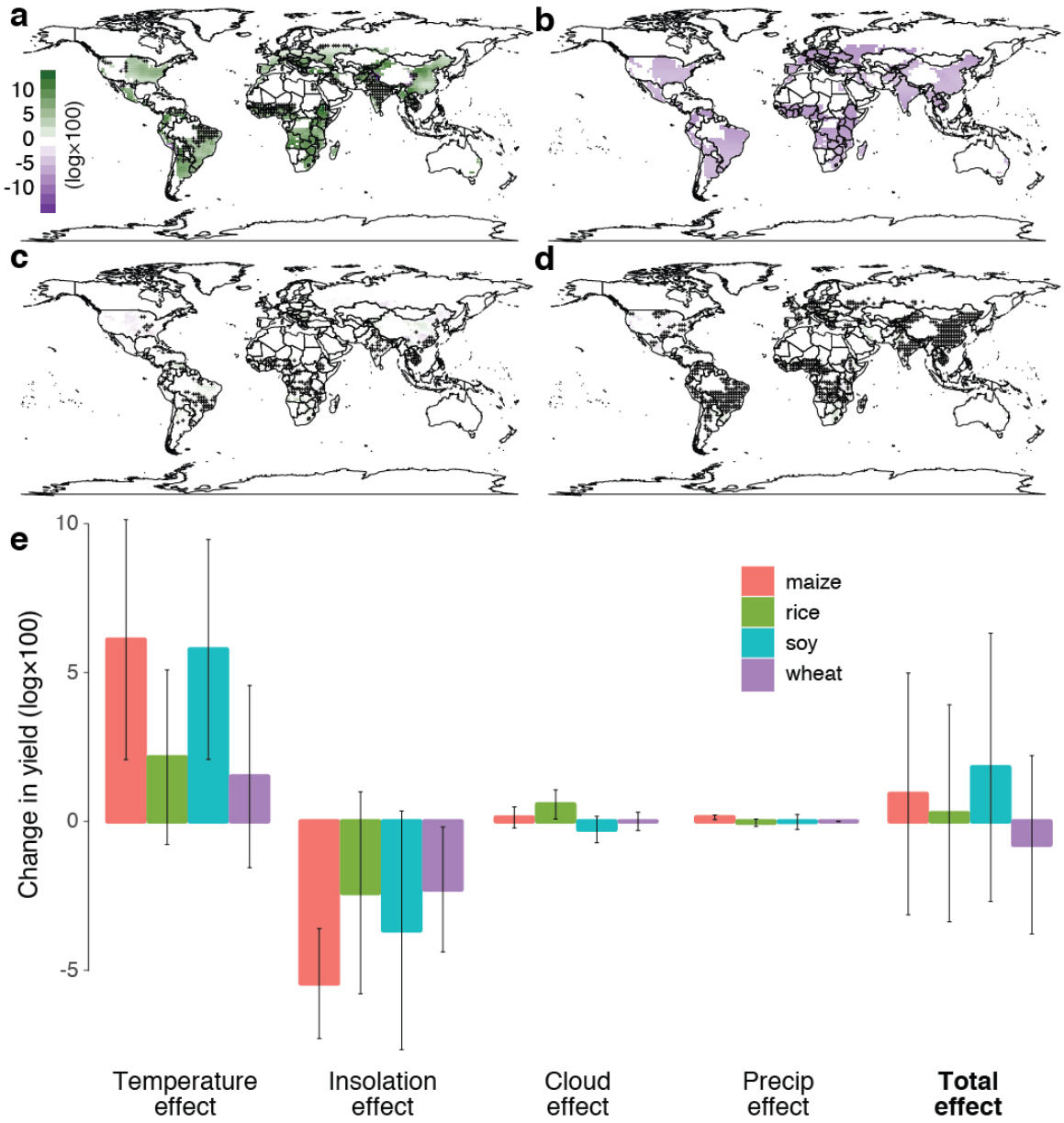


Figure 2.4: Partial and total effects of SRM on yields. The partial effects of SRM, relative to a climate-change-only scenario (RCP4.5), on expected maize yields from 2050-2069 due to changes in **a,** temperature, **b,** insolation, **c,** precipitation and **d,** cloud fraction. Statistically insignificant changes (p ¿ 0.05) are hatched. **e,** Global partial and total effects (cropped-fraction weighted average) for maize (red), soy (turquoise), rice (green) and wheat (purple). Error bars show 95% confidence intervals for the predicted effect.

Methods

To link the national annual yield data from the Food and Agricultural Organization of the United Nations to the climatological data we summarize all gridded temperature, precipitation, cloud and SAOD datasets to the annual-country level by averaging values over cropped area [88] to the growing season [100] using a similar methodology to refs. [12, 103].

Our analysis of the impact of SS on log insolation (N = 3,311,553 and 889,327 for total, and direct/diffuse insolation, respectively) models SAOD, cloud fraction[51], and ENSO (current and lagged) linearly (Supplementary Information Eqn. 2.2). We include station by day-of-year fixed effects. Our analysis of the impact of SS on atmospheric forward scattering shares the same specification (Supplementary Information Eqn. 2.5).

Our analysis of the impact of SS on log yields models the impact of SAOD linearly (non-linear estimates do not significantly differ from the linear estimate (Extended Data Fig. 2.3), the response of temperature [1], precipitation [127], and clouds [73] using restricted cubic splines, and allows the response of ENSO (current and lagged) to differ between teleconnected and non-teleconnected regions [48] (Supplementary Information Eqn. 2.16). We include country fixed effects and country-specific quadratic time trends. For all empirical insolation and yield analyses we calculate standard errors to account for serial correlation within countries across years and for spatial autocorrelation within years across countries [46].

To calculate the total effect of SRM relative to a climate change scenario we average results over three ensemble members from the Max Planck Institute Earth System Model [70]. Uncertainty in the total effect represents uncertainty in the estimated parameters of the empirical yield model (Supplementary Information IV.4). We do not consider carbon fertilization effects in calculation of the total effect because carbon dioxide levels are the same in the SRM and climate change only scenarios.

Data availability. All data used in this analysis is from free, publicly available sources and is available upon request from the corresponding author.

Code availability. Replication code is available upon request from the corresponding author.

Extended Data Table 2.1: Effect of SS on total, direct and diffuse insolation.

Radiation Type:	(1) Total	(2) Total	(3) Total	(4) Direct	(5) Direct	(6) Direct	(7) Diffuse	(8) Diffuse	(9) Diffuse
Years in Sample:	[83-09]	[79-09]	[79-09]	[83-09]	[79-09]	[79-09]	[83-09]	[79-09]	[79-09]
SAOD	−0.172***	−0.067		−1.580***	−1.395***		1.199***	1.197***	
	(0.062)	(0.058)		(0.320)	(0.295)		(0.122)	(0.125)	
SAOD x (yr≤89)			−0.024			−1.039			2.063***
[Chichón]			(0.079)			(0.760)			(0.102)
SAOD x (yr>89)			−0.100*			−1.406***			1.171***
[Pinatubo]			(0.054)			(0.301)			(0.115)
Cloud Fraction	−0.946***			−2.792***			0.499***		
	(0.041)			(0.179)			(0.085)		
Nino 3.4	0.002	0.002	0.002	0.018**	0.017	0.017	−0.001	0.0004	0.001
	(0.002)	(0.003)	(0.003)	(0.009)	(0.012)	(0.012)	(0.003)	(0.003)	(0.003)
Nino 3.4 (lagged)	0.004	−0.001	−0.001	0.006	−0.011	−0.011	−0.003***	0.001	0.0003
	(0.003)	(0.003)	(0.003)	(0.010)	(0.013)	(0.013)	(0.001)	(0.001)	(0.002)
Observations	3,311,553	4,371,586	4,371,586	889,327	1,000,776	1,000,776	889,327	1,000,776	1,000,776
Adjusted R^2	0.766	0.750	0.750	0.552	0.413	0.413	0.722	0.744	0.744

Coefficients on SAOD describe the effect of increasing SS optical depth by 1 unit on the log of total, direct or diffuse sunlight. Columns 1,4,7, show the preferred specification (Supplementary Material Eqn. 2.2). Columns 2,5,8 include data from 1979-2009 to capture the effect of both the Pinatubo and El Chichón eruptions. Columns 3,6,9 estimate the effect separately for El Chichón and Pinatubo (Supplementary Material II.1). We do not control for cloud fraction in columns 2,3,5,6,8 and 9 because the cloud data is only available beginning in 1983. All models account for station-by-day-of-year fixed effects. Standard errors of the mean, shown in parentheses, are clustered by country and by year to account for serial correlation over time within a country and for autocorrelation across space within a year. We calculate p values using a two-sided t-test; *p<0.1; **p<0.05; ***p<0.01.

Extended Data Table 2.2: Robustness of the insolation effect of SS on yields to changes in model specification, data sample, and data source.

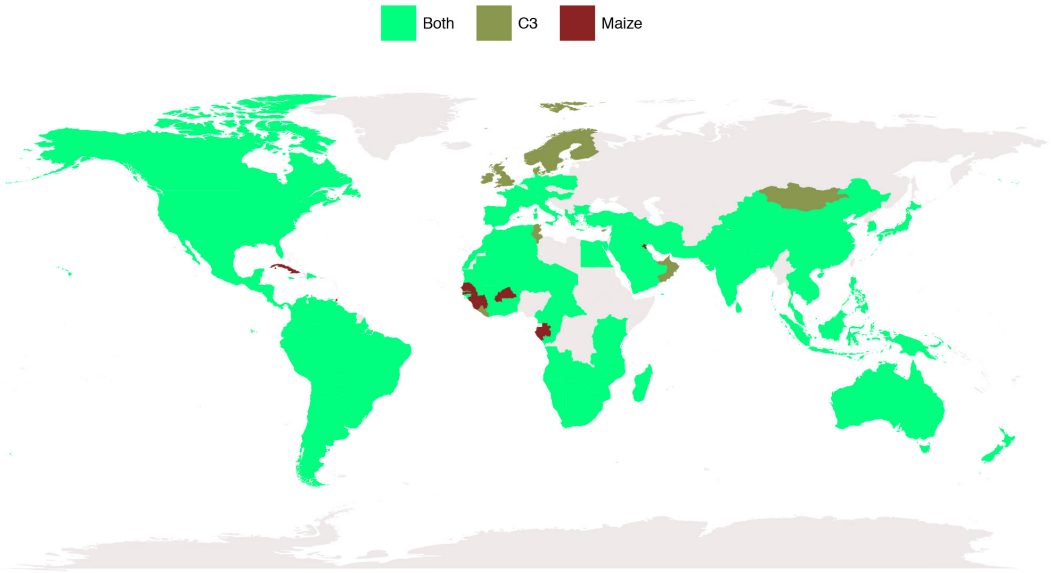
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Years in Sample	[83-09]	[83-09]	[83-09]	[83-09]	[83-09]	[83-09]	[83-03]	[83-05]	[61-09]	[83-09]	[83-09]	[79-03]	[79-03]
Climate Controls	None	T	TP	TPC	TPCE	TPCE	TPCE	TPCE	TPE	TPCE	TPCEO	TPE	TPE
SAOD Data						Cos(SZA)	SPARC	SPARC2		Drop Mex. & Phil.	Add CO_2		SPARC
Maize (C4)													
SAOD	-0.607***	-0.615***	-0.572***	-0.561***	-0.649***	-0.392***	-0.901***	-0.776***	-0.258**	-0.672***	-0.644***		
	(0.127)	(0.120)	(0.115)	(0.118)	(0.112)	(0.0656)	(0.252)	(0.162)	(0.125)	(0.116)	(0.119)		
SAOD x (yr≤89) *[Chichón]*												-0.229	-1.073
												(0.173)	(0.764)
SAOD x (yr>89) *[Pinatubo]*												-0.533***	-0.796***
												(0.142)	(0.247)
Observations	2,501	2,501	2,501	2,501	2,501	2,501	1,868	2,025	3,867	2,447	2501	2,322	2,211
R-squared	0.950	0.952	0.953	0.953	0.954	0.954	0.953	0.955	0.939	0.953	0.954	0.948	0.949
C3 - pooled													
SAOD	-0.294**	-0.286**	-0.293**	-0.297**	-0.331**	-0.183**	-0.559**	-0.439**	-0.0638	-0.349**	-0.325**		
	(0.131)	(0.131)	(0.132)	(0.129)	(0.136)	(0.0741)	(0.232)	(0.192)	(0.135)	(0.140)	(0.144)		
SAOD x (yr≤89) *[Chichón]*												0.0283	0.0669
												(0.132)	(0.606)
SAOD x (yr>89) *[Pinatubo]*												-0.362**	-0.545**
												(0.165)	(0.228)
Observations	4,828	4,828	4,828	4,828	4,828	4,828	3,618	3,916	7,431	4,694	4828	4,480	4,297
R-squared	0.940	0.941	0.941	0.942	0.942	0.942	0.946	0.946	0.928	0.941	0.942	0.942	0.943
Soy (C3)													
SAOD	-0.313	-0.327	-0.335	-0.356	-0.482*	-0.319*	-0.848**	-0.860***	-0.152	-0.541*	-0.483*		
	(0.287)	(0.282)	(0.276)	(0.288)	(0.270)	(0.160)	(0.381)	(0.268)	(0.280)	(0.270)	(0.271)		
SAOD x (yr≤89) *[Chichón]*												0.227	1.888
												(0.236)	(1.290)
SAOD x (yr>89) *[Pinatubo]*												-0.630**	-0.843**
												(0.233)	(0.324)
Observations	1,256	1,256	1,256	1,256	1,256	1,256	937	1,026	1,897	1,202	1,256	1,169	1,118
R-squared	0.883	0.888	0.889	0.890	0.890	0.890	0.905	0.903	0.868	0.891	0.890	0.894	0.894
Rice (C3)													
SAOD	-0.395*	-0.407*	-0.424**	-0.412*	-0.301	-0.158	-0.298	-0.203	-0.191	-0.321	-0.283		
	(0.196)	(0.198)	(0.201)	(0.202)	(0.217)	(0.118)	(0.372)	(0.267)	(0.143)	(0.228)	(0.225)		
SAOD x (yr≤89) *[Chichón]*												-0.149	-1.125
												(0.244)	(1.059)
SAOD x (yr>89) *[Pinatubo]*												-0.225	-0.353
												(0.251)	(0.368)
Observations	1,562	1,562	1,562	1,562	1,562	1,562	1,179	1,278	2,474	1,509	1,562	1,448	1,396
R-squared	0.935	0.935	0.935	0.935	0.936	0.935	0.941	0.941	0.907	0.935	0.936	0.932	0.933
Wheat (C3)													
SAOD	-0.201	-0.164	-0.161	-0.164	-0.257**	-0.126**	-0.594**	-0.352*	0.103	-0.253**	-.256**		
	(0.127)	(0.122)	(0.122)	(0.118)	(0.121)	(0.0600)	(0.249)	(0.200)	(0.114)	(0.123)	(0.114)		
SAOD x (yr≤89) *[Chichón]*												0.0672	0.0118
												(0.156)	(0.564)
SAOD x (yr>89) *[Pinatubo]*												-0.295*	-0.529**
												(0.158)	(0.232)
Observations	2,010	2,010	2,010	2,010	2,010	2,010	1,502	1,612	3,060	1,983	2,010	1,863	1,783
R-squared	0.939	0.940	0.940	0.941	0.941	0.941	0.944	0.944	0.934	0.940	0.941	0.942	0.943

The table above shows the insolation effect of SS for maize, C3 crops pooled, and soy, rice and wheat yields individually across a range of robustness checks (Supplementary Material III.4). The C3 response is estimated assuming that crops that share the C3 photosynthetic pathway (soy, rice, wheat) have a common insolation effect (Supplementary Material Eqn. 2.18). Columns 1-5 drop all climate controls and then add temperature (T), precipitation (P), cloud fraction (C) and ENSO (E) controls back in one at a time; column 5 is our preferred specification (Supplementary Material Eqns. 2.16); column 6 accounts for the angle at which incoming light passes through the SS layer by diving SAOD by the cosine of the solar zenith angle (SZA); columns 7 and 8 use two alternative SS datasets, SPARC and SPARC2 (Supplementary Material I.4); column 9 includes data from 1961 - 2009 to span the eruption of Agung; column 10 drops Mexico and the Philippines, where the El Chichón and Pinatubo eruptions occured, from the analysis; column 11 adds surface CO_2 concentration as a control. column 12 estimates the effects for El Chichón and Pinatubo separately; and column 13 does the same using the SPARC dataset. All models account for country fixed effects and country-specific quadratic time trends. Standard errors of the mean, shown in parentheses, are clustered by country and by year to account for serial correlation over time within a country and for autocorrelation across space within a year. We calculate p values using a two-sided t-test; *** p<0.01, ** p<0.05, * p<0.1.

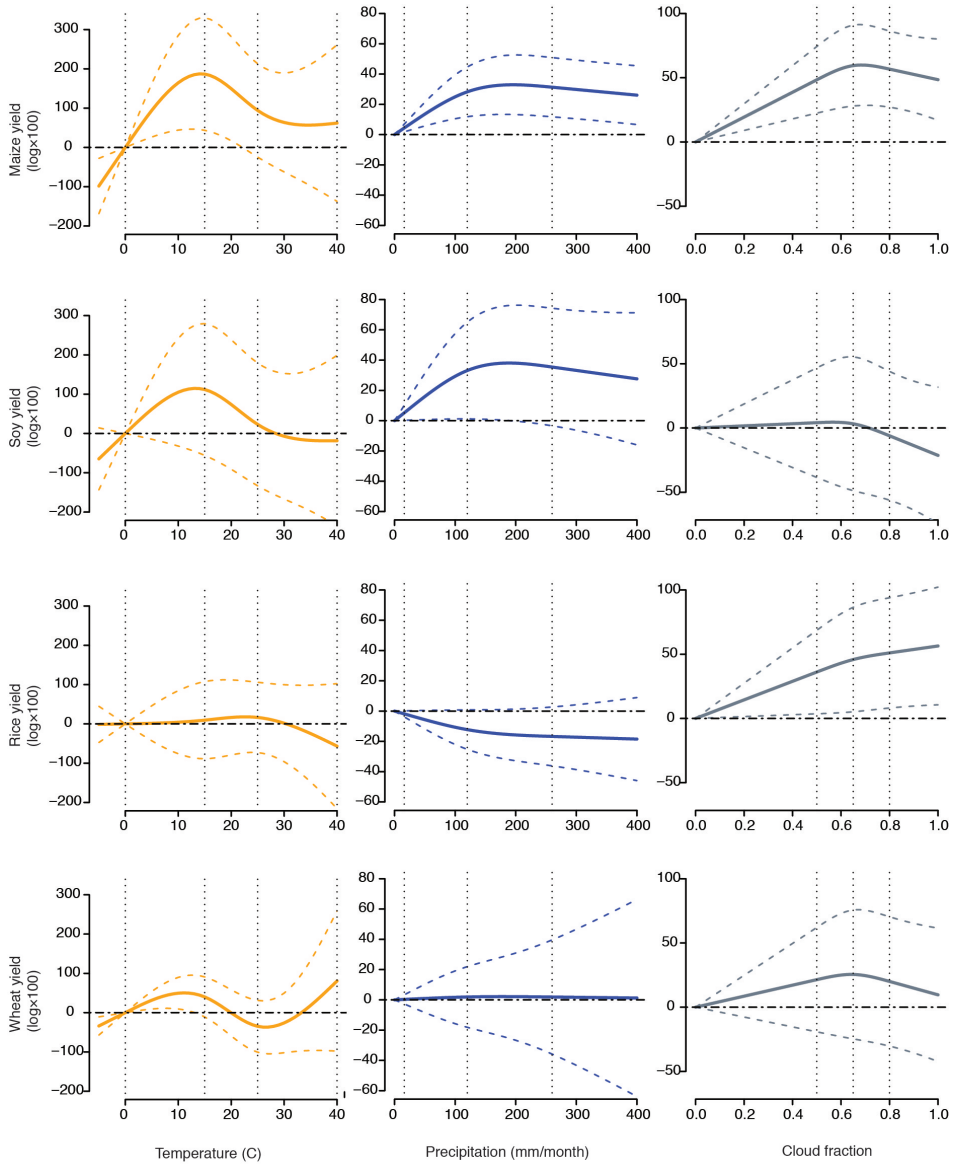
Extended Data Table 2.3: Effect of SS on atmospheric forward scattering

	(1)	(2)	(3)
Dep. Var. = Pr(photon reaches the surface — photon hits a particle)			
Year	[83-09]	[79-09]	[79-09]
SAOD	0.233***	0.243***	
	(0.030)	(0.030)	
SAOD x (yr≤89)			0.345***
[Chichón]			(0.031)
SAOD x (yr>89)			0.240***
[Pinatubo]			(0.029)
Cloud Fraction	−0.047***		
	(0.018)		
Nino 3.4	0.001	0.001	0.001
	(0.001)	(0.001)	(0.001)
Nino 3.4 (lagged)	−0.0004	−0.001	−0.001
	(0.001)	(0.001)	(0.001)
Observations	886,287	997,142	997,142
Adjusted R^2	0.228	0.227	0.227

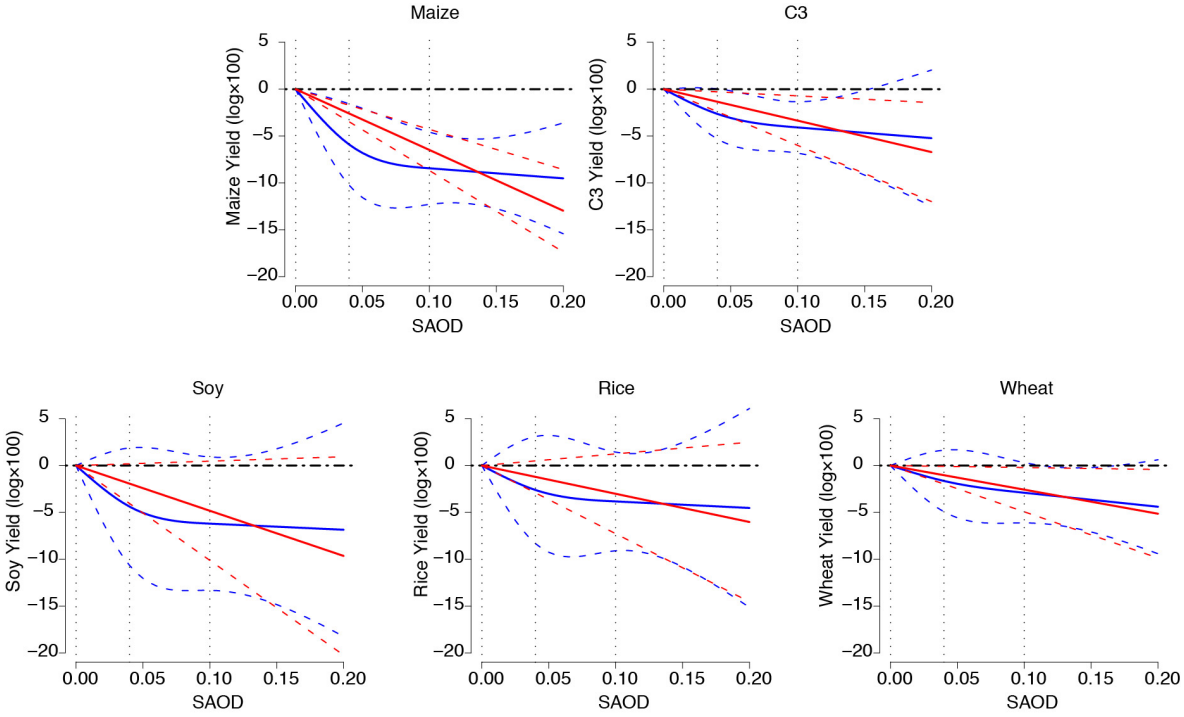
The dependent variable is the probability that a photon of light makes it to the surface, conditional on hitting a particle (w in Supplementary Material Eqn. 2.3). Coefficients on SAOD represent the effect of increasing SAOD by 1 unit on w for the entire atmospheric column. Column 1 is our preferred specification (Supplementary Material Eqn 2.5). Column 2 drops cloud controls and includes both the Pinatubo and El Chichón eruptions. Column 3 estimates the effects for El Chichón and Pinatubo separately. All models account for station-by-day-of-year fixed effects. Standard errors of the mean, shown in parentheses, are clustered by country and by year to account for serial correlation over time within a country and for autocorrelation across space within a year. We calculate p values using a two-sided t-test; *p<0.1; **p<0.05; ***p<0.01.



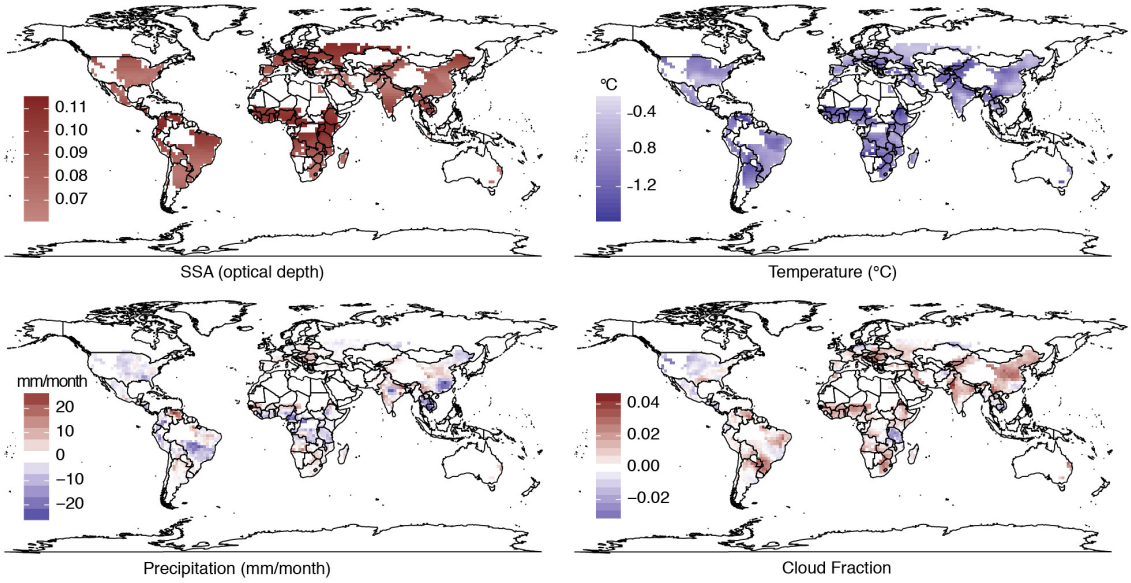
Extended Data Figure 2.1: Countries included in estimation of the insolation-mediated effect of stratospheric aerosol optical depth on crop yield. Countries in light green are included in estimation of the insolation-mediated effect of SS on yields for both C3 (soy, rice, wheat) and C4 (maize) crops. Countries in dark green are included only in estimation of the insolation effect for C3 crops, and countries in red are included only in estimation of the insolation effect for maize. Countries in grey are not included in the analysis due to missing data.



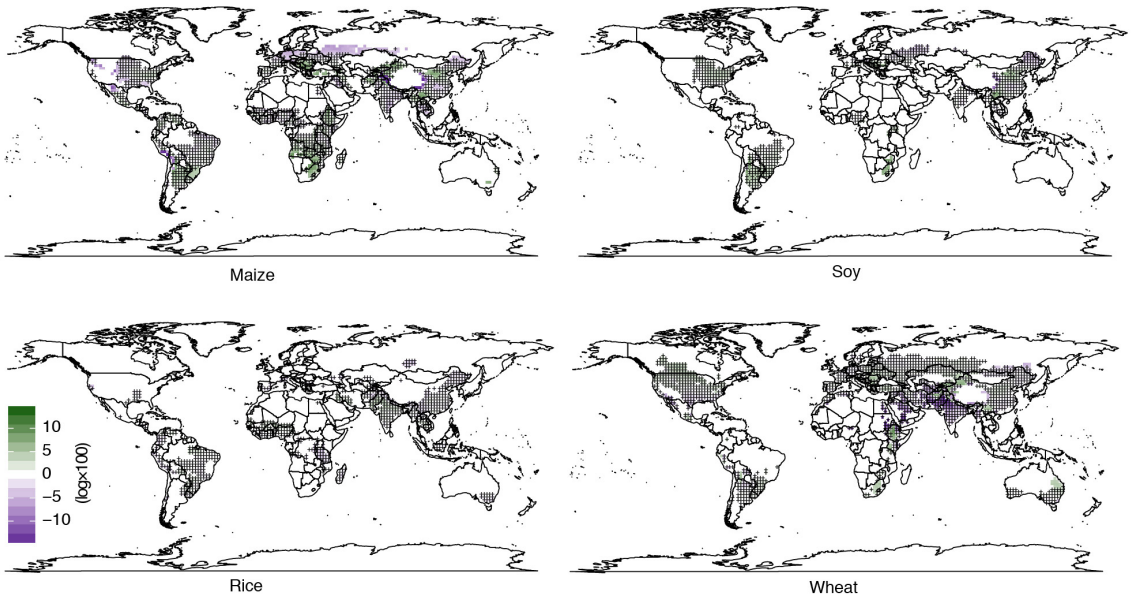
Extended Data Figure 2.2: Estimated response of yields to changes in growing season average temperature (orange), precipitation (blue), and cloud fraction (grey). Temperature, precipitation and cloud fraction axes show growing season means. The y-axes show partial effects on yield relative to a value of zero for each climatological variable ($f_T(T_{it})$, $f_P(P_{it})$, and $f_C(C_{it})$ in Supplementary Information Eqn. 2.16). Vertical dotted lines show the placement of the knots for the restricted cubic splines specification. Dashed lines show the 95% confidence intervals. N = 2,501, 1,256, 1,562 and 2,010 for maize, soy, rice and wheat, respectively.



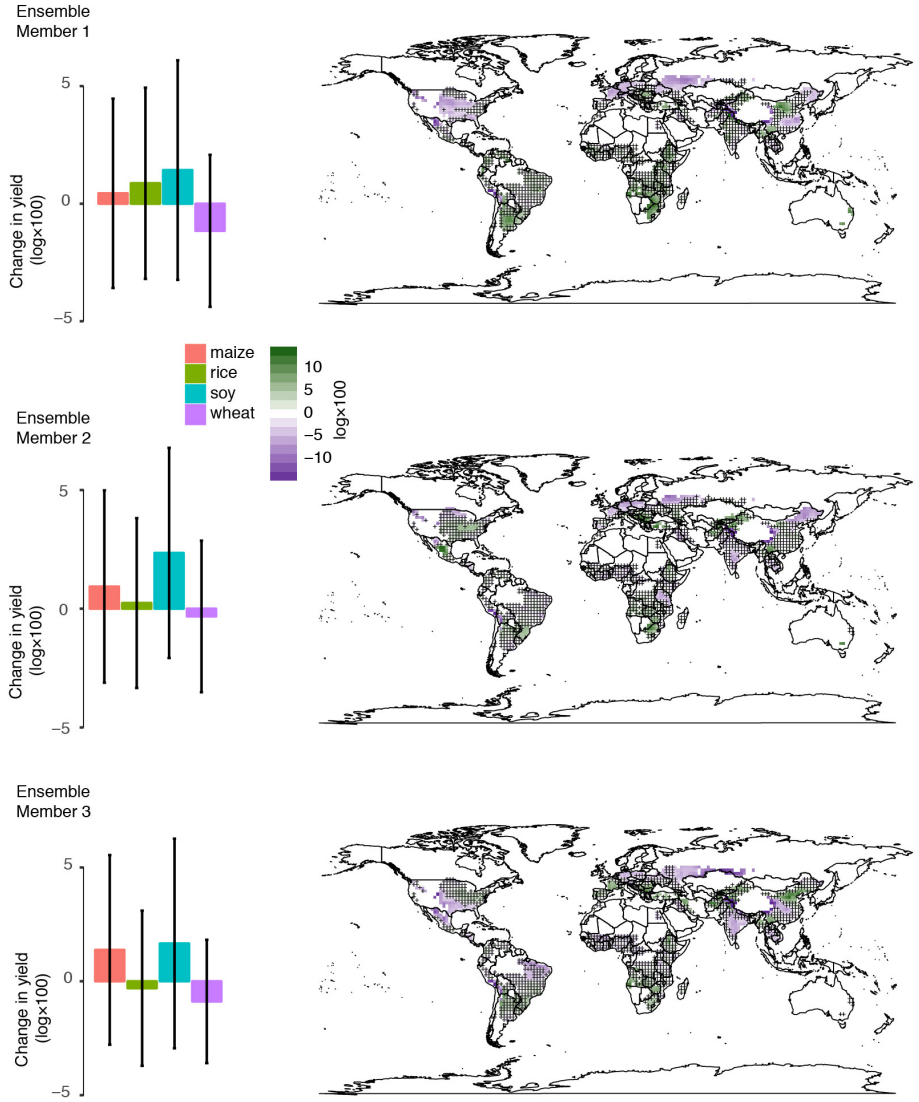
Extended Data Figure 2.3: Flexible (blue) and linear (red) estimation of the insolation-mediated impact of SS on crop yields. The SAOD axes show growing season means. Each point on a curve gives the optical effect of SAOD, relative to a value of zero (the slope of the red lines is β in Supplementary Information Eqn. 2.16). Vertical dotted lines show the placement of the knots for the restricted cubic splines specification. Dashed lines show the 95% confidence intervals.



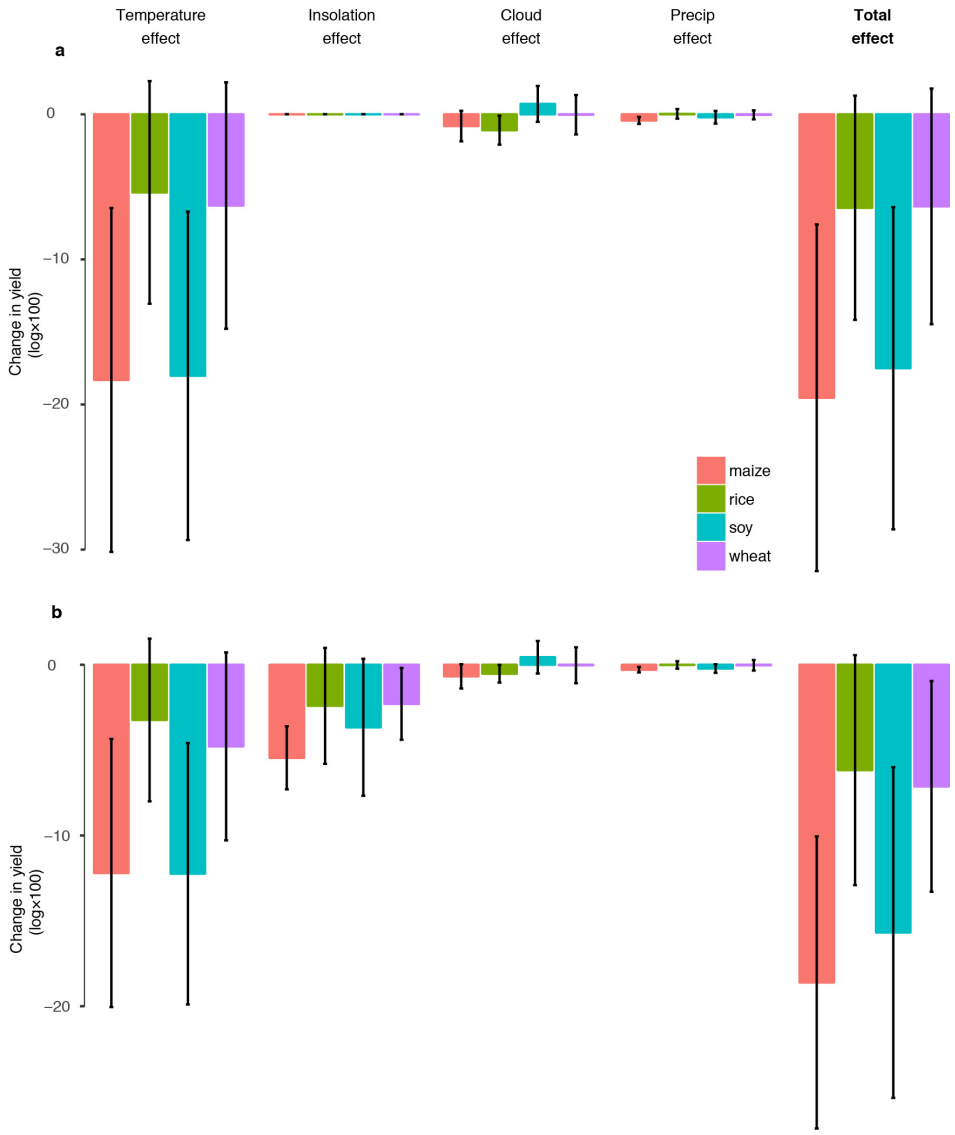
Extended Data Figure 2.4: Impact of SRM on climatological determinants of yield. SRM-induced changes in maize growing season average SAOD, temperature, precipitation, and cloud fraction, relative to the climate change only scenario.



Extended Data Figure 2.5: Total effect of SRM on maize, soy, rice, and wheat yields. Effects are relative to the climate change only scenario. Statistically insignificant effects (p ¿ .05) are hatched. We calculate p values using a two-sided t-test comparing the estimated effect of SRM to a null hypotheses of zero effect. When calculating the distribution of the estimated SRM effect, we consider only statistical uncertainty. This uncertainty is shown in Extended Data Table 2 and Extended Data Fig. 2 and the calculations are described in Supplementary Information IV.4.



Extended Data Figure 2.6: The finding that SRM mitigates little of the damages of climate change is consistent across three ensemble runs. Bar graphs show the total effect of SRM on global yields (cropped-fraction weighted average), relative to the climate change control, for each of the three ESM runs. Results are similar across ensemble member runs. Maps on the right show the total effect of SS on maize yields for each of the ensemble runs. Error bars in the bar graphs show 95% confidence intervals for estimated mean effects. Statistically insignificant effects (p ¿ .05) are hatched in the maps. We calculate P values using a two-sided t-test comparing the estimated effects to a null hypotheses of zero effect. Within each ensemble member, we calculate the distributions of the estimated effects considering only statistical uncertainty. This uncertainty is shown in Extended Data Table 2 and Extended Data Fig. 2 and the calculations are described in Supplementary Information IV.4.



Extended Data Figure 2.7: Effects of climate change and SRM relative to a historical scenario. a, Identical to Figure 4e, but comparing a climate change scenario (RCP 4.5) to a historical scenario (Supplementary Information IV.3). **b,** Identical to Figure 4e but comparing a climate change with SRM scenario to a historical scenario. Note that these calculations consider only climatological and sunlight-mediated impacts; changes in yields due to carbon fertilization, or other factors that may differ between scenarios, are not included. Error bars show 95% confidence intervals around the estimated mean effect.

Supplementary Information

Our analysis has three goals: to estimate the impact of SSAs on the quantity and quality of insolation, to estimate the impact of these changes in insolation on yields, and to estimate the total effect of changes in SSAs on yields in an SRM scenario. We discuss the estimation of these three responses, as well as the data used in the analysis, in the four supplementary information sections below.

I Data

This section describes and discusses the datasets used in this analysis.

I.1 Insolation

The World Radiation Data Centre (WRDC) provides daily insolation data from 1953 to 2013 [129]. Beginning with the entirety of their records (N = 8,394,737), we removed all data that either had missing total radiation values, was flagged as poor quality by the WRDC, or had higher reported diffuse radiation than total radiation (N = 7,334,570). In our main sample from 1983-2009 there are 3,311,553 observations from 859 stations of total insolation; of these, 889,327 observations from 324 stations split the total insolation measurements into direct and diffuse light.

I.2 Yield

The Food and Agricultural Organization of the United Nations Statistics Division provides annual country-level yields of corn, soybeans, rice and wheat [27]. These four crops make up roughly half of global caloric consumption. We included only countries that have no missing observations from 1983-2009 to balance the panel; notably, this drops countries created or dismantled by the break up of the Soviet Union in 1991. The cleaned panel contains data for maize, soy, rice and wheat in 100, 50, 61 and 80 countries, respectively (Extended Data Fig. 2.1).

I.3 Growing season and cropped fraction

We obtained global crop planting and harvesting dates from Sacks et. al. 2010 [100]. We obtained cropped area fraction (all agricultural land and crop-specific) data from Ramankutty et. al. 2008 [88] and Monfreda et. al. 2008 [67].

I.4 Stratospheric aerosol optical depth

We measure SSA concentrations using optical depth. Optical depth is a measure of opacity, defined as the natural logarithm of the ratio of incoming to transmitted direct light at a given wavelength:

$$\text{optical depth} = \ln\left(\frac{\text{incident direct light}}{\text{transmitted direct light}}\right) \tag{2.1}$$

We use three SAOD datasets in this analysis that share many similarities, but differ in the observations included in the analysis, the way they were processed, and the wavelength at which optical depth is measured.

For our main analysis we use the Goddard Institute for Space Studies (GISS) dataset, which is an updated version of the data presented in ref [101].[1] This dataset

[1]Data available at http://data.giss.nasa.gov/modelforce/strataer/; accessed 8/8/16.

provides global SAOD (550nm) from 1850-2012 at 5 degree latitudinal resolution. We use the GISS dataset as our preferred specification because it measures SAOD at 550 nm, which is at the center of the photosynthetically active range (400-700nm) and is the same wavelength as the SAOD applied in the G3 SRM scenario that we use to calculate the total effect of SSAs on yields. For the time period of our study (1979-2009), the GISS dataset incorporates observations from satellites (Stratospheric Aerosol and Gas Experiment (SAGE) I, SAGE II, Stratospheric Aerosol Measurement II) as well as aircraft, ground-based observations and balloon measurements.

We corroborate our main findings by using two additional SAOD datasets that measure optical depth at 1000 nm. These datasets are interpolated using similar methods and from similar data as the GISS dataset, but measure SAOD at a different wavelength and have different spatial resolutions and temporal extents. The first, produced by the Stratospheric Processes and their Role in Climate (SPARC) report [113], spans both the El Chichón and Pinatubo eruptions (available 1979-2003) but has only latitudinal resolution (5 degree, the same as GISS). The second, produced using a similar methodology to the SPARC dataset and which we call SPARC2,[2] spans only Pinatubo (Oct 1984 - July 2005) but is interpolated to give both latitudinal and longitudinal spatial resolution (5 x 15 degrees).

All three SAOD datasets used in our analysis combine observations from many sources because of important gaps in the satellite record. The primary gaps in the SAGE record are between November 1981 and October 1984, when there were no satellite observations of stratospheric aerosols and during the Mount Pinatubo eruption (June 1991 to 1993) when the instruments on SAGE II were saturated by the unanticipated density of particulate matter in the stratosphere. Though these gaps in the satellite record were filled and re-calibrated using non-satellite observations, they present an unavoidable source of measurement error. Measurement of SAOD before the satellite era is very coarse. Further, because there were no low-latitude observing satellites during the Chichón eruption, measurement following the Chichón eruption was much less precise and accurate than after the Pinatubo eruption. The poor observation of the Chichón eruption is why we focus our analysis on the Pinatubo eruption.

I.5 Climate variables

Our analysis uses historical temperature, precipitation, and cloud data as controls in our empirical estimation of the insolation effect. The Berkeley Earth Daily Land gridded dataset provides daily average surface temperature values from 1880 to present at 1-degree resolution [1] . The University of Delaware Air Temperature and Precipitation dataset provides monthly total precipitation over land (mm/month) from 1900 to 2010 at 0.5-degree resolution [127]. The International Satellite Cloud Climatology Project (ISCCP) gridded cloud product (D1) dataset contains 202 remotely sensed cloud parameters every 3 hours on a global 280 km equal area grid from July 1983 to December 2009 [51] . We use two versions of this dataset in our statistical estimation. Our analysis of SSA impacts on yields uses a *monthly* cloud fraction dataset that was corrected for

[2]Acquired through personal communication with the author of the SPARC Assessment of Stratospheric Aerosol Properties (Chapter 4) report, Larry Thomason (07/10/2015).

artefacts in the data [73]. Our analysis of SSA impacts on insolation uses the *3-hourly* cloud fraction records directly, which we average to daily temporal resolution (excluding nighttime observations). For our analysis of yields we use the corrected monthly observations because the outcome is annual, so having daily cloud resolution does not provide benefits relative to monthly averages.

II Insolation response

A central contribution of this study is to document the impact of SSAs on insolation at a global scale and to test how these changes in insolation impact yields. In this section we describe how we empirically estimate the effect of SSAs on insolation.

II.1 Estimating the impact of SSAs on insolation

To identify the impact of SSAs on insolation (Extended Data Table 2.1) at station i on day τ we estimate:

$$ln(I^X_{i\tau}) = \psi S_{im} + \eta C_{i\tau} + \theta_1 E_y + \theta_2 E_{y-1} + \phi_{id} + \epsilon_{i\tau} \tag{2.2}$$

Where I^X is either direct (I^D), diffuse (I^F) or total (I^T) insolation, m is month of sample, y is year of sample, and d is day of year. S_{im} are monthly SAOD measurements, $C_{i\tau}$ are daily observations of cloud fraction, and E_y and E_{y-1} are current and lagged values of the average monthly Niño 3.4 index (E) from May-December [48]. The parameter of interest is ψ, which describes the effect of SAOD on insolation. The low temporal resolution of our SAOD measurements is a source of measurement error which could attenuate our estimated coefficients towards zero. We include station-by-day-of-year fixed effects (ϕ_{id}) to non-parametrically control for all time-invariant differences between stations, such as average tropospheric aerosol optical depth, and for location-specific seasonal patterns within each location, such as dust storms in the summer or smog from heating in the winter. Any remaining variation in surface insolation that is affected by tropospheric aerosols is assumed to be orthogonal to overhead SAOD variations (measured predominately by the SAGE II satellites) and thus captured by the disturbance term ($\epsilon_{i\tau}$), which is averaged out of the conditional expectation function. Such idiosyncratic changes in insolation might be caused by, for example, the one-time closure (or opening) of a coal power plant near an insolation station. As long as these types of human-caused events are not systematically correlated with SAOD throughout our sample, both over time and coherently around the world, then our estimated effect of SAOD on insolation, ψ, will be unbiased. In this model ψ is identified by comparing SAOD and insolation measurements across years within a station-day-of-year (e.g. insolation and SAOD on March 1, *1990* in Berkeley, CA, USA is compared to insolation and SAOD on March 1 *1991* in Berkeley, CA, USA), after accounting for local cloud cover and the global ENSO state. We compute standard errors allowing for arbitrary patterns of serial correlation over time between all stations in the same country and for arbitrary patterns of autocorrelation across all station observations within the same year [46]. In cases where the estimated variance-covariance matrix is non-positive semi-definite, we apply the adjustment from Cameron, Gelbach & Miller [22].

To compare the response of insolation to SSAs from the eruptions of El Chichón and Pinatubo (Extended Data Table 2.1 Cols. 3,6,9) we make two changes to the model.

We replace $\psi S_{i\tau}$ with $\psi_{Chichon} S_{i\tau} \mathbb{1}(\tau \in [1979, 1989]) + \psi_{Pinatubo} S_{i\tau} \mathbb{1}(t \in [1990, 2009])$, which allows the effect of SSAs on insolation, ψ, to be different for the El Chichón and Pinatubo eruptions. $\mathbb{1}(.)$ is an indicator function which equals 1 if the day-of-sample is within the specified years and 0 otherwise. We drop cloud controls because the cloud data are available beginning in 1983, which is after the El Chichón eruption.

This analysis captures the average effect of SAOD on all-sky surface insolation across 859 stations (Fig. 2.1d). Using a global sample of stations enables us to average over local heterogeneity in the effect of SAOD due to interactions of SAOD scattering with further scattering and absorption by clouds or tropospheric aerosols. Though the higher density of insolation stations in northern latitudes and near urban centers means that our estimates are most representative of these areas, this global analysis is substantially more representative than previous studies, which have been limited to a few individual stations [42, 25].

This large sample of insolation stations also allows us to tease out the relatively small SAOD signal from substantial background noise due to clouds and tropospheric aerosols. This enables us to calculate the effect during all days–the crop-relevant effect–rather than limiting our analysis to clear-sky and minimally cloudy days, as previous analyses of individual stations have done [42, 25].

II.2 Estimating the impact of SSAs on atmospheric forward scattering

We define w as the proportion of light that makes it to the surface after being scattered:

$$\begin{aligned} w &= Pr(\text{photon reaches the surface}|\text{photon interacts with any particle}) \\ &= \frac{I^F}{I^F + I^R} \\ &= \frac{I^F}{I^0 - I^D} \end{aligned} \tag{2.3}$$

Where the third equality uses the identity that top of atmosphere insolation (I^0) is the sum of direct, diffuse, and reflected or absorbed light (I^R):

$$I^0 = I^D + I^F + I^R$$

Thus, w can be calculated using data on I^F (measured), I^D (measured), and I^0 (calculated as a function of latitude, longitude, and date). Rearranging Eqn. 2.3 shows that w measures how much diffuse light is gained from the loss of one unit of direct light:

$$I^F_{i\tau} = w \underbrace{\left[I^0_{i\tau} - I^D_{i\tau}\right]}_{blocked\ insolation} \tag{2.4}$$

In our sample, w is on average 0.32 meaning that about one third of insolation that does not make it to the surface as direct radiation ends up making it to the surface as diffuse radiation.

To estimate the effect of SSAs on w (Extended Data Table 2.3) we estimate:

$$w_{i\tau} = \alpha S_{im} + \eta C_{i\tau} + \theta_1 E_y + \theta_2 E_{y-1} + \phi_{id} + \epsilon_{i\tau} \tag{2.5}$$

The coefficient α describes the impact of a one unit increase in SAOD on the forward scattering probability (w) of the entire atmospheric column. The other parameters in Eqn. 2.5 are defined in the same way as in Eqn. 2.2. We model heterogeneous effects of SSAs on w across volcanic eruptions (Extended Data Table 2.3 Col. 3) in the same way that we model heterogeneous effects of SSAs on insolation across eruptions, which is described in the previous section.

III Agricultural response

In this section we present a framework for empirically estimating the insolation effect and the total effect of SSAs on yields.

III.1 Empirical framework

We consider a situation where yield (Y) is a function of inputs: insolation (I), temperature (T), precipitation (P), cloud fraction (C), and other unobservable variables (U) such as labor supply, or surface ozone concentration. Indexing observations by country i and year t we have:

$$Y_{it} = Y(T_{it}, P_{it}, C_{it}, I_{it}, U_{it}) \tag{2.6}$$

A challenge in this context is that these inputs are themselves functions of SSAs (S), ENSO (E), and other factors (Z):

$$\mathrm{T}_{it} = T(E_t, S_{it}, Z_{it}) \tag{2.7}$$
$$\mathrm{P}_{it} = P(E_t, S_{it}, Z_{it}) \tag{2.8}$$
$$\mathrm{C}_{it} = C(E_t, S_{it}, Z_{it}) \tag{2.9}$$
$$\mathrm{I}_{it} = I(E_t, S_{it}, Z_{it}) \tag{2.10}$$
$$\mathrm{U}_{it} = U(E_t, S_{it}, Z_{it}) \tag{2.11}$$

We discuss ENSO explicitly because the eruptions of both El Chichón and Pinatubo coincided to some degree with strong El Niño events, which could confound our estimation of the SSA insolation effect, and thus total effect, if not appropriately controlled for. ENSO is a global phenomenon so E_t itself does not vary by location; though its impacts on T, P, C, I and U, and in turn on Y, may vary across space (Supplementary Information III.2, III.3).

To consider the impacts of SSAs on yield, we differentiate Y (Eqn. 2.6) with respect to S, which decomposes the effect of SSAs into an insolation term, a temperature term, a precipitation term, a cloud term, and a term capturing any residual effects of SSAs:

$$\frac{dY}{dS} = \underbrace{\frac{\partial Y}{\partial I}\frac{\partial I}{\partial S}}_{\text{insolation effect}} + \underbrace{\frac{\partial Y}{\partial T}\frac{\partial T}{\partial S}}_{\text{temperature effect}} + \underbrace{\frac{\partial Y}{\partial P}\frac{\partial P}{\partial S}}_{\text{precipitation effect}} + \underbrace{\frac{\partial Y}{\partial C}\frac{\partial C}{\partial S}}_{\text{cloud effect}} + \underbrace{\frac{\partial Y}{\partial U}\frac{\partial U}{\partial S}}_{\text{residual effect}} \tag{2.12}$$

Ideally, to empirically recover the insolation effect ($\frac{\partial Y}{\partial I}$) and the total effect ($\frac{dY}{dS}$) we would observe the changes in these climatic, insolation, and residual variables due to SSAs (i.e. $\frac{\partial I}{\partial S}$, $\frac{\partial T}{\partial S}$, $\frac{\partial P}{\partial S}$, $\frac{\partial C}{\partial S}$, $\frac{\partial U}{\partial S}$), but we cannot measure these directly. What we are able to observe instead is how some of these variables and SSAs change over time (i.e.

$\frac{\partial S}{\partial t}$, $\frac{\partial T}{\partial t}$, $\frac{\partial P}{\partial t}$, $\frac{\partial C}{\partial t}$). To see how these observations are useful, we differentiate our original yield equation (Eqn. 2.6) with respect to time to get:

$$\frac{dY}{dt} = \frac{\partial Y}{\partial T}\frac{dT}{dt} + \frac{\partial Y}{\partial P}\frac{dP}{dt} + \frac{\partial Y}{\partial C}\frac{dC}{dt} + \frac{\partial Y}{\partial I}\frac{dI}{dt} + \frac{\partial Y}{\partial U}\frac{dU}{dt} \tag{2.13}$$

Changes in yield over time are the combined effect of changes over time due to each input. Differentiating the climate, insolation, and residual equations (Eqns. 2.7-2.11) with respect to time and then substituting into Eqn. 2.13 gives the change in yield over time due to changes over time in the factors that determine each input.

$$\begin{aligned}\frac{dY}{dt} = &\frac{\partial Y}{\partial T}\underbrace{\left[\frac{\partial T}{\partial E}\frac{dE}{dt} + \frac{\partial T}{\partial S}\frac{dS}{dt} + \frac{\partial T}{\partial Z}\frac{dZ}{dt}\right]}_{\frac{dT}{dt}} + \frac{\partial Y}{\partial P}\underbrace{\left[\frac{\partial P}{\partial E}\frac{dE}{dt} + \frac{\partial P}{\partial S}\frac{dS}{dt} + \frac{\partial P}{\partial Z}\frac{dZ}{dt}\right]}_{\frac{dP}{dt}} \\ &+ \frac{\partial Y}{\partial C}\underbrace{\left[\frac{\partial C}{\partial E}\frac{dE}{dt} + \frac{\partial C}{\partial S}\frac{dS}{dt} + \frac{\partial C}{\partial Z}\frac{dZ}{dt}\right]}_{\frac{dC}{dt}} + \frac{\partial Y}{\partial I}\underbrace{\left[\frac{\partial I}{\partial E}\frac{dE}{dt} + \frac{\partial I}{\partial S}\frac{dS}{dt} + \frac{\partial I}{\partial Z}\frac{dZ}{dt}\right]}_{\frac{dI}{dt}} \\ &+ \frac{\partial Y}{\partial U}\underbrace{\left[\frac{\partial U}{\partial E}\frac{dE}{dt} + \frac{\partial U}{\partial S}\frac{dS}{dt} + \frac{\partial U}{\partial Z}\frac{dZ}{dt}\right]}_{\frac{dU}{dt}}\end{aligned} \tag{2.14}$$

Because we observe and are able to control for them directly, we condense the $\frac{dT}{dt}$, $\frac{dP}{dt}$, and $\frac{dC}{dt}$ terms:

$$\begin{aligned}\frac{dY}{dt} = \frac{\partial Y}{\partial T}\frac{dT}{dt} + \frac{\partial Y}{\partial P}\frac{dP}{dt} + \frac{\partial Y}{\partial C}\frac{dC}{dt} + \frac{\partial Y}{\partial I}\left[\frac{\partial I}{\partial E}\frac{dE}{dt} + \frac{\partial I}{\partial S}\frac{dS}{dt} + \frac{\partial I}{\partial Z}\frac{dZ}{dt}\right] \\ + \frac{\partial Y}{\partial U}\left[\frac{\partial U}{\partial E}\frac{dE}{dt} + \frac{\partial U}{\partial S}\frac{dS}{dt} + \frac{\partial U}{\partial Z}\frac{dZ}{dt}\right]\end{aligned}$$

Re-arranging further shows the terms that will be represented directly in our empirical model:

$$\begin{aligned}\frac{dY}{dt} = &\underbrace{\frac{\partial Y}{\partial T}}_{\frac{\partial}{\partial T}f_T(.)}\mathbf{\frac{dT}{dt}} + \underbrace{\frac{\partial Y}{\partial P}}_{\frac{\partial}{\partial P}f_P(.)}\mathbf{\frac{dP}{dt}} + \underbrace{\frac{\partial Y}{\partial C}}_{\frac{\partial}{\partial C}f_C(.)}\mathbf{\frac{dC}{dt}} \\ &+ \underbrace{\left[\frac{\partial Y}{\partial I}\frac{\partial I}{\partial E} + \frac{\partial Y}{\partial U}\frac{\partial U}{\partial E}\right]}_{\frac{\partial}{\partial E}g(E)}\mathbf{\frac{dE}{dt}} + \underbrace{\left[\frac{\partial Y}{\partial I}\frac{\partial I}{\partial S} + \frac{\partial Y}{\partial U}\frac{\partial U}{\partial S}\right]}_{\beta}\mathbf{\frac{dS}{dt}} + \underbrace{\left[\frac{\partial Y}{\partial I}\frac{\partial I}{\partial Z} + \frac{\partial Y}{\partial U}\frac{\partial U}{\partial Z}\right]\frac{dZ}{dt}}_{\epsilon}\end{aligned} \tag{2.15}$$

Observed quantities are in bold. Quantities estimated in our empirical model, discussed in the following section, are shown in underbraces.

III.2 Empirical model

We estimate the structure of Eqn. 2.15 to recover the insolation-mediated impact of SSAs on yields ($\frac{\partial Y}{\partial I}\frac{\partial I}{\partial S}$) (Fig. 2.3 Model 1, Extended Data Table 2.1) in country i year t by fitting the model:

$$ln(Y_{it}) = f_T(T_{it}) + f_P(P_{it}) + f_C(C_{it}) + g(E_t, E_{t-1}, i) + \beta S_{it} + \lambda_i + \phi_{i1}t + \phi_{i2}t^2 + \epsilon_{it}. \tag{2.16}$$

β is the insolation-mediated effect of SSAs on yields and is the coefficient of interest in our study. It captures both the impact of decreasing total radiation and increasing the diffuse fraction, as well as any residual impacts of SSAs not mediated through T,P, or C (i.e. $\frac{\partial Y}{\partial U}\frac{\partial U}{\partial S}$), if they exist. Y_{it} are yields of either maize, wheat, soy, or rice. $f_T(.)$, $f_P(.)$, and $f_C(.)$ are restricted cubic splines which allow for estimation of a flexible response [41]. λ_i are country fixed effects which account for all time invariant differences between countries, such as topography or cultural history. $\phi_{i1}t + \phi_{i2}t^2$ are country-specific quadratic time-trends which control for gradual country-specific developments, such as smooth changes in technology or income. [3][46].

$g(.)$, defined as

$$g = (\theta_1 E_t + \theta_2 E_{t-1})\mathbb{1}(i \in \text{tele}) + (\theta_3 E_t + \theta_4 E_{t-1})\mathbb{1}(i \notin \text{tele}), \tag{2.17}$$

controls for current and lagged values of the average monthly Niño 3.4 index (E) from May-December, and allows the responses of teleconnected regions to differ from those of non-teleconnected regions [48]. $\mathbb{1}$ is an indicator function, so $\mathbb{1}(i \in \text{tele}) = 1$ if country i is teleconected and 0 otherwise. ϵ_{it} captures factors that impact yield, and vary over time within a country but are not captured by the quadratic country-specific time trend or the climate controls, and are uncorrelated with SAOD, such as the price of oil. In all yield regressions we compute our standard errors allowing for arbitrary patterns of serial correlation within countries over time and for arbitrary patterns of autocorrelation within years across countries [46]. In cases where the estimated variance-covariance matrix is non-positive semi-definite, we apply the adjustment from Cameron, Gelbach & Miller [22]. We model the impact of SAOD on yields linearly because flexible estimation of the response suggests that a linear fit is appropriate over the support of SAOD values observed during the eruptions of El Chichón and Pinatubo as well as those used in the SRM projections (Extended Data Fig. 2.3).

[3]Controlling for spatial-unit-specific polynomial time trends is a commonly used approach to account for gradual changes in yields, such as those driven by adoption of new technology. Though aggregation up to the national level will likely smooth out changes that are discontinuous at the local level, some abrupt or discontinuous changes in national crop production technologies will not be accounted for by these polynomial trends. Directly measuring and accounting for each change in agricultural technology is not possible. Thus, instead, to address these unobserved variables we leverage the notion that SSA concentrations were as good as randomly assigned by the unanticipated eruptions of El Chichón and Pinatubo to obtain an unbiased estimate of the impact of SSAs on yields, despite the potential presence of these omitted variables. With this strategy, in order for a discontinuous change in technology to bias our estimated effects of SSAs, it would need to be the case that technology all around the world changed abruptly, discontinuously, and coherently at the time of these eruptions. Given the gradual nature of agricultural technological diffusion, we do not find this plausible. See Supplementary Information III.3 for an in-depth discussion of identification.

III.2.1 Pooled C3 response Soy, rice and wheat share the same metabolic pathway for carbon fixation and thus their yields may respond similarly to changes in sunlight induced by SSAs. To estimate a pooled insolation effect for these C3 crops (Fig. 2.3, Extended Data Table 2.2) we fit the model:

$$ln(Y_{itj}) = \beta_{C3}S_{itj} + f_{Tj}(T_{itj}) + f_{Pj}(P_{itj}) + f_{Cj}(C_{itj}) + g_j(E_t, E_{t-1}, i) + \lambda_{ij} + \phi_{i1j}t + \phi_{i2j}t^2 + \epsilon_{itj} \quad (2.18)$$

where j indexes crop (soy, rice, wheat). This is the same model as for individual crops except that here, changes in SSAs, and in turn insolation, are assumed to have a common effect (β_{C3}) across all C3 crops. Temperature effects, precipitation effects, cloud effects, ENSO effects and time trends are still allowed to be crop-specific. $g_j(.)$ shares the same form as Eqn. 2.17 for each crop j. Thus, all parameters other than β_{C3} are estimated 'as if' individual versions of Eqn. 2.16 were estimated for each crop.

III.2.2 Heterogeneous effects across eruptions To compare the response of crops to the eruptions of El Chichón and Pinatubo (Fig 2.3a Models 7-8) we make two changes to Eqn. 2.16. We replace βS_{it} with $\beta_{Chichon}S_{it}\mathbb{1}(t \in [1979, 1989]) + \beta_{Pinatubo}S_{it}\mathbb{1}(t \in [1990, 2003])$, which allows the estimated insolation effect to differ across the two eruptions. We also drop cloud controls because the cloud data are only available beginning in 1983 (after the El Chichón eruption). We similarly alter Eqn. 2.18 to allow the C3 insolation effect to vary across the two eruptions.

III.2.3 Leads and lags To estimate the effect of SSAs in years prior to and after the current growing season on the current growing season's yield (Fig 2.3d) we use the same form as Eqn. 2.16 and include two years of lags and leads for SAOD, temperature, precipitation, cloud fraction and ENSO into the model.[4] Coefficients for leads, lags and the contemporaneous effects are estimated simultaneously.

III.3 Identification strategy

Our experimental design and model specification address three main challenges to empirically estimating the insolation effect of SSAs on yields. First, the countries most heavily treated with SSAs (e.g. those in the tropics) may have different average yields than those that were less heavily treated. Second, there may be time-varying factors that are correlated with SSA exposure and affect yield. Third, in addition to affecting yields by changing sunlight, SSAs may affect yields by altering a variety of other yield inputs including temperature, precipitation, and cloud cover (as expressed in Eqn. 2.14).

Our fixed-effects panel estimation strategy estimates the impact of SSAs on yields by comparing countries to themselves over time with varying exposure to SSAs. This is achieved by including an indicator variable for each country in the model, which accounts for all time-invariant differences between countries. We identify the insolation effect using the remaining variation in yields and SSAs within individual countries over

[4]Because ENSO is already lagged, the creation of two leads and lags for each variable creates duplicated ENSO controls in the model; we drop these before estimation.

time. We also account for time-trending differences between countries using country-specific quadratic time trends.

We address potential confounding from unobserved time-varying determinants of yield by leveraging the quasi-random variation in SSAs induced by unanticipated volcanic eruptions and subsequent stratospheric aerosol transport. These natural experiments cause SSA concentrations over countries and over time to vary in a way that is arguably as good as randomly assigned. This quasi-random assignment of SSAs, like random assignment in a randomized control trail, balances unobserved covariates within countries before and after volcanic treatment.

A further complication specific to this study is that the SSA distribution engendered by the eruptions of El Chichón and Pinatubo happened, by chance, to coincide temporally and spatially with strong El Niño events, which are global climatic phenomena that affect yields through similar pathways to SSAs (Eqn. 2.14). We address this by both conditioning on the pathways through which ENSO is likely to affect yields (temperature, precipitation, and cloud cover) and by conditioning directly on the current and lagged ENSO index to account for any remaining pathways through which ENSO variation might impact yields (i.e. $\frac{\partial U}{\partial E}\frac{dE}{dt}$ or $\frac{\partial I}{\partial E}\frac{dE}{dt}$).

In addition to accounting for these time-varying factors, we address potential confounding due to the impacts of SSAs on other determinants of yield by directly controlling for observable determinants of yield that SSAs affect, such as temperature, precipitation, and clouds (Section III.1).

III.3.1 Identifying assumptions This identification strategy rests on two assumptions. First, we assume that the aerosol treatment is independent of other variables that impact yield, conditional on the model's controls. Put another way, we assume that a country before and after an eruption is a good "control" for the same country during overhead SSA "treatment," after conditioning on climate controls, ENSO, fixed effects, and time trends. [5] Second, we assume that SSAs affect yields through changing only insolation, temperature, precipitation, and clouds:

$$\frac{\partial Y}{\partial U}\frac{\partial U}{\partial S} = 0 \tag{2.20}$$

If both of these assumptions hold, then our estimate of the insolation effect will not be confounded. If Eqn. 2.19 (in footnote) holds but 2.20 does not, then instead of capturing only the insolation effect, β will additionally capture any residual effects of SSAs on yields that are localized to a country and that are not mediated though changes in temperature, precipitation, or clouds (i.e. $\frac{\partial Y}{\partial I}\frac{\partial I}{\partial S} + \frac{\partial Y}{\partial U}\frac{\partial U}{\partial S}$). Such potential channels include increases in ultraviolet radiation (UV), [122] and decreases in tropospheric ozone [130]. It is unlikely that changes in tropospheric ozone are the mechanism that explain the estimated negative effect of SAOD on yields. Because tropospheric ozone reduces crop

[5] More formally, we assume that:

$$S_{it} \perp\!\!\!\perp \epsilon_{it} | f_T(T_{it}), f_P(P_{it}), f_C(C_{it}), g(E_t, E_{t-1}, i), \lambda_i, \phi_{i1}t, \phi_{i2}t^2 \tag{2.19}$$

where notation in the above equation is borrowed from our empirical model (Eqn. 2.16).

yields [65], concentrations of tropospheric ozone would have to be positively correlated with SAOD to explain the negative impacts of SSAs on yields. To the contrary, there is evidence that Pinatubo did [110] and SRM would [130] decrease tropospheric and surface ozone concentrations. Though it is unknown whether this had a measurable effect on yields during the Pinatubo eruption, if anything these changes in ozone would cause us to under-estimate insolation-mediated SAOD damages, rather than over-estimate them. It is possible that changes in UV radiation may explain part of our estimated "insolation effect" because UV radiation damages crop yields and there is some evidence that the decrease in ozone following the eruption let in more UV light than the sulfate aerosols directly blocked [122]. Future research should further explore the mechanisms responsible for these empirically estimated effects of SSAs on yields.

III.4 Robustness of the SSA insolation effect on yields

Here we examine the sensitivity of our estimation of the insolation effect of SSAs on yields to different model specifications, data samples, and data sources (Fig. 2.3, Extended Data Table 2.2).

To test the sensitivity of our results to the temperature, precipitation, cloud and ENSO controls we run the analysis without these climate controls (Extended Data Table 2.2 Column 1) and then add in temperature, precipitation, clouds and ENSO one at a time (Columns 2-5). Column 5 is our preferred specification. We see that the effect of SSAs is robust to removing these climate controls; all coefficients across all 5 specifications and all 5 crops and crop groups are negative, and tend to have similar magnitudes and precisions to our preferred specification.

To test the sensitivity of our results to different measures of SSA exposure, we divide SAOD by the cosine of the solar zenith angle (SZA) (Column 6). Motivation for this model stems from the insolation mechanism: if a ray of light passes through an aerosol layer at an angle, it interacts with more of the aerosol layer than if it passes through perpendicularly. Accounting for the angle of incident sunlight mechanically increases the treatment values (i.e. $SAOD < \frac{SAOD}{cos(SZA)}$) and thus decreases the size of the estimated coefficient. The predicted average treatment effect between the CSZA and preferred specifications, however, is essentially unchanged.[6] We avoid using the CSZA transformation as our preferred specification because it gives similar predictions and increases the conceptual and computational complexity of our model.

To test whether local, direct volcanic damages (e.g mudflows or avalanches of ash) are driving the global result, we drop the countries where these major eruptions occurred (Mexico and the Philippines) from our analysis; we find little difference in the results from our preferred specification (Column 10).

To test whether interannual fluctuations in carbon dioxide (CO_2) concentrations are driving the main result, we measure and control for surface CO_2 concentrations in our empirical model. We calculate the CO_2 concentration (ppm) for each year, country and crop by averaging monthly average CO_2 concentrations over the growing season. We

[6] In 1992, the average SAOD at 550nm across countries is 0.13 and the SAOD coefficient is -.649, which gives a predicted effect on yields of -8.2%. Similarly, the average $\frac{SAOD}{cos(SZA)}$ across countries is 0.21 and the coefficient for maize is -.392, which gives a predicted effect on yields of -8.3%.

use data from Mauna Loa [3] and Cape Grim [2] to measure CO_2 concentrations in the Northern and Southern Hemispheres, respectively. [7] We find that controlling for surface CO2 concentrations has little impact on the main results (Column 11).

We test the robustness of the results across different SAOD datasets, by re-estimating the model with the SPARC and SPARC2 datasets, described in Supplementary Information IV (Columns 7,8,12). Though the coefficients are larger–because for a given concentration of SSAs, SAOD measured at 1000nm is smaller than SAOD measured at 550nm due to the fact that SSA particles block light at 550nm more efficiently than they do at 1000nm–we calculate predicted effects using these datasets that are similar to predicted effects using the preferred GISS dataset. [8]

III.5 Interpretation of the insolation effect

Here we analyze what our estimation of the insolation effect ($\frac{\partial Y}{\partial I}\frac{\partial I}{\partial S} = \beta$) suggests about the strength of the diffuse fertilization effect for the production of crop yield. From our most general yield model in the previous section we have:

$$Y = Y(T, P, C, I, U)$$

Allowing the effects of direct and diffuse light to differ, and taking a first order Taylor approximation gives:

$$Y = \gamma_1 I^D + \gamma_2 I^F + f(T, P, C)$$

Differentiating with respect to S to get the marginal impacts of SSAs gives:

$$\frac{\partial Y}{\partial S} = \underbrace{\gamma_1 \frac{\partial I^D}{\partial S} + \gamma_2 \frac{\partial I^F}{\partial S}}_{\text{insolation effect}} + \frac{\partial}{\partial S} f(T, P, C) \tag{2.21}$$

Here, the insolation effect ($\frac{\partial Y}{\partial I}\frac{\partial I}{\partial S}$) is represented by the sum $\gamma_1 \frac{\partial I^D}{\partial S} + \gamma_2 \frac{\partial I^F}{\partial S}$. Equation 2.4 showed how w for the entire atmospheric column determines how much diffuse light is gained from the scattering of a unit of direct light. Differentiating Eqn. 2.4 with respect to S gives:

$$\frac{\partial I^F}{\partial S} = \frac{\partial w}{\partial S}\left[I^0 - I^D\right] + w\frac{\partial}{\partial S}\left[I^0 - I^D\right] \tag{2.22}$$

[7] We assume that CO_2 concentrations are homogeneous within each hemisphere since within-hemisphere variation in CO_2 is relatively small (¡3ppm) [61] and finer-scale measurements, such as those from remote sensing, are not available during our study period.

[8] In 1992, the average SPARC SAOD at 1020nm across countries is 0.076 and the SAOD coefficient is -.901, which gives a predicted effect on yields of -6.9%. Similarly, the average SPARC2 SAOD at 1020nm across countries is 0.099 and the coefficient for maize is -.776, which gives a predicted effect on yields of -7.8%. Tests for difference between these predicted effects and the predicted effect using the GISS dataset (-8.2%, see previous footnote) fail to reject the null hypothesis that the effects are the same (p¿0.1).

which describes how the optical behavior of the entire atmospheric column changes with respect to the concentration of SSAs. To limit the focus of this equation to the effects of scattering due to only SSAs, we define a new variable:

$$w_{ss} = Pr(\text{photon reaches the surface}|\text{photon hits } \textit{a SSA particle}) \tag{2.23}$$

and substitute it in for w in Eqn. 2.22 to get:

$$\frac{\partial I^F}{\partial S} = \frac{\partial w_{ss}}{\partial S}\left[I^0 - I^D\right] + w_{ss}\frac{\partial}{\partial S}\left[I^0 - I^D\right] \tag{2.24}$$

Note that the definition of w_{ss} is identical to that of w except that w_{ss} is specific to SSA particles and w describes the scattering properties of the entire atmospheric column. Because $\frac{\partial w_{ss}}{\partial S}$ and $\frac{\partial I^0}{\partial S}$ equal zero[9] we can simplify this to:

$$w_{ss} = -\frac{\frac{\partial I^F}{\partial S}}{\frac{\partial I^D}{\partial S}} \tag{2.25}$$

Thus, similar to w, w_{ss} is both the probability that a ray of light makes it to the surface conditional on hitting a particle *of SSA* and the amount of diffuse light that is gained from the loss of one unit of direct light *due to SSA scattering*. Now, solving for $\frac{\partial I^F}{\partial S}$ in Eqn. 2.25 and substituting into Eqn. 2.21 we get:

$$\frac{\partial Y}{\partial S} = \underbrace{(\gamma_1 - \gamma_2 w_{ss})\frac{\partial I^D}{\partial S}}_{\text{insolation effect}} + \frac{\partial}{\partial S} f(T, P, C) \tag{2.26}$$

This formulation of the yield model allows us analyze the relative contribution of direct and diffuse light to yield using our estimations of the insolation effect and the optical properties of SSA. Here, the insolation effect is $(\gamma_1 - \gamma_2 w_{ss})\frac{\partial I^D}{\partial S}$, where w_{ss} describes the optical properties of the SSAs and γ_1 and γ_2 describe how direct and diffuse light impact yields. Because scattering blocks the transmission of direct light, $\frac{\partial I^D}{\partial S}$ is negative. Thus, the insolation effect is positive if $\gamma_1 - \gamma_2 w_{ss} < 0$ and negative if $\gamma_1 - \gamma_2 w_{ss} > 0$. The diffuse light fertilization effect suggests that diffuse light is used more efficiently in the production of yield than direct light, which means that $\gamma_2 > \gamma_1$. However, a unit of diffuse light comes at the cost of $\frac{1}{w_{ss}}$ units of direct light. If the benefit of diffuse light (γ_2) after accounting for the loss of direct light (w_{ss}) exceeds the benefit of direct light (γ_1), then $\gamma_1 < \gamma_2 w_{ss}$ and scattering will increase yields. If w_{ss} is too small, or the relative benefit of diffuse to direct light is too small, then $\gamma_1 > \gamma_2 w$ and scattering decreases yields.

[9] Stating that $\frac{\partial w_{ss}}{\partial S} = 0$ rests on the assumption that adding more SSAs to the atmosphere does not impact the optical behavior of the SSAs already there (i.e. no multiple scattering). This assumption is likely to hold reasonably well over the range of optical depths observed during the Pinatubo and El Chichon eruptions as well as the SRM simulations [28].

Our finding that the insolation effect is negative suggests that $\gamma_1 > \gamma_2 w$. Calculation of w_{ss} using our insolation data, suggests that $w_{ss} \approx 0.85$.[10] This means that $\gamma_2 < 1.2\gamma_1$, meaning that diffuse light is less than 1.2 times more efficiently used for the production of yield than direct light. This finding is in contrast to previous studies, which have found that scattering light tends to substantially increase plant growth [66, 36], though there are exceptions [107]. Our results suggest that either increasing the diffuse fraction increases photosynthesis less for crops than for other ecosystems,[11] or that scattering light reduces the harvest index,[12] or that past estimates have confounded the impacts of diffuse light with the effects of cooling. Though this calculation gives a rough upper bound for the strength of the diffuse fertilization effect in the production of crop yield, further research should more rigorously estimate the magnitude of this effect as well as how it depends on environmental conditions.

III.6 Discussion of potential mechanisms causing the differences in the estimated insolation effect between eruptions

Here, we expand our discussion in the main text of whether the differences in the insolation effect between the El Chichón and Pinatubo eruptions should be interpreted as a product of either attenuation bias or of differences in SSA radiative properties. We add two points to those made in the main text. First, if attenuation bias were the only mechanism at work, we might expect the estimated impact of SSAs on insolation (e.g. Extended Data Tables 2.1,2.3) to be attenuated towards zero as well. Rather, we estimate larger marginal effects of SSAs on diffuse radiation and on forward scattering during the El Chichón eruption than the Pinatubo eruption, which supports the argument that differences in SSA radiative properties may have played a role. The distribution of insolation stations, however, is concentrated in areas, such as Europe, where non-satellite measurements of SSAs would also be most available and accurate. Thus, it is possible that the measurement error of SSA concentrations is smaller over the insolation stations than over the countries included in the yield regressions, which include many tropical and subtropical countries with limited insolation and SSA monitoring. If that is the case, then the observed differences in the insolation effect could still be due to attenuation bias.

Second, as mentioned in the paper, we find that including data from the El Chichón and Agung eruptions, consecutively, makes the estimated insolation effect progressively closer to zero, though still negative. If SSA radiative properties are driving this, we would expect the radiative properties of the SSAs from the El Chichón and Agung eruptions to be more similar to each other than to the radiative properties of the SSAs from the Pinatubo eruption. Though we are unable to directly estimate the insolation impacts of the SSAs erupted by Agung using data from insolation stations, other measurements suggest that the SSAs from Agung were roughly similar in size

[10] To estimate w_{ss}, we re-estimate Eqn. 2.2 using I^D and I^F, rather than $ln(I^D)$ and $ln(I^F)$ as outcome variables. We get $\frac{\partial I^F}{\partial S} = 1,909 \frac{Wh}{m^2 day}$ per unit SAOD (p¡.01) and $\frac{\partial I^D}{\partial S} = -2,238 \frac{Wh}{m^2 day}$ per unit SAOD (p¡.01), which gives a w_{ss} of 0.85 using Eqn. 2.23.

[11] Selective breeding for upright architectures and high photosynthetic capacity may be one reason for the relatively weaker diffuse fertilization effect in crops.

[12] The harvest index is the mass of edible yield divided by the total plant mass.

to those of Chichón [101], and particle size is a key determinant of SSA radiative characteristics. This is consistent with SSA radiative properties contributing to the differences in the insolation effect between eruptions, though it should be noted that measurements of the particle size of Agung's SSAs were likely imprecise. Future research will hopefully clarify the degree to which attenuation bias or differences in SSA radiative properties explain the differences in the insolation effect across eruptions.

IV Total effect estimation in SRM simulations

The last step of our analysis is to estimate the total effect of SSAs on yields in a SRM scenario. The SRM scenario we examine is the G3 experiment from the Geoengineering Model Intercomparison Project [57]. In the G3 scenario, SO_2 is injected into the stratosphere to keep the top of atmosphere radiation flux at 2020 levels by offsetting anthropogenic forcing from RCP4.5 [111]. The climate data we use in our analysis comes from the Max Planck Institute Earth System Model (MPI-ESM) [70]. We use stratospheric aerosol optical depth, surface air temperature, precipitation and cloud fraction fields generated by the experiment. We chose the MPI-ESM model because it participated in the GeoMIP G3 experiment and specifically modeled SRM with sulfur injection rather than direct dimming of the solar constant.[13] To calculate the total effect of SSAs on yields we compare a scenario with climate change – following RCP 4.5 – and SRM to a scenario with climate change only (CC) from 2050-2069:

$$\text{Total effect} = \text{Yields with SRM and climate change} - \text{Yields with climate change only} \tag{2.27}$$

IV.1 Model

The total effect of SRM on agriculture is the sum of SRM's effects through each determinant of yield that SRM affects. Our agricultural model from the previous section (Eqn. 2.12) breaks the total effect of SSAs down into an insolation effect, a temperature effect, a precipitation effect, a cloud effect and a residual effect:

$$\frac{dY}{dS} = \underbrace{\frac{\partial Y}{\partial I}\frac{\partial I}{\partial S}}_{\text{insolation effect}} + \underbrace{\frac{\partial Y}{\partial T}\frac{\partial T}{\partial S}}_{\text{temperature effect}} + \underbrace{\frac{\partial Y}{\partial P}\frac{\partial P}{\partial S}}_{\text{precipitation effect}} + \underbrace{\frac{\partial Y}{\partial C}\frac{\partial C}{\partial S}}_{\text{cloud effect}} + \underbrace{\frac{\partial Y}{\partial U}\frac{\partial U}{\partial S}}_{\text{residual effect}}$$

To get the change in yields due to SRM we multiply through by the change in SAOD due to a given SRM scenario (G, for geoengineering):

$$\begin{aligned}\frac{dY}{dG} = \frac{dY}{dS}\frac{dS}{dG} = \underbrace{\frac{dS}{dG}}_{\text{ESM}}\underbrace{\left[\frac{\partial Y}{\partial I}\frac{\partial I}{\partial S} + \frac{\partial Y}{\partial U}\frac{\partial U}{\partial S}\right]}_{\beta_{SAOD}} + \underbrace{\frac{dS}{dG}\frac{\partial T}{\partial S}}_{\text{ESM}}\underbrace{\frac{\partial Y}{\partial T}}_{\frac{\partial}{\partial T}f_T(.)} \\ + \underbrace{\frac{dS}{dG}\frac{\partial P}{\partial S}}_{\text{ESM}}\underbrace{\frac{\partial Y}{\partial P}}_{\frac{\partial}{\partial P}f_P(.)} + \underbrace{\frac{dS}{dG}\frac{\partial C}{\partial S}}_{\text{ESM}}\underbrace{\frac{\partial Y}{\partial C}}_{\frac{\partial}{\partial C}f_C(.)}\end{aligned} \tag{2.28}$$

[13] The MPI model does not internally calculate the evolution of SO_2 in the stratosphere; rather it intakes aerosol radiative properties calculated from an aerosol microphysical model that simulates equatorial injection of SO_2 at 60hPa.

Here, the change in yields due to SRM is expressed in terms that we can calculate using an ESM and our fitted empirical yield model. We model the SRM-induced changes in SAOD ($\frac{dS}{dG}$), temperature ($\frac{dS}{dG}\frac{\partial T}{\partial S}$), precipitation ($\frac{dS}{dG}\frac{\partial P}{\partial S}$) and cloud fraction ($\frac{dS}{dG}\frac{\partial C}{\partial S}$) induced by SRM using the MPI-ESM, and we estimate the impacts of these changes ($\beta, f_T(.), f_P(.), f_C(.)$) using our statistical yield model (Eqn. 2.16).

IV.2 Calculation of the total effect of SRM on crop yield

To calculate the total effect of SRM on crop yield we first use bilinear interpolation to regrid all of the historical climate observations to the same resolution as the climate model output. We then mean-debias the climate model data by month [24], average the climate model data over the growing season for each crop for each year and then evaluate our statistical crop model at these growing season values to get an estimate of the yield for each year. We apply the statistical crop model fitted to data from 1983-2009 (Fig. 3a Model 1) because the SSAs from Pinatubo were substantially more accurately measured than those of any other major eruption [113], and the cloud data does not reach back to the Chichón eruption. We do the steps above for each crop (maize, soy, rice, wheat), each scenario (SRM, CC), each ensemble member (each scenario is run three times with slightly different initial conditions), and each year. We then average across years and ensembles to get an estimate of the future expected yield for each crop under each scenario. To compare scenarios, we subtract their average yields.

Specifically, we calculate the effect of SRM on yields for each crop j in each year t and pixel p as the difference in yields in the SRM and CC scenarios:

$$\begin{aligned}\frac{\Delta Y}{\Delta G}_{tpj} &= Y_{tpj}^{SRM} - Y_{tpj}^{CC} \\ &= \beta_j S_{tpj}^{SRM} - \beta_j S_{tpj}^{CC} + f_{Tj}(T_{tpj}^{SRM}) - f_{Tj}(T_{tpj}^{CC}) \\ &\quad + f_{Pj}(P_{tpj}^{SRM}) - f_{Pj}(P_{tpj}^{CC}) + f_{Cj}(C_{tpj}^{SRM}) - f_{Cj}(C_{tpj}^{CC}),\end{aligned} \tag{2.29}$$

where superscripts denote the scenario.[14] We then calculate the expected change in yields due to SRM for each pixel (Extended Data Fig. 2.5) by averaging over 2050-2069:

$$\frac{\Delta Y}{\Delta G}_{pj} = \sum_{t \in 2050-2069} \frac{1}{20} \frac{\Delta Y}{\Delta G}_{tpj} \tag{2.30}$$

And finally we calculate the global average total effect of SRM on yields (Fig. 2.4e) by averaging expected yield changes over pixels:

$$\frac{\Delta Y}{\Delta G}_{world,j} = \sum_{p \in world} \nu_{pj} \frac{\Delta Y}{\Delta G}_{pj} \tag{2.31}$$

Here, ν_{pj} is a cropped-fraction and pixel-area weight such that $\sum_{p \in world} \nu_{pj} = 1$. We calculate the total effect separately for each ensemble member, as shown in Eqns. 2.29-2.31 and then average the effects. We do not include ENSO in our analysis of the

[14] Thus, $f_{Tj}(T_{tpj}^{SRM}) - f_{Tj}(T_{tpj}^{CC})$ calculates the temperature-mediated impact of SRM on the yield of crop j in year t and pixel p.

total effect because its long term average effect is mean zero by construction. CO_2 fertilization effects will be equal in the SRM and climate change only scenarios, and thus do not affect our calculation of the total effect of SRM.

We note that calculation of the total effect will be accurate even if the assumption in Eqn. 2.19 is not satisfied. This is because β captures all effects of SSAs on yields not mediated by temperature, precipitation or clouds and thus captures both the insolation and residual effects ($\frac{\partial Y}{\partial I}\frac{\partial I}{\partial S} + \frac{\partial Y}{\partial U}\frac{\partial U}{\partial S}$) in Eqn. 2.28, if any residual effects exist.

IV.3 Comparisons to historical scenarios

To corroborate our methodology, we compare the CC scenario to a historical scenario (1940-1959). Using a similar methodology to estimation of the total effect of SRM on yields, we estimate that changes in temperature (2.6 °C), precipitation (1.0 mm/month) and cloud fraction (-0.03) from 1940-1959 to 2050-2069 under RCP4.5 decrease yields by 18%, 6%, 16%, and 6% for maize, rice, soy, and wheat, respectively (Extended Data Fig. 2.7a). These estimates are within the range of estimates reported by the IPCC [83].

Similarly, we compare the SRM scenario to the historical scenario. We find that SRM and climate change together decrease yields by 17%, 6%, 14%, and 7% for maize, rice, soy, and wheat, respectively (Extended Data Fig. 2.7b). Comparing the SRM - historical impacts to the CC - historical impacts, we see that SRM mitigates yield losses from rising temperatures, but imposes equally-sized sunlight-mediated damages. Thus, we see again that, on net, SRM mitigates little to none of the damages from climate change to global agricultural production.

We note that we have not considered the effects of carbon fertilization when comparing SRM and climate change only scenarios to the historical scenario. Thus, we interpret these calculations as the impacts of changes in temperature, precipitation, clouds and sunlight only.

IV.4 Uncertainty of the total effect

We analyze uncertainty in the total effect of SSAs on yield from two sources. Statistical uncertainty comes from the fact that our crop model was estimated using historical data and thus its coefficients are not perfectly known due to sampling variability (this uncertainty is shown Extended Data Table 2 Column 5 and Extended Data Fig. 2). Variances depicted in maps of the total effect are of $\frac{\Delta Y}{\Delta G}_{pj}$ and variances depicted in the bar graphs are of $\frac{\Delta Y}{\Delta G}_{world,j}$. The error bars and hatchings in all figures that display analyses of the total SSA effect represent statistical uncertainty only. Model uncertainty comes from the fact that climate models do not perfectly predict the future climate. We estimate model uncertainty by calculating the effects separately for each of the three ensemble members rather than for their average (Extended Data Fig. 2.6) and find that our result that SRM mitigates little of the damages of climate change is stable across the three ensemble runs.

Chapter 3

Estimating the effect of cloud optical scattering on global crop yield

Anthropogenic emissions of air pollutants and greenhouse gases alter the amount, distribution and properties of cloud cover; yet the economic impacts of these manipulations remain largely unknown. Changing cloudiness may impact crop productivity by altering temperature, precipitation and sunlight. While the impacts of temperature and precipitation on crop productivity are relatively well understood, the impacts of changes in sunlight from cloud scattering remain poorly constrained because of the potentially offsetting effects of changes in total and diffuse sunlight. Here I leverage remotely-sensed cloud observations and subnational crop yield data from the United States, Europe, Brazil, and China to provide the first empirical estimates of the sunlight-mediated effect of cloud optical scattering on maize and soy yields. I find a consistent concave response of yields to cloud optical thickness across crops and regions. Changing ten days in the growing season from clear to the optimal cloud thickness increases maize and soy yields by 4.0% and 4.4%, respectively; further increasing cloud thickness to the 95th growing season percentile decreases maize and soy yields by 3.4% and 3.5%. Mechanistically, I find that the concavity in the cloud response is driven by concavity in the response to total sunlight as well as – in some regions – benefits from increased diffuse light. Applying these empirical estimates to earth system model simulations, I find that changes in sunlight, due to anthropogenic air pollution-induced changes in clouds, are suppressing maize and soy yields by as much as 5% in heavily polluted areas of India and China by increasing the frequency of days with extremely high cloud optical depths. This costs Chinese maize farmers roughly US$1 billion a year. Changes in sunlight due to changes in clouds from a quadrupling of CO_2 relative to pre-industrial tend to decrease global maize yields and redistribute soy yields. The methodology developed in this paper could be extended study the impact of changes to the global optical environment on other global-scale economic outcomes.

Anthropogenic activity such as emissions of air pollutants and greenhouse gases alter the amount, distribution and properties of global cloud cover; yet the agricultural impacts of these manipulations due to the resulting changes in sunlight remain largely unknown [11, 72, 45, 125, 126]. While the impacts of changes in temperature and precipitation on yields have been well documented [104, 118], the impacts of changes in sunlight on yields are less well understood.

A key uncertainty regarding the radiative impacts of cloud scattering on yields is to what degree the benefits from increased diffuse light outweigh losses from reductions in total light. Since each leaf of a crop has diminishing marginal photosynthetic returns to sunlight, atmospheric scattering is thought to increase radiation use efficiency by redistributing light from the sun-saturated leaves at the top of the canopy to the shaded leaves below [56, 60]. The strength of this diffuse fertilization effect (DFE) – and, in turn, atmospheric scattering's impact on yield – is highly debated, with some analyses showing large benefits of atmospheric scattering to plant productivity and crop yields [102, 60, 135, 36] while others find substantial costs [38, 84, 108, 7]

The impact of scattering on agricultural yield depends primarily on the optical properties of the scattering particle, and the relative contributions of direct and diffuse light to the production of yield. This, in turn, can depend on crop type, climate conditions, nutrient availability, and importantly, the background optical environment upon which additional scattering occurs [56, 35]. Previous empirical studies show positive effects of solar brightening in the U.S. [114] and increased sunlight hours [136] in China, and negative effects of tropospheric [38] and stratospheric [84] aerosol scattering on Indian and global crop yields, respectively. Simulation studies [102], controlled experiments [60], and studies using flux tower measurements [135, 36], however, find that optical scattering increases plant productivity and crop yields.

Part of this discrepancy may be explained by a non-linear relationship between optical scattering and crop yields. At low levels, scattering may benefit yields by increasing light diffusivity and reducing damages from extreme insolation such as those from photoinhibition and other cellular damage, water stress, or high leaf temperatures [56]. With more intense scattering, however, total light availability can be so low that it suppresses yields. Such a non-linear response has been posited and described for un-managed ecosystem productivity [56, 75] but never before empirically estimated for agricultural yields, which may respond differently than edible yield [84]. Empirically quantifying the impact of optical scattering on crop yields provides an improved, generalized understanding of agricultural response to changes in insolation and, in turn, informs the impacts of historical, current, and future changes in the optical environment on global agricultural productivity such as those due to air pollution, climate change, or potentially solar geoengineering [11, 125, 84].

This study combines remotely-sensed cloud observations from the International Satellite Cloud Climatology Project (ISCCP) [98] (Fig. 3.1a) and subnational yield data from the United States, European Union, China and Brazil (Extended Data Fig. 3.1) to provide, to my knowledge, the first empirical estimates of the sunlight-mediated impact of clouds on international maize and soy yields. Estimating the non-linear impact of cloud optical scattering on yields tests the strength of the diffuse fertilization

effect at a global scale and over a range of scattering environments. In turn, this enables the first empirically-based estimates of the cloud-mediated effects of climate change and air pollution on yields due to changes in sunlight. The study's large sample (N = 166,651 administrative-unit-by-year observations) enables estimation of an internationally-representative response by averaging out potentially unrepresentative heterogeneous effects and noise. Studying clouds as a source of atmospheric scattering enables precise estimation of a nonlinear response because clouds are the primary determinant of growing season to growing season variation in insolation and span a range of optical depths an order of magnitude larger other sources of atmospheric scattering, such as sulfate or black carbon aerosols.

The theoretically ideal experiment would measure the non-linear impact of cloud scattering on yield by seeding clouds of different thicknesses above a set of administrative units (i.e. counties), cooling and drying these units to hold all variables other than sunlight fixed, and comparing the resulting yields to a set of identical and un-manipulated control units. In practice, I approximate this ideal experiment using historical growing season to growing season variation in cloudiness. I identify the sunlight-mediated impact of cloud scattering on yield by comparing administrative units (e.g. counties) to themselves over time with differing cloud amounts and cloud optical depths – a measure for how difficult it is for light to pass directly through the cloud without being scattered or absorbed. I use a multivariate fixed-effects panel estimation strategy to account for unobserved time-invariant factors, such as soil type, as well as administrative-unit-specific time-trending variables, such as access to agricultural technologies (Supplementary Information Eqn. 3.12). I isolate the effects of changes in sunlight by flexibly accounting for potentially confounding climate variables including temperature, precipitation, wind speed and aerosol optical depth (Supplementary Information Section III.3). These variables are correlated with both cloud cover (Fig. 3.1 and Extended Data Fig. 3.2) and yield (Extended Data Fig. 3.6) and thus would bias estimates if not accounted for (Extended Data Fig. 3.4). I validate the model by verifying that the estimated regional responses of crop yield to temperature and precipitation are consistent with previous estimates in the literature (Extended Data Fig. 3.6) [104].

I empirically estimate the impact of cloud optical depth on total, direct and diffuse shortwave insolation using 859 global insolation monitors (N = 3,428,474 station-days) to inform the mechanisms driving the impact of cloud scattering on yield and to validate the remotely sensed cloud data (Supplementary Information Section II). I find that, given an average cloud amount of 70 %, diffuse insolation peaks at 13.5 optical depth, increasing 36.5% or 400 $\frac{Wh}{m^2 day}$ from baseline; total light decreases across the entire support (Fig. 3.1, Extended Data Fig. 3.3, Extended Data Table 3.1). These findings globally generalize previous analyses that use only a handful of stations [56, 75].

In turn, I find that the sunlight-mediated effect of cloud scattering on maize and soy yields is concave, with a yield-maximizing optical depth of 15 (Fig. 3.2). Increasing cloud optical depth for ten days during the growing season from 0 (clear-sky) to 15 increases maize and soy yields by 4.0% and 4.4%, respectively, due to changes in sunlight

– given the growing season average cloud amount of 70% (Supplementary Information Section III.2). Relative to a clear day, an increase in 15 optical depth coincides with an increase in diffuse light of 36% or 391 $\frac{Wh}{m^2 day}$ and a decrease in total light of 51 % or 2,284 $\frac{Wh}{m^2 day}$. Further increasing cloud scattering to 30 optical depth – roughly the 95th percentile growing season thickness – for 10 days decreases growing season maize and soy yields by 3.4% and 3.5%, relative to an optical depth of 15.

Compared to an optically ideal growing season with a constant optical depth of 15, realized optical depths lower cropped-fraction-weighted-global-average maize yields by 51% and soy yields by 56% (Supplementary Information section III.8). This is a similar reduction to that caused by realized temperatures, which previous analyses have shown suppress yields by 48% relative to the thermally ideal growing season [14].

The finding of a concave cloud scattering effect peaking near 15 optical depth is consistent across the two crops and four regions analyzed. Regional heterogeneity in the strength of this effect may be due either to heterogeneous impacts of clouds on sunlight – potentially due to differing solar zenith angles or cloud optical properties – or to heterogeneous sensitivities of crops to sunlight – potentially due to differing varietals, climate conditions, or nutrient and water availability [56]. The estimated impact of cloud optical scattering is robust to adding or altering the climate controls (i.e. additionally controlling for vapor pressure deficit, multiple flexible forms of average, maximum and minimum daily temperature, a different precipitation data set, and nighttime cloud cover), changing the fixed-effects in the panel regression (to include cubic adm-2 level trends or to use quadratic adm-1 level trends), estimating the response with alternative functional forms (i.e. non-parametrically using bins and with additional knots in the restricted cubic spline), and weighting the response by planted area (Extended Data Fig. 3.5) .

Mechanistically, the damages from extreme cloudiness are likely due to light limitation. The benefits of scattering at optical depths less than 15, however, could be due to the diffuse fertilization effect, to reduced damage from extreme insolation, or to a combination of the two [56]. To empirically test the relative contributions of these potential mechanisms, I estimate the effect of photosynthetically active radiation (PAR) [30] on yields, allowing the effect of PAR to vary as a function of the diffuse fraction (Supplementary Information Section IV, Extended Data Table 3.2).

I find that both maize and soy yields – in the pooled sample and regional subsamples – have a concave response to PAR when the impact of PAR is evaluated at the average diffuse fraction observed for each level of PAR (Fig. 3.3 a). This is consistent with the finding of a concave response to cloud scattering. I find that this concavity – and thus the concavity in the response to cloud scattering – is driven by a concave response to total sunlight and not by the diffuse fertilization effect in the pooled sample (Fig. 3.3 c). I do, however, find evidence for the diffuse fertilization effect in the Chinese maize, U.S. maize and E.U. soy subsamples. For U.S., Chinese and E.U. maize, a positive diffuse fertilization effect is the primary driver of concavity in the PAR, and thus cloud scattering, response (Extended Data Fig. 3.7). For Brazilian maize and all regions of soy the response to PAR is concave regardless of the diffuse fraction and only E.U. soy

has a positive diffuse fertilization effect. Future research should investigate the drivers of these regional differences in the mechanisms driving the concave responses to cloud scattering.

Anthropogenic activities such as the emission of aerosol precursors or greenhouse gasses shape the distribution and properties of global clouds and, in turn, the global optical environment. To estimate how anthropogenic changes in clouds alter sunlight and, in turn, yields I apply empirical relationships between cloud optical scattering and crop yield (Supplementary Information Section III.7 and Extended Data Fig. 3.9) to output from five global climate models and compare yields in two scenerios. (1) I estimate the sunlight-mediated effect of anthropogenic air pollution by comparing the contribution of cloud scattering to yields in identical worlds with global aerosol distributions set to pre-industrial (1860) and year 2000 levels, following [133]; (2) I estimate the sunlight-mediated effect of climate change by comparing the contribution of cloud scattering to yields in identical worlds with global carbon dioxide concentrations set to pre-industrial and quadrupled-pre-industrial levels, following [133] (Supplementary Information section V).

I find that changes in sunlight due to changes in clouds from to anthropogenic aerosols have, on average, decreased maize and soy yields (Fig. 3.4a-b) by decreasing the frequency of medium optical depth clouds ($3.6 < \text{OD} < 23$) – which tend to increase yields – and increasing the frequency of extremely thick ($\text{OD} > 23$) clouds – which tend to decrease yields (Extended Data Fig. 3.8). In heavily polluted areas such as China, maize and soy yields are reduced by 2.7% and 1.1%, respectively, amounting to roughly US$1 billion a year of Chinese maize production. I find that climate-change-induced changes in cloud optical scattering tend to decrease maize yields and redistribute soy yields (Fig. 3.4c-d, Supplementary Information Section V). Though these projections are not predictions of climate change or air pollution impacts because they assume away adaptation, estimate only the sunlight-mediated effects due to changes in cloud cover thereby omitting other effects such as direct damages from air pollutants or benefits from carbon dioxide fertilization, and rely on notoriously difficult-to-predict changes in cloud distributions [11] that incite substantial inter-model variation into predicted effects (Extended Data Fig. 3.10); they do provide the first demonstration of how empirical and physical models can be paired to quantify the agricultural externalities stemming from anthropogenic dimming and brightening of the global optical environment due to changes in cloud cover [125].

The finding of a statistically significant and economically substantial non-linear effect of cloud optical scattering on yield suggests that the impact of future anthropogenically-induced atmospheric scattering – due to air pollution, climate change, or potentially geoengineering – will depend on the background optical depth. Farmer adaptation, which is not modelled explicitly, could theoretically lessen any negative impacts of future changes to the global optical environment, yet the degree to which farmers have historically adapted to changes in insolation, and their ability to adapt to future changes is unknown.

The finding of a concave response of yields to atmospheric scattering is consistent with theoretical predictions and observations of unmanged ecosystem productivity [56],

yet the estimated damages from high cloud optical depths are not as severe as past measurements of damages due to increased tropospheric and stratospheric aerosol optical depth [37, 84]. This difference may be explained by either increasing cloudiness being correlated with some omitted variable that both acts to increase yield and is uncorrelated with the model's controls in both the primary specification and robustness checks, or by differing impacts of clouds and aerosols on the intensity, diffusivity and spectral distribution of insolation [56, 44] causing different impacts on crops.

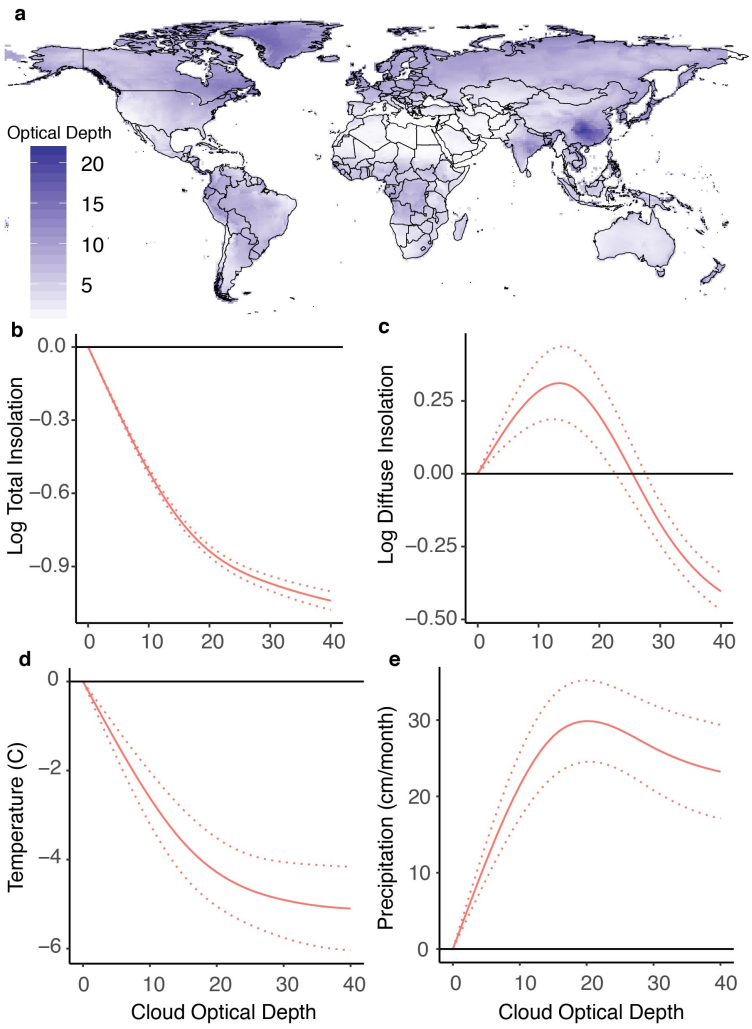


Figure 3.1: Clouds alter the global optical environment. a, Average growing season cloud optical depth 1983-2009 from the International Satellite Cloud Climatology Project. **b-c,** Empirical estimates of the effect of clouds on total (N = 3,428,474 station-days) and scattered (N= 928,202) sunlight using a network of global stations (Supplementary Information Section II). **d-e,** Empirical correlation between growing season cloud optical depth and growing season temperature and precipitation. Dashed lines in b-e show 95% confidence intervals. Responses to optical depth are shown for the average growing season cloud amount of 70%.

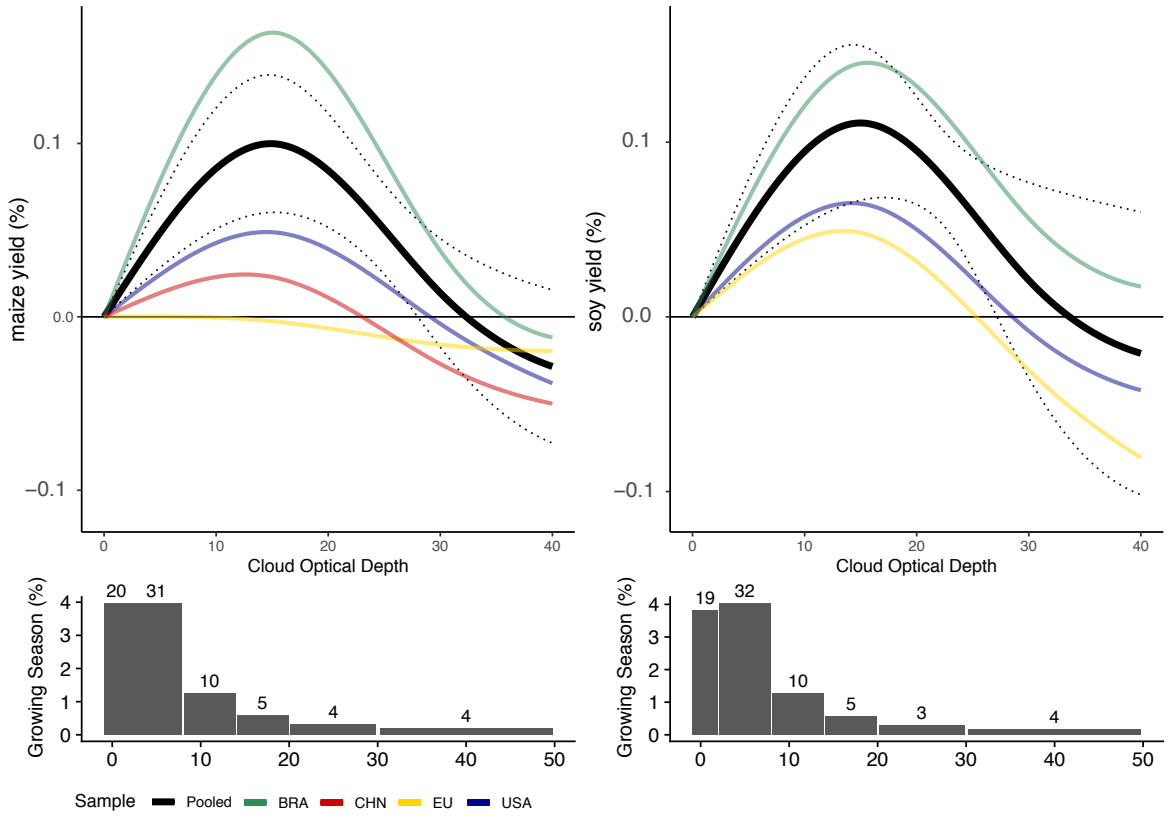


Figure 3.2: Empirical estimates of the sunlight-mediated effect of cloud scattering on crop yield. Curves show the estimated effect of increasing the cloud optical depth of cloudy areas from zero to a given value for three daytime hours during the growing season on growing season yield. A consistent concave response of yields to cloud scattering is recovered in the pooled sample (black, N = 166,651 for maize and 96,727 for soy), as well as in Brazil (green, N = 93,468 for maize and 61,480 for soy), China (red, N = 27,451 for maize), the European Union (yellow, N = 2,248 for maize and 703 for soy), and the United States (blue, N = 43,484 for maize and 34,544 for soy). Regional impacts are estimated independently. Models include climate controls and adm-2 (e.g. county) specific fixed effects and adm-2 specific quadratic time trends. Responses to optical depth are shown for the average growing season cloud amount of 70%. Dotted lines represent the 95% confidence interval for the pooled effect, which is calculated allowing for arbitrary temporal and spatial correlation within adm-1 (e.g. state) units. The histogram shows the distribution of daytime 3-hourly cloud optical depth during the growing season.

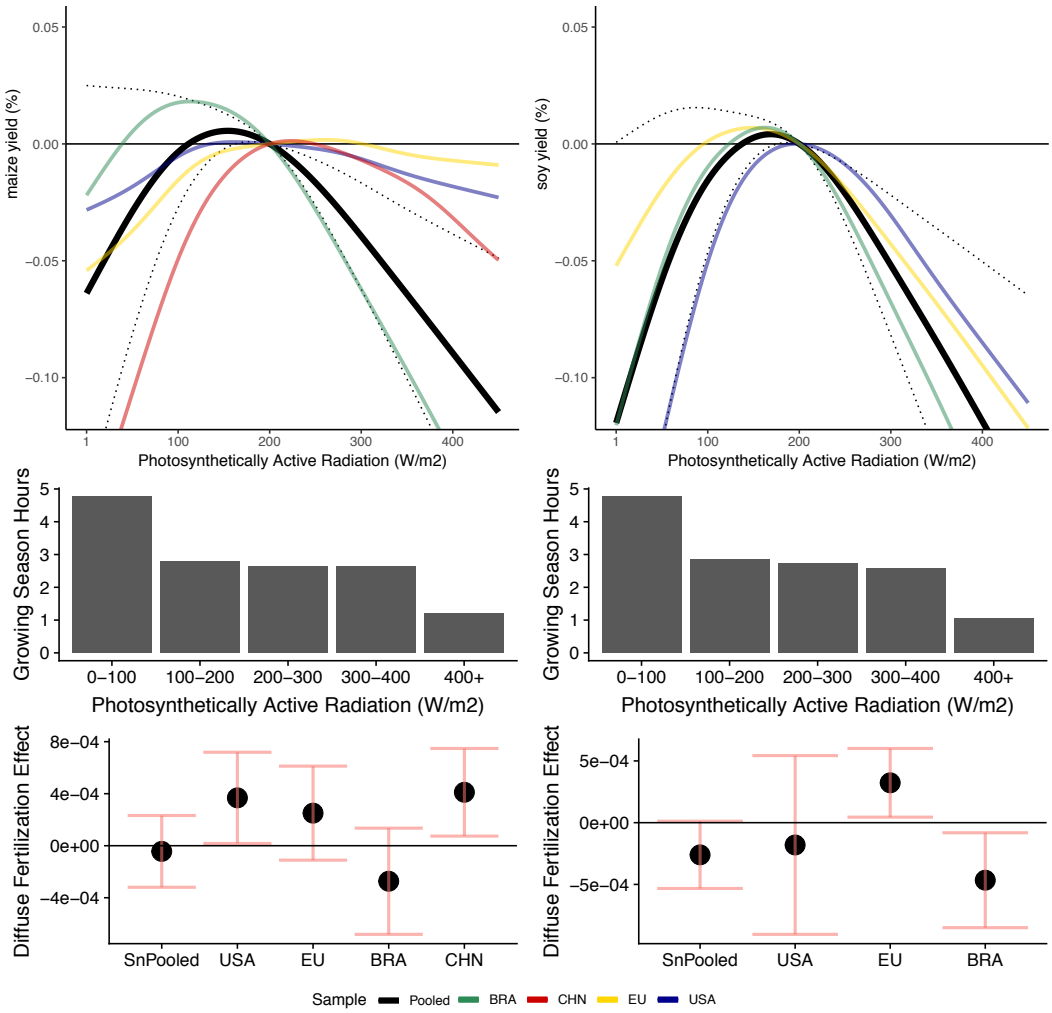


Figure 3.3: Empirical estimates of the effect of sunlight on crop yield. The estimated effect of changing photosynthetically active radiation (PAR) on maize and soy yields for an hour during the growing season. A consistent concave response of yields to PAR is recovered in the pooled sample (black), as well as in Brazil (green), China (red), the European Union (yellow), and the United States (blue). The effect of changing PAR is evaluated at the average diffuse fraction for each level of PAR. Models include climate controls and adm-2 (e.g. county) specific fixed effects and adm-2 specific quadratic time trends. Dotted lines represent the 95% confidence interval for the pooled effect, which is calculated allowing for arbitrary temporal and spatial correlation within adm-1 (e.g. state) units. The histogram shows the distribution of hourly PAR during each day in the growing season in the pooled sample. Whisker plots show empirical estimates of the impact of the diffuse fraction on the marginal impact of PAR on crop yield, which is a measure for the strength of the diffuse fertilization effect (α_3 in Supplementary Information Equation 3.22). If $\alpha_3 = 0.004$, as in U.S. maize, increasing the diffuse fraction for an hour from 0 to 1 when PAR is at 200 $\frac{W}{m^2}$ increases growing season yields by 0.08 %.

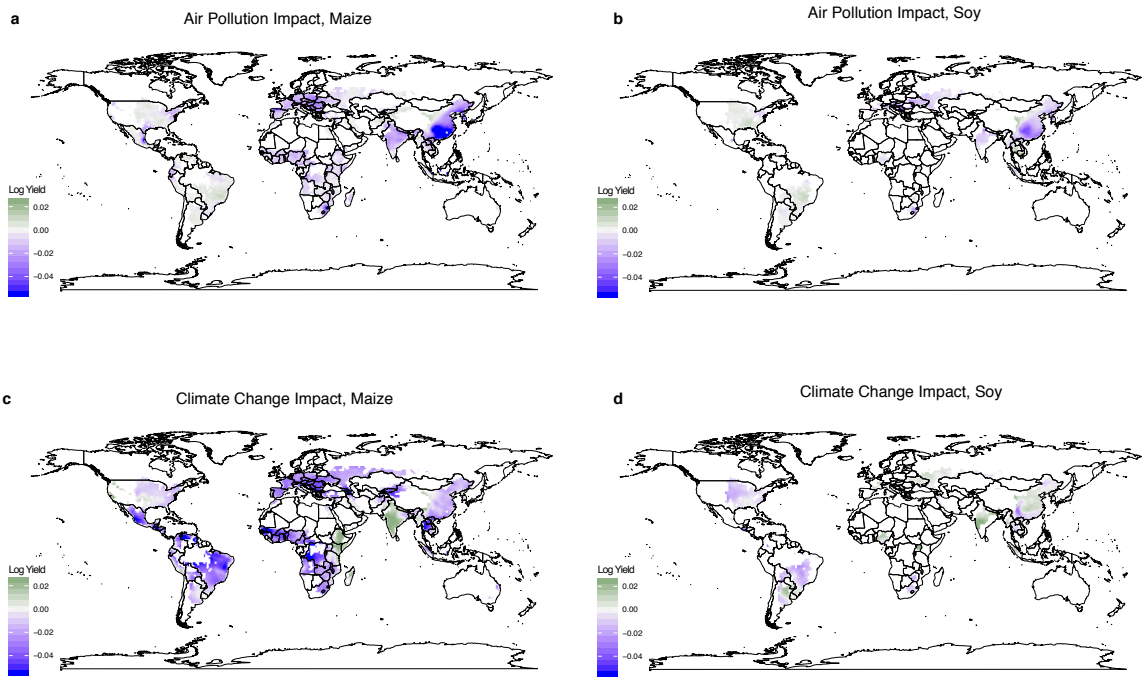


Figure 3.4: The sunlight-mediated impact of anthropogenic changes in cloud cover on global crop yield. **a**,**b** The sunlight-mediated effect due to changes in cloud cover of year-2000 relative to pre-industrial air pollution on maize and soy yields. **c**,**d** The sunlight-mediated effect due to changes in cloud cover of a quadrupling of pre-industrial atmospheric carbon dioxide concentrations on maize and soy yields.

Methods

To match the administrative yield data from the United States (obtained from the United States Department of Agriculture's National Agricultural Statistics Service), Brazil (from the Brazilian Institute of Geography and Statistics) the European Union (from [47]) and China (from [47]) to the climate data, I summarize all gridded cloud [98], temperature[1], precipitation [127], aerosol optical depth [30], wind speed [30], and insolation [30] data to the growing season administrative unit level by taking the mean of values over cropped area [67] and the growing season [100] using a methodolgy similar to previous studies [104] and tailored to the analysis in this paper (Supplementary Information Section III.5). In all aggregation and analyses, clear sky is defined as having a cloud optical depth of 0.

To estimate the impact of cloud scattering on insolation I pair the ISCCP cloud data with station measurements of total, direct and diffuse insolation [129]. I model the effect of cloud scattering on log yields using restricted cubic splines, and account for station-by-day-of-year fixed effects to account for any season-specific differences between stations (Supplementary Information Equation 3.1). The calculation of standard errors allows for arbitrary correlation across stations within a country over time as well as within a year across stations.

To estimate the response of cloud optical scattering to crop yields I model the impacts of cloud optical depth, temperature and precipitation using restricted cubic splines, and the effect of maximum hourly wind speed and aerosol optical depth using cubic polynomials (Supplementary Information Equation 3.13). To estimate the response of crop yields to PAR I model the impact of PAR using restricted cubic splines, and allow that impact to vary linearly with the diffuse fraction (Supplementary Information Equation 3.22); I model the climate controls identically to those in the model of cloud optical scattering on crop yields. The calculation of standard errors for the regressions of crop yield on clouds and sunlight allow for arbitrary correlation of observations across space and over time within an administrative level-1 unit (e.g. state). Analyses are conducted separately for maize and soy, as well as for estimation of regional effects.

Data used to calculate the sunlight-mediated impact of anthropogenic changes in cloud distributions on crop yield come from the Coupled Model Intercomparison Project models: CanESM2, HadGEM2 (-A and -ES for the aerosol and climate change scenarios, respectively), IPSL-CM5A-LR, MIROC5 and MRI-CGCM3.

	Log Insolation		
	Total	Diffuse	Direct
	(1)	(2)	(3)
Cloud Optical Depth RCS Feat. 1	-0.077^{***}	0.050^{***}	-0.271^{***}
	(0.001)	(0.009)	(0.007)
Cloud Optical Depth RCS Feat. 2	0.163^{***}	-0.331^{***}	0.719^{***}
	(0.006)	(0.043)	(0.031)
Cloud Optical Depth RCS Feat. 3	-0.307^{***}	0.802^{***}	-1.444^{***}
	(0.014)	(0.101)	(0.071)
Climate Controls	None	None	None
Projected R2	0.628	0.182	0.589
Observations	3,428,474	928,202	928,202
R^2	0.887	0.757	0.766
Note:			*p<0.1; **p<0.05; ***p<0.01

Extended Data Table 3.1: Effect of cloud optical depth on total, diffuse, and direct insolation Coefficients are the maringal impact of the restricted cubic spline features of cloud optical depth (plotted in Fig. 3.1). All models include station-by-day-of-year fixed effects. Standard errors, shown in parentheses, are clustered by country and by year to account for serial correlation over time within a country and for autocorrelation across space within a year.

	Insolation		
	Total	Diffuse	Direct
	(1)	(2)	(3)
MERRA2 Total PAR	2.002***		
	(0.071)		
MERRA2 Diffuse PAR		1.188***	
		(0.091)	
MERRA2 Direct PAR			1.864***
			(0.079)
Climate Controls	None	None	None
Projected R2	0.466	0.159	0.45
Observations	4,696,137	1,149,146	1,149,146
R^2	0.848	0.713	0.747
Note:			*p<0.1; **p<0.05; ***p<0.01

Extended Data Table 3.2: Comparison of total, diffuse and direct MERRA2 PAR to WRDC station measurements. Coefficients represent the increase in measured station insolation that correlates with a 1 unit increase in MERRA2 predicted photosynthetically active radiation. Coefficients tend to be larger than 1 because only a portion of the shortwave solar spectrum measured by the WRDC stations is photosynthetically active. All models include station-by-day-of-year fixed effects. Standard errors, shown in parentheses, are clustered by country and by year to account for serial correlation over time within a country and for autocorrelation across space within a year.

	Maize yield (log)				
	Subnational Pooled	US	EU	China	Brazil
Cloud Optical Depth RCS Feat. 1	0.101***	0.056***	0.001	0.027	0.128***
	(0.021)	(0.015)	(0.022)	(0.021)	(0.035)
Cloud Optical Depth RCS Feat. 2	−0.549***	−0.322***	−0.045	−0.204	−0.679**
	(0.133)	(0.107)	(0.139)	(0.131)	(0.268)
Cloud Optical Depth RCS Feat. 3	1.306***	0.733***	0.129	0.504	1.650**
	(0.333)	(0.284)	(0.358)	(0.327)	(0.703)
Climate Controls	T,P,A,W	T,P,A,W	T,P,A,W	T,P,A,W	T,P,A,W
Projected R2	0.065	0.254	0.167	0.066	0.045
Observations	166,651	43,484	2,248	27,451	93,468
R^2	0.993	0.721	0.864	0.713	0.817

	Soy yield (log)			
	Subnational Pooled	US	EU	Brazil
Cloud Optical Depth RCS Feat. 1	0.116***	0.068***	0.065	0.132***
	(0.028)	(0.018)	(0.047)	(0.045)
Cloud Optical Depth RCS Feat. 2	−0.619***	−0.405***	−0.432	−0.654*
	(0.209)	(0.099)	(0.285)	(0.390)
Cloud Optical Depth RCS Feat. 3	1.483***	0.978***	0.990	1.561
	(0.538)	(0.246)	(0.720)	(1.029)
Climate Controls	T,P,A,W	T,P,A,W	T,P,A,W	T,P,A,W
Projected R2	0.131	0.293	0.174	0.1
Observations	96,727	34,544	703	61,480
R^2	0.994	0.708	0.725	0.792

Note: *p<0.1; **p<0.05; ***p<0.01

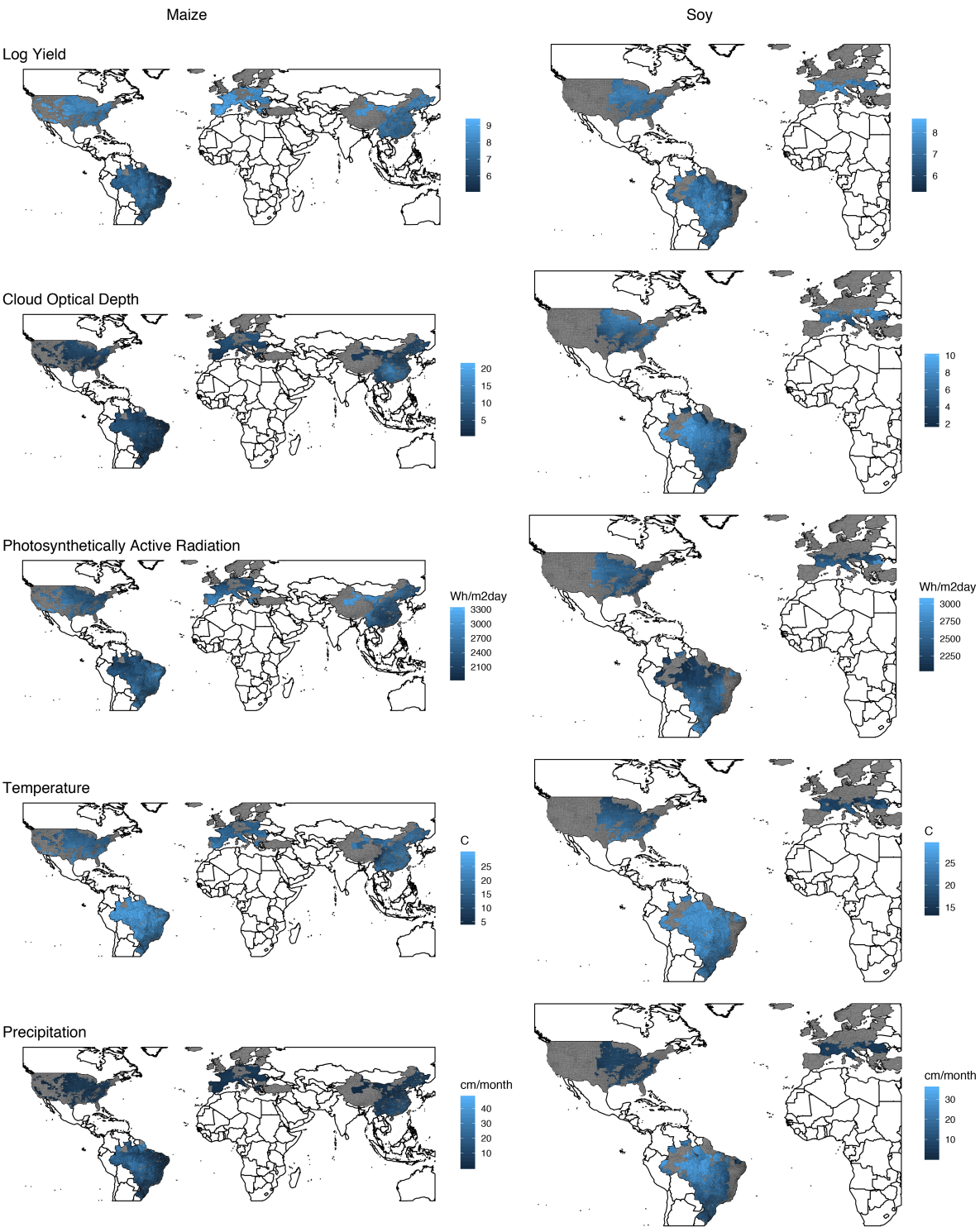
Extended Data Table 3.3: The impact of cloud optical scattering on maize and soy yields. Coefficients represent the marginal impact of the changing the growing season average of restricted cubic spline features of cloud optical depth on yield (plotted in Fig. 3.2). All models include temperature (T), precipitation (P), aerosol optical depth (A) and precipitation (P) controls as well as administrative unit fixed effects and administrative-unit-specific quadratic time trends. Standard errors, shown in parentheses are clustered by adm-1 (e.g. state) to account for correlation over time and space within adm-1 units.

	Maize yield (log)				
	Subnational Pooled	US	EU	China	Brazil
	(1)	(2)	(3)	(4)	(5)
Sunlight (PAR) RCS Feat. 1	0.027	−0.003	−0.002	0.027*	0.025
	(0.016)	(0.019)	(0.020)	(0.015)	(0.039)
Sunlight (PAR) RCS Feat. 2	−0.051**	−0.002	−0.001	−0.042*	−0.053
	(0.024)	(0.032)	(0.031)	(0.023)	(0.057)
Diffuse Fert. Eff. (PAR x DF)	−0.002	0.010	0.015**	0.013**	−0.008
	(0.005)	(0.007)	(0.007)	(0.005)	(0.006)
Climate Controls	T,P,A,W	T,P,A,W	T,P,A,W	T,P,A,W	T,P,A,W
Projected R2	0.065	0.252	0.173	0.067	0.044
Observations	166,651	43,484	2,248	27,451	93,468
R^2	0.993	0.720	0.865	0.713	0.817

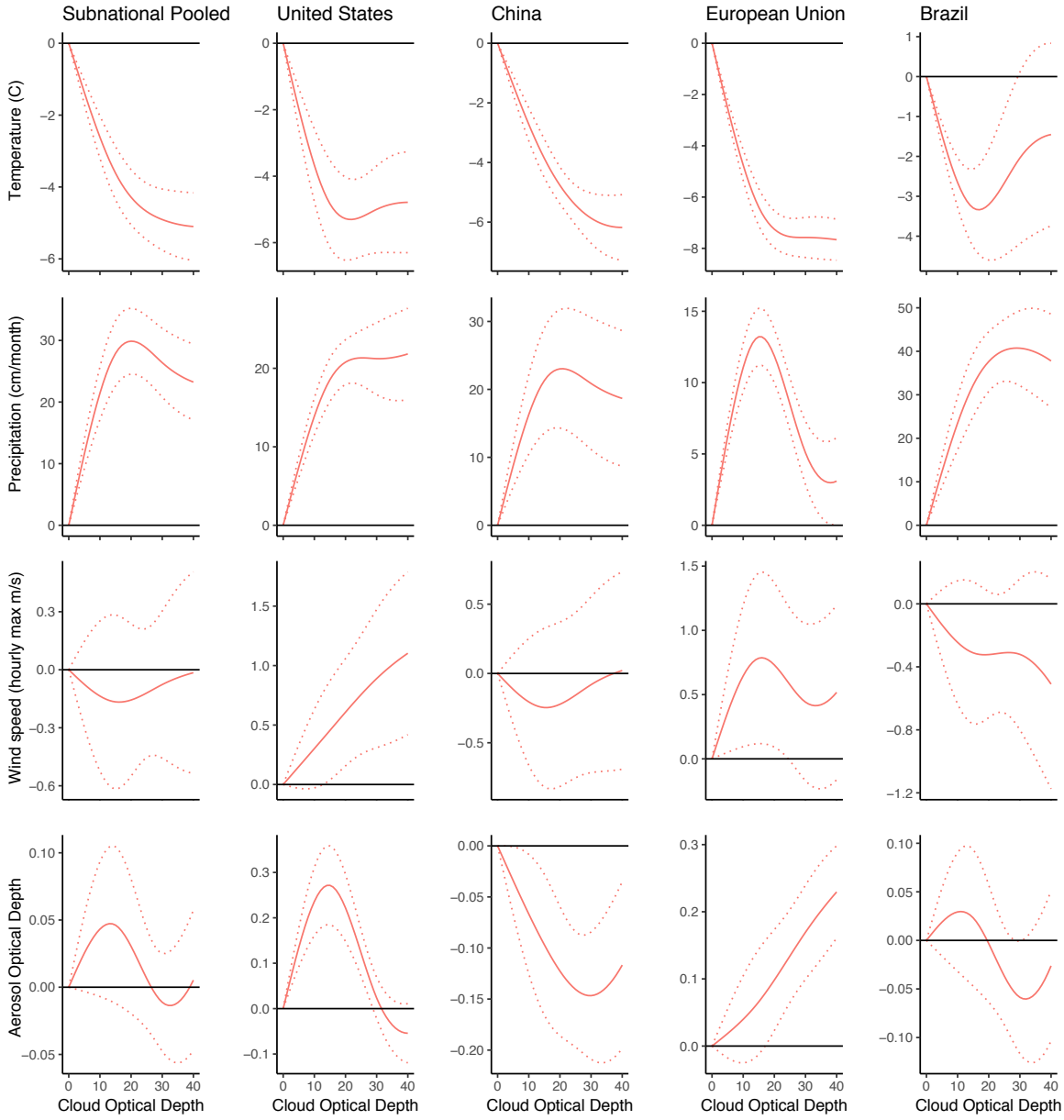
	Soy yield (log)			
	Subnational Pooled	USA	EU	Brazil
	(1)	(2)	(3)	(4)
Sunlight (PAR) RCS Feat. 1	0.057***	0.048***	0.035	0.066
	(0.022)	(0.018)	(0.030)	(0.042)
Sunlight (PAR) RCS Feat. 2	−0.095***	−0.078***	−0.065	−0.111*
	(0.033)	(0.028)	(0.054)	(0.063)
Diffuse Fert. Eff. (PAR x DF)	−0.010*	0.011**	−0.007	−0.016**
	(0.005)	(0.005)	(0.015)	(0.007)
Climate Controls	T,P,A,W	T,P,A,W	T,P,A,W	T,P,A,W
Projected R2	0.128	0.301	0.159	0.097
Observations	96,727	34,544	703	61,480
R^2	0.994	0.712	0.720	0.791

Note: *p<0.1; **p<0.05; ***p<0.01

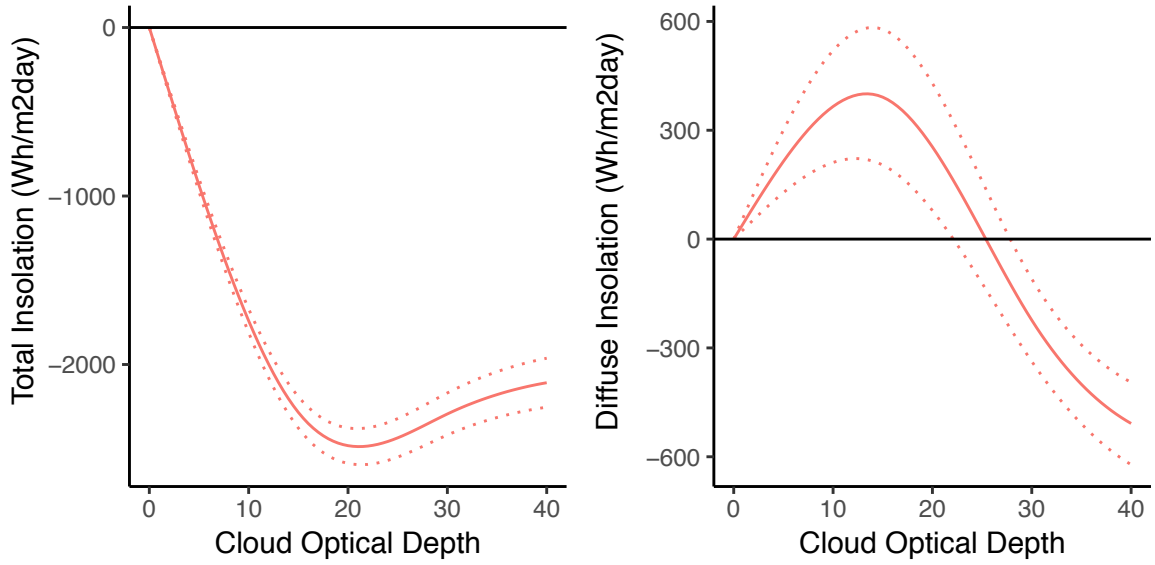
Extended Data Table 3.4: The impact of PAR on maize and soy yields. Coefficients represent the marginal impact of changing the growing season average of the restricted cubic spline features of PAR as well as the interaction between PAR and the diffuse fraction (plotted in Fig. 3.3). All models include temperature (T), precipitation (P), aerosol optical depth (A) and precipitation (P) controls as well as administrative unit fixed effects and administrative-unit-specific quadratic time trends. Standard errors, shown in parentheses are clustered by adm-1 (e.g. state) to account for correlation over time and space within adm-1 units.



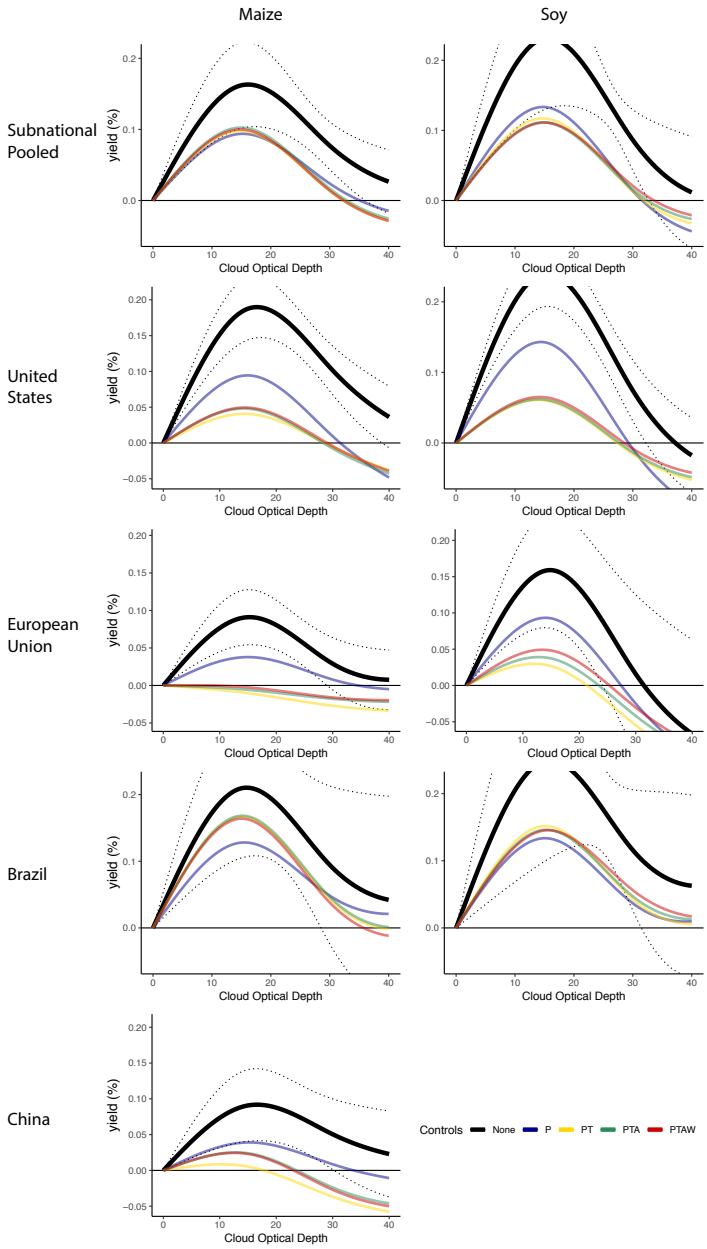
Extended Data Figure 3.1: Measurements of maize and soy yield and some of their climatological determinants. Twenty-five-year maize and soy growing-season-average yield, cloud optical depth, photosynthetically active radiation, temperature, and precipitation in the subnational areas included in the analysis.



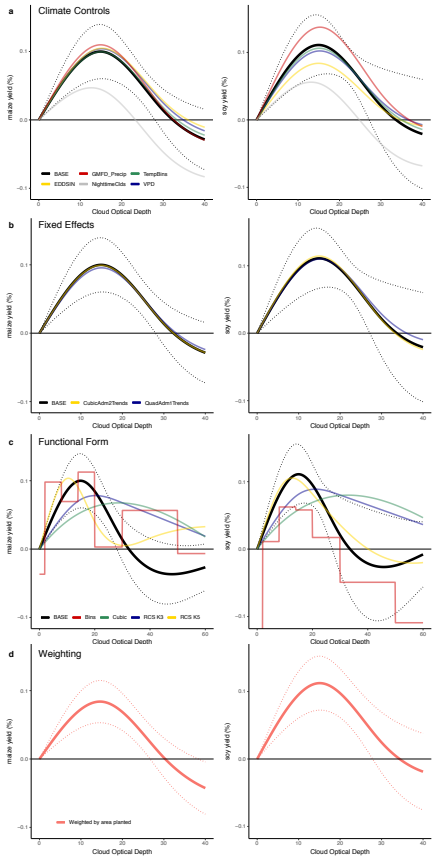
Extended Data Figure 3.2: Correlation of cloud optical depth with other climatological determinants of yield. The estimated correlation of temperature, precipitation, wind speed and aerosol optical depth with cloud optical depth in the pooled sample. Regressions to estimate the effect of cloud optical depth on each variable include the other three climate variables and administrative level-2 fixed effects and quadratic time trends as controls to mirror the identifying variation in the estimation of cloud impacts on yields. Dotted lines represent the 95% confidence interval, which is calculated allowing for arbitrary temporal and spatial correlation within adm-1 (e.g. state) units.



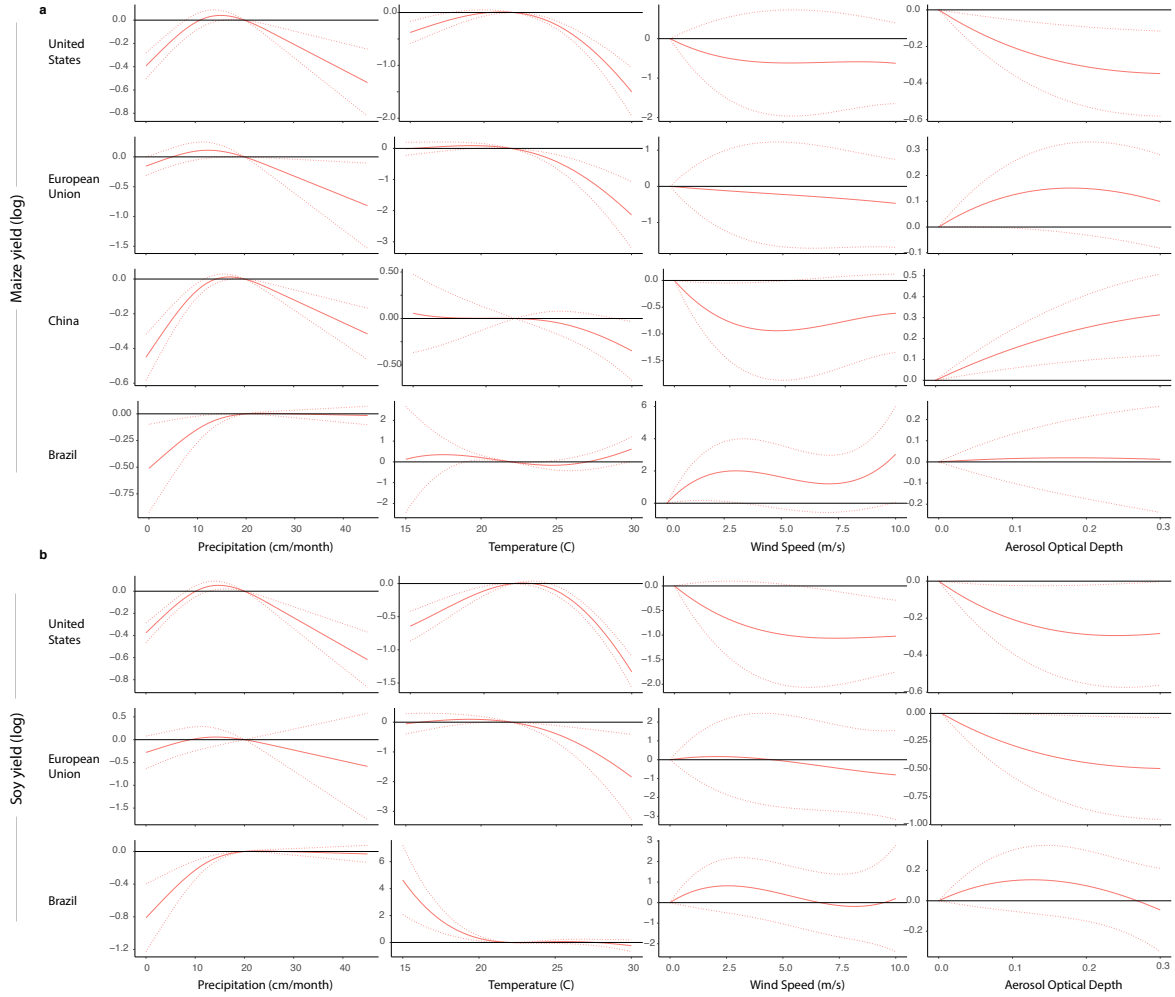
Extended Data Figure 3.3: Empirical estimates of the impact of clouds on shortwave insolation. Estimates are identical to those in Fig. 3.1 except that sunlight is modeled in levels rather than in logs (Supplementary Information Section II).



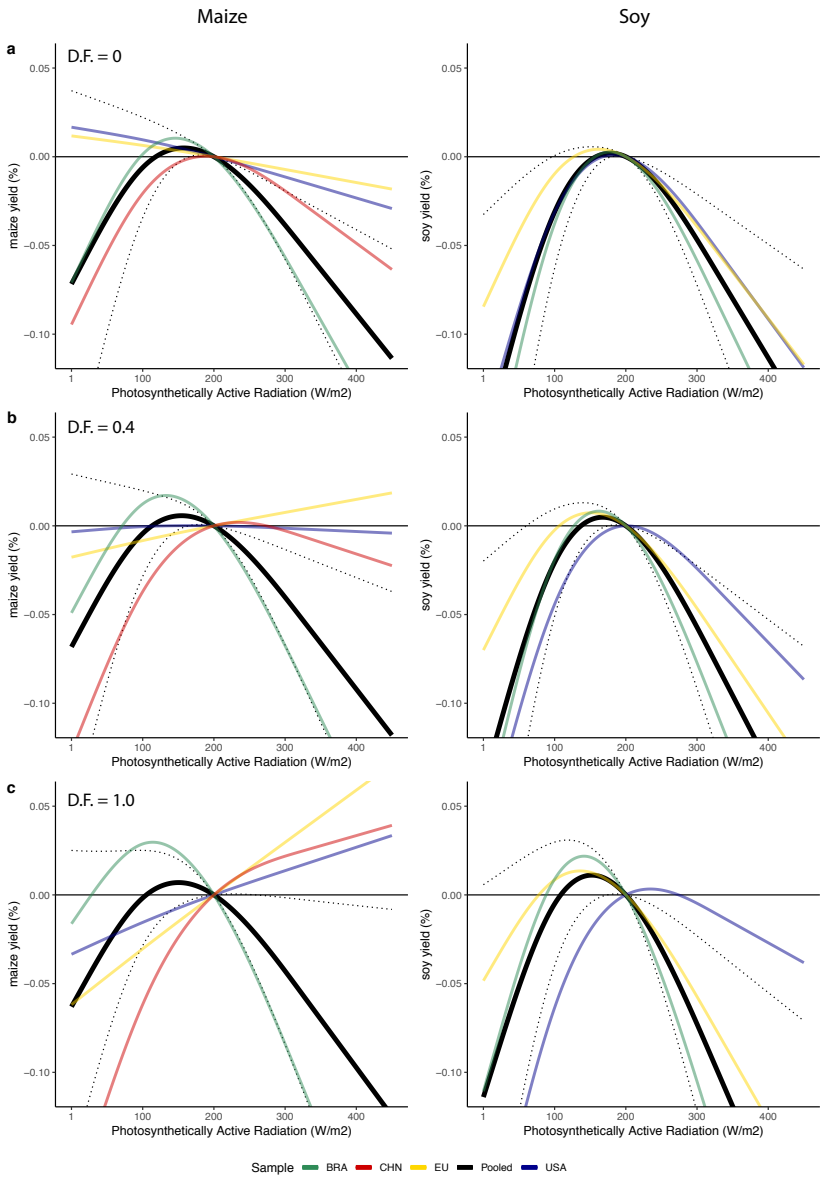
Extended Data Figure 3.4: The effect of adding controls to the estimated effect of cloud scattering on maize and soy yields. The estimated effect of cloud scattering on maize and soy yields in the pooled and regional samples adding controls in one-at-a-time. Generally, adding precipitation (P) and temperature (T) into the model decrease the benefits of cloudiness while adding aerosol optical depth (A) and wind speed (W) have little effect. Dotted lines show the 95% confidence interval for the cloud response in the model with no controls. The confidence interval is calculated allowing for arbitrary temporal and spatial correlation within adm-1 (e.g. state) units.



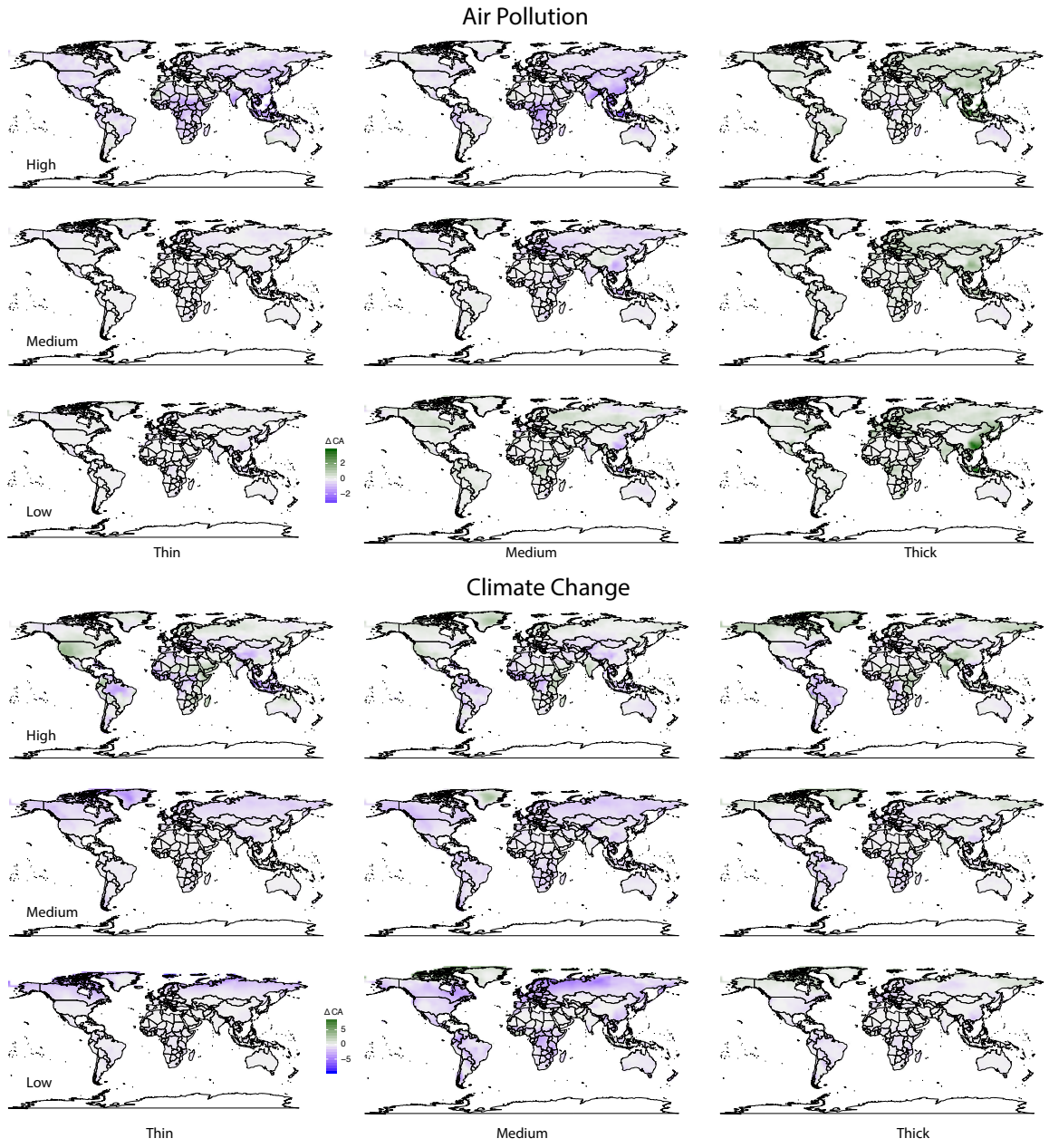
Extended Data Figure 3.5: Robustness of empirical estimates of the sunlight-mediated effect of cloud scattering on crop yield. Each curve shows the estimated effect of increasing the cloud optical depth of cloudy areas from zero to a given value for three hours during the growing season in the pooled sample. In all panels, dotted lines represent the 95% confidence interval for the pooled effect. **a**, Climatic controls in the primary specification ("BASE", black) (Supplementary Information Equation 3.13) are altered to use a different precipitation dataset (red), model temperature non-parametrically using bins (green), model temperature using degree days calculated from a sinusoidal interpolation of daily maximum and minimum temperature (yellow), control for nighttime cloud cover (grey) and for vapor pressure deficit (blue) (Supplementary Information, section III.4). **b**, The fixed-effects in the primary specification (black) are changed from administrative unit level-2 specific quadratic trends, to administrative unit level-2 cubic trends (yellow) and administrative unit level-1 quadratic trends (blue). **c**, The functional form used to calculate the cloud response is altered from the preferred specification which uses a restricted cubic spline with four knots (black) to a restricted cubic spline using 3 knots (blue) and 5 knots (yellow) as well as a cubic polynomial (green), and non-parametric bins of optical depth (red). **d**, The observational weights are changed from standard OLS weights to planted-area weights (red).



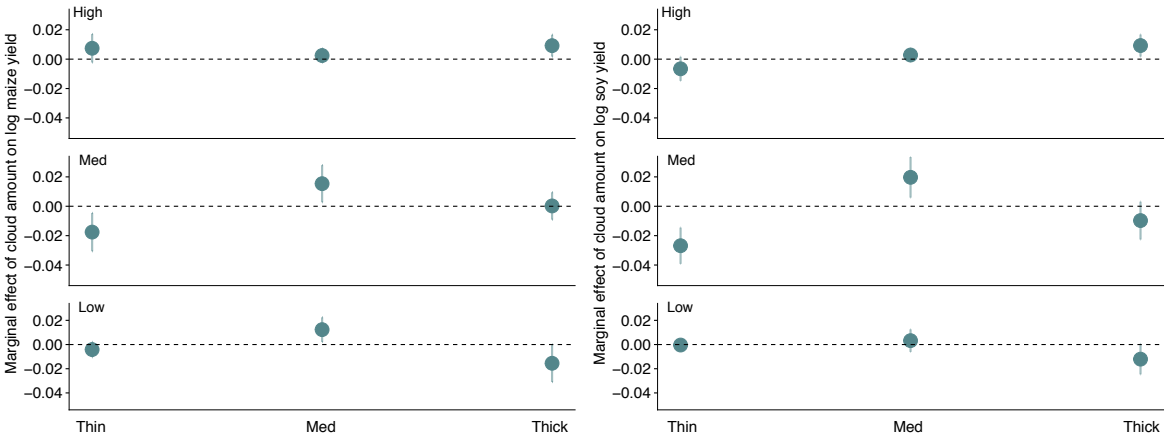
Extended Data Figure 3.6: Estimated climate control functions from the model of cloud scattering impacts on maize and soy yield. Regional climate response functions from the pooled model (Supplementary Information Equation 3.14) for maize (**a**) and soy (**b**). Responses show the effect of changing the entire growing season precipitation, temperature, wind speed and aerosol optical depth on growing season log yield. Daily effects can be calculated by dividing the growing season response by the number of days within the growing season (roughly 150 days, depending on the region).



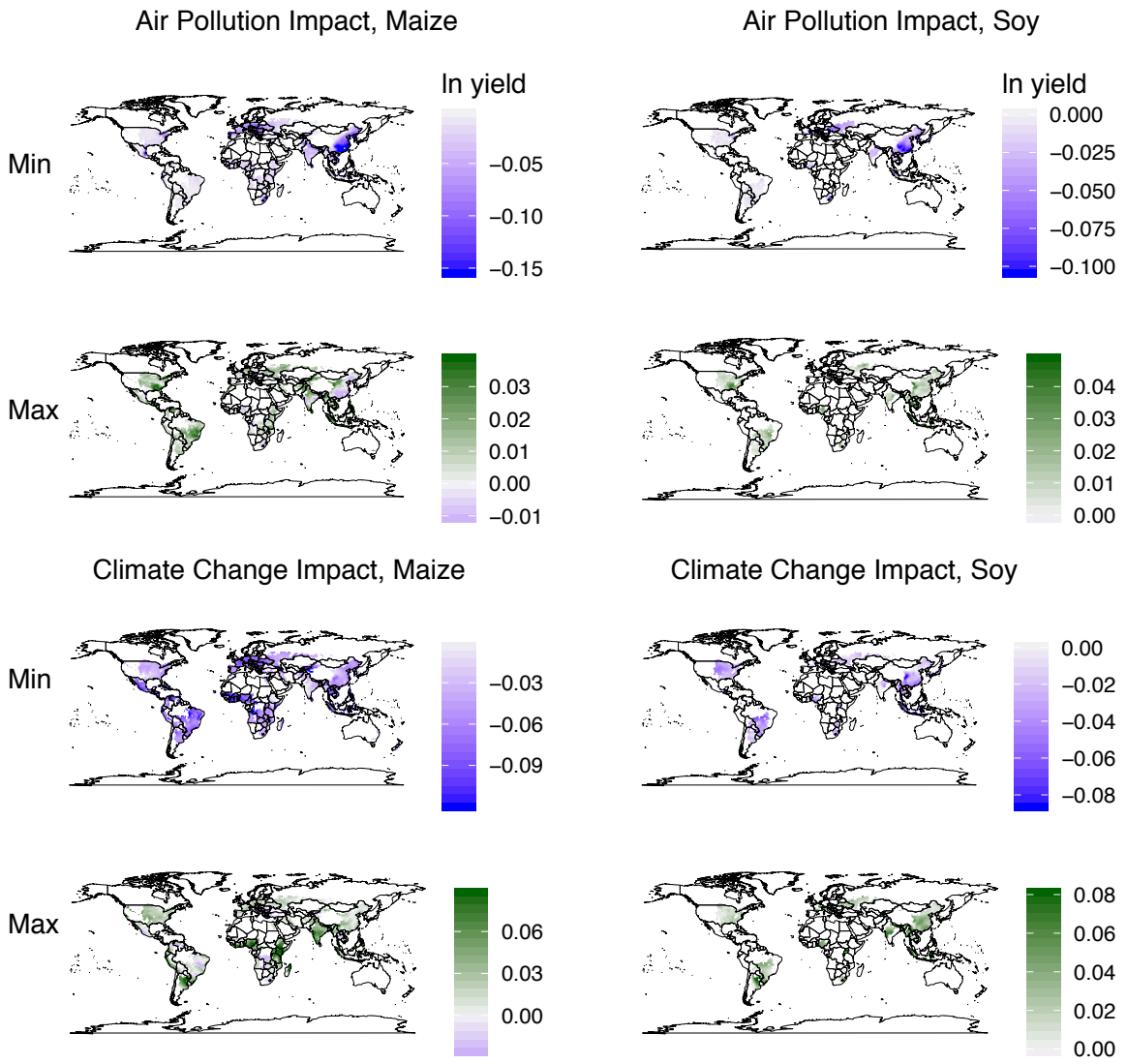
Extended Data Figure 3.7: Empirical estimates of the effect of changing the diffuse fraction on the response of maize and soy yields to photosynthetically active radiation. The estimated effect of changing photosynthetically active radiation (PAR) on maize and soy yields for an hour during the growing season in the pooled sample (black), Brazil (green), China (red), the European Union (yellow), and the United States (blue). The effect of changing PAR is evaluated at a diffuse fraction of 0 (**a**), the average diffuse fraction of 0.4 (**b**) and a diffuse fraction of 1 (c) (Supplementary Information Section IV). Dotted lines represent the 95% confidence interval for the pooled effect, which is calculated allowing for arbitrary temporal and spatial correlation within adm-1 (e.g. state) units.



Extended Data Figure 3.8: Changes in cloud distributions due to anthropogenic aerosol and climate change emissions. Simulated changes in cloud amount (CA) for nine cloud types in the air pollution and climate changes scenarios. Growing season changes are averages over 5 climate models and 30 years of data (Supplementary Information Section V).



Extended Data Figure 3.9: Marginal impact of cloud optical scattering by cloud type on maize and soy yields. Circles show the estimated marginal impact of growing season cloud amount on yield for nine cloud types in the pooled sample (Supplementary Information Section III.7) used to project the sunlight-mediated impact of anthropogenic changes in cloud distributions on yields. For maize, the model R2 is 0.99 and the projected R2 is 0.07. For soy, the R2 is 0.99 and the projected R2 is 0.14. Whiskers show the 95% confidence interval, which is calculated allowing for arbitrary temporal and spatial correlation within adm-1 (e.g. state) units.



Extended Data Figure 3.10: Model spread of the projected effect of air pollution and climate change on crop yields due to cloud-induced changes in sunlight. Maps show the pixel-wise minium and maximum impacts projected by the 5 climate models for maize and soy yields.

Supplementary Information

I Data

Insolation data The World Radation Data Centre (WRDC) provides daily measurements of total, direct and diffuse insolation from 1953-2013 [129]. The data is cleaned as in [84]. The sample has 3,428,474 station-day observations of total insolation and 928,202 station-day observations of diffuse light.

Yield data I use subnational maize and soy yields from the United States, Brazil, China and the European Union in the analysis. I use data from 1985 - 2009 because yields recorded in 1984 have growing seasons that include data from 1983 which is partially missing in the cloud data, which begins July 1, 1983. I drop any administrative unit-year observation that is has missing data, that is in the top or bottom 1% of the national yield to remove outliers, and that is missing over 5% of within growing season cloud or temperature measurements to reduce measurement error. Subsequently, I drop any administrative unit which has fewer than 13 (i.e. half) remaining observations to increase the balance of the final panel. I calculate yield as $\frac{\text{production}}{\text{area planted}}$ in all regions other than the EU, where only harvested area is available and thus I calculate yield as $\frac{\text{production}}{\text{area harvested}}$. The resulting panel has 93,468/61,480 municipio-year obsevations from Brazil, 43,484/34,544 county-year observations from the United States, 27,451/NA county-year observations from China, and 2,248/703 NUTS1 and NUTS2 regions from Europe for maize/soy, respectively.

Cloud data Data on cloud amount, cloud optical depth and cloud type are from the International Satellite Cloud Climatology Project (ISCCP) Climate Data Record, H-Series (HGG). These data have global 3-hourly coverage, from July 1983 to December 2009 on a 1-degree equal area grid.[1],[2] [98]. The ISCCP data were selected for the length of the series (25+ years), the high temporal resolution, and the ability to match consistently with output of global climate models through ISCCP simulators (Supplementary Information Section V).

From the ISCCP data I used data on cloud amount, cloud optical depth of cloudy pixels, and cloud amount by cloud type. Details on how these data were processed from raw geostationary and polar orbiting satellites with a native resolution of around 0.1 degree and measurements of surface conditions can be found in the Climate Data Record Program's Climate Algorithm Theoretical Basis Document for the International Satellite Cloud Climatology Project, Cloud Properties, H-Series.[3]

Primary specification controls The Berkeley Earth Daily Land gridded dataset gives daily average surface temperature values from 1880 to present at a resolution of 1-degree [1]. The University of Delaware Air Temperature and Precipitation dataset provides monthly total precipitation over land (mm/month) from 1900 to 2010 at 0.5-degree resolution [127]. The Modern-Era Retrospective Analysis for Research and Applica-

[1]Data available at: https://www.ncei.noaa.gov/data/international-satellite-cloud-climate-project-sccp-h-series-data/access/isccp-basic/hgg/.

[2]Data extended to June 2015 as of December 2018; these data could be included in future analyses.

[3]https://www1.ncdc.noaa.gov/pub/data/sds/cdr/CDRs/Cloud_Properties-ISCCP/AlgorithmDescription_01B-29.pdf

tions, version 2 (MERRA-2) gives hourly data on total column aerosol optical depth (TOTEXTTAU) and hourly maximum wind speed (SPEEDMAX) at 0.625 degrees longitude by 0.5 degrees latitude resolution [30].

Photosynthetically Active Radiation MERRA-2 gives hourly measurements of total (PARTOT), direct (PARDR) and diffuse (PARDF) photosynthetically active insolation [30].

Additional controls for robustness tests The Berkeley Earth Daily Land gridded dataset gives daily maximum and minimum temperature, which are used to calculate vapor pressure deficit as well as hourly growing degree days and killing degree days (Supplementary Information Section III.4). The Global Meteorological Forcing Dataset for land surface modeling gives monthly precipitation from 1901-2012 at 0.5 degree resolution [105].

Growing season and cropped fraction data Growing season planting and harvesting dates are from [100]; cropped fraction data (crop-specific) are from [67].

II Estimating the effect of cloud scattering on sunlight

To estimate the impact of cloud scattering on total (I^T), direct (I^D), and diffuse (I^F) insolation I follow [84] and pair the 3-hourly (h) cloud amount (CA) and cloud optical depth τ data with daily WRDC insolation data and estimate:

$$ln(I^X_{id}) = f_{C^{I^X}}(C_{id}) + \epsilon_{id} = \psi_1 \sum_{h \in \text{day}} \frac{\text{CA}_{ih} r_1(\tau_{ih})}{\#h \in \text{day}} + \psi_2 \sum_{h \in \text{day}} \frac{\text{CA}_{ih} r_2(\tau_{ih})}{\#h \in \text{day}} + \psi_3 \sum_{h \in \text{day}} \frac{\text{CA}_{ih} r_3(\tau_{ih})}{\#h \in \text{day}} + \phi_{ij} + \epsilon_{id} \tag{3.1}$$

Where d is day of sample, h is 3-hour of sample, j is day of year, day is daytime (i.e. positive insolation), and ϵ is an error term. That is, I regress daily insolation on the daytime average restricted cubic spline expansions of cloud optical depth to estimate ψ_1, ψ_2, and ψ_3, which paramatarize the response of insolation to clouds, $f_{C^{I^X}}$. $r_1(.)$, $r_2(.)$ and $r_3(.)$ are restricted cubic spline feature expansions [41] with knots at $\tau_{ih} = 0,\ 15,\ 30,\ 60$. Impacts on total ($I^T$), direct ($I^D$), and diffuse ($I^F$) insolation are estimated separately. Station-by-day-of-year fixed effects, ϕ_{id}, account for all time-invariant differences between stations, such as latitude, as well as for all location-specific seasonal patterns such as summer dust or smoke. $f_{C^{I^X}}$ is identified by comparing insolation measurements and cloudiness across years within the same day-of-year. For example, diffuse light and cloudiness on November 26, 1991 is compared to diffuse light and cloudiness November 26, 1992. Standard errors are computed allowing for arbitrary patterns of serial correlation over time between all stations in the same country and for arbitrary patterns of autocorrelation across all station observations within the same year [46].

III Estimating the sunlight-mediated effect of cloud scattering on crop yield

The primary goal of this analysis is to estimate the sunlight-mediated effect of cloud scattering on yield. In this section I describe how I estimate this effect.

III.1 Empirical framework

Consider yield Y in administrative unit i and year t as a function of temperature (T), precipitation (P), sunlight (I), air pollutants (A) and other variables (U):

$$Y_{it} = Y(T_{it}, P_{it}, I_{it}, A_{it}, U_{it}) \tag{3.2}$$

A challenge in this context is that these inputs are themselves functions of Clouds (C) and other factors (Z):

$$\text{Temp}_{it} = T(C_{it}, Z_{it}) \tag{3.3}$$
$$\text{Precip}_{it} = P(C_{it}, Z_{it}) \tag{3.4}$$
$$\text{Insolation}_{it} = I(C_{it}, A_{it}, Z_{it}) \tag{3.5}$$
$$\text{Air Pollution}_{it} = A(C_{it}, Z_{it}) \tag{3.6}$$
$$\text{Wind Speed}_{it} = W(C_{it}, Z_{it}) \tag{3.7}$$
$$\text{Unobservable}_{it} = U(C_{it}, Z_{it}) \tag{3.8}$$

Further, clouds themselves are dependent on temperature and air pollution [11]: $C_{it} = C(T_{it}, A_{it}, Z_{it})$.

Differentiating yield with respect to C, following [84], decomposes the effects of clouds into a temperature term, a precipitation term, an insolation, and an air pollution term:

$$\begin{aligned}\frac{dY}{dC} = &\underbrace{\frac{\partial Y}{\partial T}\frac{\partial T}{\partial C}}_{\text{temperature effect}} + \underbrace{\frac{\partial Y}{\partial P}\frac{\partial P}{\partial C}}_{\text{precipitation effect}} + \underbrace{\frac{\partial Y}{\partial I}\frac{\partial I}{\partial C}}_{\text{insolation effect}} + \underbrace{\frac{\partial Y}{\partial A}\frac{\partial A}{\partial C}}_{\text{air pollution effect}} \\ &+ \underbrace{\frac{\partial Y}{\partial W}\frac{\partial W}{\partial C}}_{\text{wind speed effect}} + \underbrace{\frac{\partial Y}{\partial U}\frac{\partial U}{\partial C}}_{\text{other effect}}\end{aligned} \tag{3.9}$$

The insolation-mediated effect, $(\frac{\partial Y}{\partial I}\frac{\partial I}{\partial C})$ is difficult to recover directly because we do not observe the change in yields due directly to changes in insolation. Rather, we observe only the changes in these variables over time. To see how we can use these, we take the derivative of Y with respect to time:

$$\frac{dY}{dt} = \frac{\partial Y}{\partial T}\frac{dT}{dt} + \frac{\partial Y}{\partial P}\frac{dP}{dt} + \frac{\partial Y}{\partial I}\frac{dI}{dt} + \frac{\partial Y}{\partial A}\frac{dA}{dt} + \frac{\partial Y}{\partial W}\frac{dW}{dt} + \frac{\partial Y}{\partial U}\frac{dU}{dt} \tag{3.10}$$

Then, taking the derivatives of the variables we don't observe directly (equations 3.5,3.8) with respect to t and substituting into equation 3.10, we get:

$$\begin{aligned}\frac{dY}{dt} = &\frac{\partial Y}{\partial T}\frac{dT}{dt} + \frac{\partial Y}{\partial P}\frac{dP}{dt} + \frac{\partial Y}{\partial A}\frac{dA}{dt} + \frac{\partial Y}{\partial W}\frac{dW}{dt} \\ &+ \frac{\partial Y}{\partial I}\underbrace{\left[\frac{\partial I}{\partial C}\left[\frac{\partial C}{\partial T}\frac{dT}{dt} + \frac{\partial C}{\partial A}\frac{dA}{dt} + \frac{\partial C}{\partial Z}\frac{dZ}{dt}\right] + \frac{\partial I}{\partial A}\frac{dA}{dt} + \frac{\partial I}{\partial Z}\frac{dZ}{dt}\right]}_{\frac{dI}{dt}} \\ &+ \frac{\partial Y}{\partial U}\underbrace{\left[\frac{\partial U}{\partial C}\frac{dC}{dt} + \frac{\partial U}{\partial Z}\frac{dZ}{dt}\right]}_{\frac{dU}{dt}}\end{aligned} \tag{3.11}$$

Variables whose changes over time we observe are left as is because they can be directly accounted for in the empirical model. Re-arranging and simplifying terms gives:

$$\begin{aligned}\frac{dY}{dt} = &\left[\underbrace{\frac{\partial Y}{\partial T}}_{\frac{\partial}{\partial T}f_T(.)} + \frac{\partial Y}{\partial I}\frac{\partial I}{\partial C}\frac{\partial C}{\partial T}\right]\mathbf{\frac{dT}{dt}} + \underbrace{\frac{\partial Y}{\partial P}}_{\frac{\partial}{\partial P}f_P(.)}\mathbf{\frac{dP}{dt}} + \underbrace{\frac{\partial Y}{\partial W}}_{\frac{\partial}{\partial W}f_W(.)}\mathbf{\frac{dW}{dt}} \\ &+ \left[\underbrace{\frac{\partial Y}{\partial A} + \frac{\partial I}{\partial A}}_{\frac{\partial}{\partial A}f_A(.)} + \frac{\partial Y}{\partial I}\frac{\partial I}{\partial C}\frac{\partial C}{\partial A}\right]\mathbf{\frac{dA}{dt}} \\ &+ \underbrace{\left[\frac{\partial Y}{\partial I}\frac{\partial I}{\partial C} + \frac{\partial Y}{\partial U}\frac{\partial U}{\partial C}\right]}_{\frac{\partial}{\partial I}f_C(.)}\underbrace{\mathbf{\frac{\partial C}{\partial Z}\frac{dZ}{dt}}}_{\text{exogeneous cloud variation}} \\ &+ \underbrace{\left[\frac{\partial I}{\partial Z} + \frac{\partial Y}{\partial U}\frac{\partial U}{\partial Z}\right]\frac{dZ}{dt}}_{\epsilon}\end{aligned} \tag{3.12}$$

which leads directly to the structure of the empirical model, whose terms are shown in bold.

Note that cross-variable terms such as the effect of temperature on yields mediated by cloud-induced changes in sunlight ($\frac{\partial Y}{\partial I}\frac{\partial I}{\partial C}\frac{\partial C}{\partial T}$) will not contribute to the estimation of either $f_T(.)$ or $f_C(.)$ because when estimating each, the effect of the other variable is projected out, as can be seen by the Frisch–Waugh–Lovell theorem. Thus, when estimating the impact of clouds all variation due to temperature is projected out, and when estimating the impact of temperature all the variation due to changes in clouds is projected out; in practice both are estimated simultaneously in a single multivariate regression, as described below.

III.2 Empirical model

Following equation 3.12 I recover the insolation-mediated effect of clouds on yield ($\frac{\partial Y}{\partial I}\frac{\partial I}{\partial C}$) in administrative unit i and year t by fitting:

$$Y_{it} = f_T(T_{it}) + f_P(P_{it}) + f_A(A_{it}) + f_W(W_{it}) + f_C(C_{it}) + \lambda_i + \phi_{i1}t + \phi_{i2}t^2 + \epsilon_{it} \quad (3.13)$$

Here, $f_C(C_{it})$ is the sunlight-mediated effect of clouds on yields, which includes the effects of decreasing total insolation, increasing (or decreasing) diffuse insolation, and inducing changes in the relative intensities of insoaltion across the solar spectrum. It also includes any residual impacts of clouds on yields not mediated through or correlated with T, P, A, or W (i.e. $\frac{\partial Y}{\partial U}\frac{\partial U}{\partial C}$), if they exist. Y_{it} are log yields of maize or soy (each crop estimated separately). $f_T(.)$ and $f_P(.)$ are restricted cubic splines of daily average temperature and monthly precipitation and $f_A(.)$ and $f_W(.)$ are cubic polynomials of hourly daytime aerosol optical depth and hourly maximum wind speed, which control flexibly for these variables [46]. The administrative-unit-specific fixed effects, λ_i, control for any time-invariant differences between administrative units such as soil type, and the unit-specific quadratic time trends, $\phi_{i1}t + \phi_{i2}t^2$, control for unit-specific trending variables such as tchnological adoption or income. The residual, ϵ_{it}, captures all factors that vary within administrative units over time but are uncorrelated with cloud scattering as well as the other factors in the model, such as the price of fertilizer. I compute standard errors clustering by administrative level-1 unit (e.g. state), which allows for arbitrary patterns of correlation between all observations from administrative level-2 units (e.g. counties) within each administrative level-1 unit. This allows for both spatial and temporal correlation of errors [46].

All models using the pooled sample estimate a single cloud scattering effect across regions and allow the effect of climate controls to vary in each region, r:

$$Y_{it} = f^r_T(T_{it}) + f^r_P(P_{it}) + f^r_A(A_{it}) + f^r_W(W_{it}) + f_C(C_{it}) + \lambda_i + \phi_{i1}t + \phi_{i2}t^2 + \epsilon_{it} \quad (3.14)$$

When plotting the estimated response of yields to cloud scattering, $f_C(C_{it})$, we assume a cloud fraction of 0.7, roughly the global average. Plotting the response for a different assumed cloud amount is a simple linear re-scaling of the curve. To plot the response to changes in optical depth for a fully cloudy day (a cloud fraction of 1), for example, one would simply multiply the shown curves by $\frac{1}{0.7}$, or roughly 1.4.

III.3 Identification strategy

A key challenge to identifying the sunlight-mediated effect of cloud scattering on yields is that clouds can both affect and be affected by variables that impact yield including temperature, precipitation, aerosols and wind speed. The model accounts for these potentially confounding variables by directly observing them and flexibly controlling for them in the model. Similarly, the model's unit-specific fixed effects remove any time-invariant differences between observations, and thus account for all potentially confounding correlations between climatological cloudiness and average yields. The

model's unit-specific time trends account for any potentially confounding correlation between tends in yields and trends in cloudiness, such as those potentially due to industrialization.

Thus, to identify the sunlight-mediated effect of cloud scattering on yields I examine how yields within an admistriative unit vary year-to-year with varying cloud cover after accounting for changes in yields and clouds due to temperature, precipitation, wind speed, aerosols, and unit-specific fixed effects and time trends. The identifying assumption is that the variation in yields (after partitioning out variation due to fixed effect and controls) that is correlated with variation in cloudiness (similarly conditioned – $\frac{\partial C}{\partial Z}\frac{dZ}{dt}$) is due to changes in sunlight (i.e. $C_{it} \perp\!\!\!\perp \epsilon_{it} | f_T(T_{it}), f_P(P_{it}), f_A(A_{it}), f_W(W_{it}), \lambda_i, \phi_{i1}t, \phi_{i2}t^2$). Put another way, I assume that clouds impact yield only through changes in sunlight, after potential cloud impacts due to temperature, precipitation, air pollution and wind speed have been accounted for, and that there are no variables omitted from the model that are correlated with both clouds and yield.

III.4 Robustness of the sunlight-mediated impact of clouds on yields

To test the robustness of the estimated sunlight-mediated impact of clouds on yields I re-estimate the model (equation 3.13) using alternative and additional climate controls, functional forms for $f_C(.)$, fixed effects, and observation weights.

Climate controls: To test for potential bias due to mis-measurement of precipitation I replace the station-based precipiation dataset with a reanalysis product, the Global Meteorological Forcing Dataset for land surface modeling [105]. To test for potential miss-specification of temperature I estimate the effect replacing the restricted cubic spline temperature controls with a flexible non-parametric function using 1-degree bins of temperature [46] as well as a "degree-day" piecewise linear function of hourly temperature (from sinusoidal interpolation of daily maximum and minimum temperature) [104]. To test for potential confounding from relative humidity and, in turn, water stress I additionally control for vapor pressure deficit [90]. To test for whether the effect is driven by factors correlated with the creation of clouds and not already accounted for by the controls in the model I add a cubic in nighttime cloud optical depth into the model. The model is robust to each of these changes other than the addition of nighttime cloud optical depth (Extended Data Fig. 3.5). Including nighttime optical depth into the model preserves the non-linear shape of the response but attenuates the benefits of cloudiness, which may be due to either the existence of confounding variables or to attenuation bias from removing a substantial amount of the identifying variation in daytime cloudiness.

Functional Form: To test sensitivity to the functional form of $F_C(.)$, I re-estimate the model using a flexible non-parametric "binned" function of optical depth (Supplementary Information Section III.5), restricted cubic splines of optical depth using three and five knots (the base specification has four knots), and a cubic polynomial in optical depth. All of these models find a concave response to cloud scattering though the optimal amount of scattering and the magnitude of the benefits and damages from low and high amounts of scattering differ.

Fixed Effects: To test the sensitivity to model specification, I re-estimate the

model using a set of less flexible (quadratic administrative level-1 time trends) and more flexible (cubic administrative level-2 time trends) non-parametric controls, and find similar responses.

Weighting: To test the sensitivity of the model to weighting, I weight each observation by planted area, and find similar results.[4]

III.5 Derivation of the estimating equation for the non-linear impact of cloud optical depth

Here, I derive and interpret the estimating equation for the non-linear response of annual yield to sub-daily changes in cloud optical depth. Following the models in [104, 10] for temperature impacts, I assume that in each county i, log productivity is some function of the average cloud optical depth, τ_{ih} during the 3-hour period h.

$$ln(y_{ih}) = f_C(\tau_{ih}) + \epsilon_{ih} \tag{3.15}$$

In this paper, I use flexible linear models, $\sum_{k=1}^{K} \theta_k(\tau_{ih})$, to estimate f – the non-linear effect of cloud optical depth on productivity:

$$ln(y_{ih}) = \sum_{k=1}^{K} \theta_k(\tau_{ih}) + \epsilon_{ih} \tag{3.16}$$

For the simplicity of exposition, here I use a second order polynomial, though I use restricted cubic splines and other flexible functional forms in the analysis.

$$ln(y_{ih}) = \beta_1 \tau_{ih} + \beta_2 \tau_{ih}^2 + \epsilon_{ih} \tag{3.17}$$

Summing over the growing season, t, I get:

$$\sum_{h \in t} ln(y_{ih}) = \beta_1 \sum_{h \in t} \tau_{ih} + \beta_2 \sum_{h \in t} \tau_{ih}^2 + \sum_{h \in t} \epsilon_{ih} \tag{3.18}$$

Or equivalently:

$$ln(\prod_{h \in t} y_{ih}) = \beta_1 \sum_{h \in t} \tau_{ih} + \beta_2 \sum_{h \in t} \tau_{ih}^2 + \sum_{h \in t} \epsilon_{ih} \tag{3.19}$$

Following [104, 9], we assume that the growing season yield, Y_{it} is the product of withing-growing season productivity, $\prod_{h \in t} y_{ih} = Y_{it}$. Put another way, we assume plant growth compounds over the growing season. This gives:

[4] For samples within the EU, observations are weighted by *harvested* area because planted area data is unavailable.

$$ln(Y_{it}) = \beta_1 \sum_{h \in t} \tau_{ih} + \beta_2 \sum_{h \in t} \tau_{ih}^2 + \sum_{h \in t} \epsilon_{ih} \tag{3.20}$$

or equivalently:

$$ln(Y_{it}) = \frac{\beta_1}{\#_{h \in t}} \tau_{it} + \frac{\beta_2}{\#_{h \in t}} \tau_{it}^2 + \epsilon_{it} \tag{3.21}$$

which is our estimating equation. $\#_{h \in t}$ is the number of 3-hour daytime observations of clouds during the growing season, τ_{it} is the growing season average optical depth, and τ_{it}^2 is the growing season average of squared 3-hourly optical depth.

In practice, we measure τ_{ih}, the optical depth of a county during a 3-hour period, as the cropped-area-weighted average of ISCCP-pixel-average optical depth values. Similarly, we measure τ_{ih}^2, the squared optical depth of a county during a 3-hour period, as the cropped-area-weighted average of ISCCP-pixel-average *squared* optical depth values. ISCCP pixels are roughly 111km by 111km, but measure optical depth at the 10km by 10km level before averaging over cloudy pixels, so our analysis closely approximates the average squared optical depth at the 10km resolution.[5]

ISCCP provides measurements of cloud amount and cloud optical depth for cloudy pixels. We calculate the average cloud optical depth τ_{it}^2 by noting that it is equivalent to the cloud fraction times the optical depth of cloudy pixels.

III.6 Functional Forms for $f_C(\tau)$

To test robustness of the main result, I estimate $f_C(\tau)$ using a variety of specifications (Extended Data Fig. 3.5). Each of these is estimated by computing a different set of non-linear transformations of τ_{ih} – i.e. $\sum_{k=1}^{K} \theta_k(\tau_{ih})$ – and then learning an approximation of $f_C(\tau)$ using linear regression.

To estimate a model using restricted cubic splines I calculate $\sum_{k=1}^{K} \theta_k(\tau_{ih})$ using the rcspline.eval function from the R package Hmisc V4.2-0 and as described in [41]. I used four knots placed at τ_{ih} = 0,15,30,60 chosen to span the distribution of τ_{ih} during the growing season.

To estimate a binned model, I calculate $\sum_{k=1}^{K} \theta_k(\tau_{ih}) = \sum_{k=1}^{K} \mathbb{1}(\tau_{ih} \in B_k)$ for bins B_k in (0,2], (2,8], (8,14], (14,20], (20,30], (30,50], (50,100], (100, ∞), where $\mathbb{1}$ is the indicator function. Bins were chosen to span the distribution of τ_{ih} during the growing season.

To estimate a model using a cubic polynomial I calculate $\sum_{k=1}^{K} \theta_k(\tau_{ih}) = \sum_{k=1}^{3} \tau_{ih}^k$.

[5] If all 10km by 10km *cloudy* pixels within the 111km by 111km pixel had the same optical depth then the approximation would be exact. Heterogeneity of cloud optical depth within cloudy pixels will lead to measurement error, which could attenuate the estimated model coefficients.

III.7 Cloud type model for $f_C(.)$

As a complement to paramatarizing the impact of cloud cover as a non-linear function of cloud optical depth, I estimate a model paramatarizing cloud optical scattering impacts as a function of cloud type. Cloud types in the ISCCP data are defined by thresholds of cloud height and cloud optical depth. Thus, paramatarizing clouds by cloud type is essentially a flexible non-parametric binned model of cloud optical depth that allows for clouds of different heights to have different effects. Cloud height could affect cloud optical impacts because clouds of different heights – and thus different temperatures – may have different optical properties.

In this model, I paramatarize $f_C(C_it)$ as $\sum_{v=1}^{9} \beta_v CA_v$ where CA is the cloud amount for each cloud type v. The 9 cloud types are defined by the cross of three cloud optical depth bins: (0-3.6],(3.6,23],(23-∞) and three cloud top heights (measured as pressure in mb) bins (1000, 680], (680, 440], (440, 0].

This paramatarization, though more difficult to interpret, has similar predictive skill as the model using the optical depth paramatarization (Extended Data Fig. 3.9).

III.8 Comparison to the ideal cloud distribution

To calculate the suppression of yields due to the present climate relative to an optically ideal climate for each crop, I first estimate the contribution of cloud optical scattering to global actual and ideal yields using the trained statistical yield model (Supplementary Information Eqn. 3.13). Actual yields are evaluated using observed cloud optical depths from 1984-2009, and the ideal climate assumes the empirically estimated optimal optical depth of 15 during the entire growing season. I then subtract the ideal from the actual, calculate a global cropped-fraction weighted average and convert to percent to arrive at the yield supression due to the actual global optical scattering environment.

IV Estimating the strength of the diffuse fertilization effect and response to insolation for crop yield.

Cloud scattering changes the total amount of insolation, the fraction of that insolation that is diffuse, as well as the distribution of insolation across the solar spectrum. Flexibly estimating the impact of optical scattering on yields captures all of these effects as a function of optical depth. Here, I specify and estimate a functional form for how total light and the diffuse fraction impact yields to learn the potential mechanisms that are driving the impact of optical scattering on yields.

A common way to model the diffuse fertilization effect is to allow a higher diffuse fraction to increase the radiation use efficiency of crops [102]. I model the impact of photosynthetically active radiation (PAR, or R) on crops by allowing the effect of R to flexibly impact yield, and allowing the marginal effect of R to vary linearly with the diffuse fraction (DF):

$$Y_{it} = f_T(T_{it}) + f_P(P_{it}) + f_A(A_{it}) + f_W(W_{it}) + f_R(R_{it}) + \lambda_i + \phi_{i1}t + \phi_{i2}t^2 + \epsilon_{it} \quad (3.22)$$

with

$$f_R(R_{it}) = \alpha_1 r_1(R_{it}) + \alpha_2 r_2(R_{it}) + \alpha_3 R_{it}\mathrm{DF}_{it} + \epsilon_{it}$$

where r_1 and r_2 are restricted cubic spline features of hourly insolation with knots at 10, 100, and 325 $\frac{W}{m^2}$.[6] The impact of total PAR on yields is defined by α_1, α_2, and α_3. α_3 describes how the marginal effect of total PAR on yields changes with increasing diffuse fraction. A positive value of α_3 is evidence of the diffuse fertilization effect. Since most of the variation in sunlight is due to changes in clouds, identifying the effect of PAR on yields faces the same challenges as identifying the insolation-mediated impacts of clouds on yields – thus we similarly account for potentially confounding variables as in equation 3.13 when estimating equation 3.22.

IV.1 Validation of MERRA2 PAR using the WRDC station data

To validate the MERRA2 observations of total, direct and diffuse PAR, I compare the MERRA2 reanalysis measurements to station-measurements of shortwave irradiance from the WRDC. I compare the two by regressing the WRDC station measurements on the MERRA2 PAR measurements. I control for station-by-day-of-year fixed effects to account for season-specific differences across locations and to test ability of the MERRA2 measurements to explain day-to-day variation in PAR within a location, which is similar to the variation I use when estimating the effect of insolation on crop yields. I find that the MERRA2 data explain 46%, 16% and 45% of the WRDC measured day-to-day variation in shortwave insolation (Extended Data Table 3.2). I find that each $\frac{Wh}{m^2\mathrm{day}}$ increase in PAR correlates with a 2 $\frac{Wh}{m^2\mathrm{day}}$ increase in shortwave insolation, which is consistent with previous findings that slightly less than half of shortwave solar irradiance is photosynthetically active [77]. The lack of perfect correlation between these two measurements of PAR may be due to differences in attenuation between shortwave irradiance and PAR, to heterogeneity of insolation within the MERRA2 grid cell, or to other sources of error in the MERRA2 or WRDC station measurements.

V Calculating the sunlight-mediated impact of anthropogenic changes in clouds on global maize and soy yields

To place the impact of cloud optical scattering on yields in context, I calculate the sunlight-mediated effects of anthopogenic changes in clouds on yields in two scenarios. First, I calculate the how changes in sunlight due to changes in cloud distributions from anthropogenic aerosols impact yield (the "aerosol-cloud effect"). And second, I calculate how changes in sunlight due to changes in cloud distributions from anthropogenic climate change impact yield (the "carbon-cloud effect").I calculate these effects by combining climate model estimates of how cloud distributions change due to human activity with statistical estimates of how changes in sunlight due to clouds impact yield (Supplementary Information Section III.7 and Extended Data Fig. 3.9). These estimates inform the long-standing question of how observed "global dimming and brightening"

[6] Similarly to the regression features for estimating the effect of cloud optical depth, I calculate the restricted cubic spline features before averaging over the growing season.

due to anthropogenic emissions of air pollutants impacts crop productivity [125, 126]. While these estimates give valuable context for the statistical analyses of cloud impacts on yield, they should not be interpreted as predictions of future impacts because they abstract away from some potentially important factors such as heterogeneity of local effects and adaptation.

To estimate the aerosol-cloud effect, I compare yields in a control world with pre-industrial (1860) aerosol levels (sstClim, following the nomenclature of [112])) to an identical world with aerosol levels set to more modern (2000) levels (sstClimAerosol), following [133]. To estimate the carbon-cloud effect, I compare yields in a control world with pre-industrial CO2 levels (piControl) with an identical world with quadrupled CO2 levels (1pctCO2), following [134].

I calculate the impact of air pollution and climate change on growing season cloud distributions using the output of five climate models participating in phase 5 of the Coupled Model Intercomparison Project (CMIP5): CanESM2, HadGEM2[7], IPSL-CM5A-LR, MIROC5 and MRI-CGCM3. These models were selected because each ran the four experiments described above (i.e. sstClim, sstClimAerosol, piControl, 1pctCO2), implemented the ISCCP simulator to output cloud variables in a manner consistent with the observed ISCCP products [124], was available at `https://esgf-node.llnl.gov/search/cmip5/`, and simulated 30 years of monthly cloud data.

These models do not report cloud optical depth, which prevents direct application of estimates from equation 3.13. Instead, the models report cloud amount by cloud type, with cloud types defined by thresholds of optical depth and cloud height. To link changes in cloud distributions measured in cloud amount by cloud type (Extended Data Fig. 3.8) to changes in yields, I estimate a non-parametric empirical model of yield as a function of cloud amount by cloud type using a form very similar to equation 3.13 (Supplementary Information Section III.7). I then average the cloud data over the growing season and evaluate the statistical crop model at these growing season values to get an estimate of yield for each crop, year and experiment. I then calculate the expected yields in each experiment by averaging yields across years. To calculate the aerosol-cloud effect I subtract expected yields in the sstClim experiment from those in the sstClimAerosol experiment; and to calculate the carbon-cloud effect I subtract expected yields in the piControl experiment from those in the 1pctCO2 experiment.

In this analysis I consider only changes in yields due to cloud-induced changes in sunlight. Estimates of the full impact of anthropogenic aerosol emissions and greenhouse gas emissions on yields would have to consider both the impacts of these emissions on crops through changes in variables other than clouds (such as carbon fertilization) and impacts through changes in clouds due to variables other than sunlight (such as temperature or precipitation).

[7] I use HadGEM2-A and HadGEM2-ES for the aerosol and climate change scenarios, respectively.

Chapter 4

Generalizing Earth observation with satellite imagery and machine learning

Numerous global challenges, such as managing planetary resources, require globally comprehensive observation of many variables simultaneously. Combining satellite imagery with machine learning (SIML) presents an opportunity for assembling such observations [19, 64, 140], but current case-by-case solutions require custom systems, extensive expert knowledge, access to imagery, and major computational resources in order to estimate a single variable (a *task*) using regional or global imagery [40, 78, 50, 52, 91]. Here, we develop a general solution to constructing global observations via SIML, where a single method for transforming satellite imagery is sufficiently descriptive that it should be able to predict nearly any ground-level variables that are recoverable through inspection of a satellite image, including previously unstudied tasks. Our approach is task-independent, allowing centralized computation of features to be executed only once ever per image, then distributed and applied to potentially unlimited future tasks by users who require neither domain expertise nor access to underlying imagery. We demonstrate this generalizability across tasks by constructing high resolution ($\sim$1km$\times$1km) estimates for forest cover, population density, elevation, nighttime lights, household income, total road length, and housing prices across the entire US using exclusively daytime images that are processed only once and in advance. Our system outperforms spatial extrapolation of ground-truth data, especially over large distances, and matches or exceeds performance of a state-of-the-art deep convolutional neural network that is much more costly to implement. Our approach requires only that users download a tabular data set, merge it to geolocated labels, and implement a single regression on a personal computer. We demonstrate that our design scales globally with no alterations and naturally achieves super-resolution, where estimates are more spatially granular than the original labels used for training. Generalization enables democratization of SIML, potentially increasing the pace of planet-scale observation and research, accelerating our understanding of global processes and enabling progress towards tackling planetary challenges.

This chapter is joint work with Esther Rolf, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht and Solomon Hsiang; it has been submitted for publication.

Introduction

Addressing complex global challenges—such as managing global climate changes, population movements, ecosystem transformations, or economic development—requires high-resolution, continuous-time, planet-scale observational systems of many variables simultaneously, including some that are not yet specified. Ground-based monitoring systems are generally prohibitively costly for this purpose, and most remain incomplete. Satellite imagery presents a viable alternative for gathering globally comprehensive data, with over 700 earth observation satellites currently in orbit [117], and application of machine learning could be an effective approach for transforming these vast quantities of unstructured imagery data into structured observations that can support research and decision-making. However, observing a single new variable globally using state-of-the-art SIML currently requires a major research program and is inaccessible to non-specialists, facts that restrict the deployment of SIML to solve global challenges. At present, there exists no unified system that generalizes SIML-based Earth observation such that non-specialists can study unlimited new variables, relevant to their context, with no alteration of method.

Current applications of SIML either *(i)* provide many users with access to a small number of pre-defined output variables or *(ii)* give a limited number of experts tools and data to study individual variables of their choosing. Most research falls in the first category *(i)*, where specialists develop custom methods to measure specific variables—such as forest cover [40], surface water [78], land use [50], poverty rates [52] and population density [91]—and the output of the analysis is then distributed in the form of maps for broader use. This approach enables widespread application of these specific maps, but observing a single global snapshot of any new variable (hereafter, a *task*) demands a major enterprise involving a combination of task-specific domain knowledge, remote sensing and engineering expertise, customization and tuning of sophisticated machine learning architectures, specialized data storage systems, and large computational resources. The resulting high costs of development and deployment limit the scope, frequency, and availability of current SIML-based global observations. Recognizing this bottleneck, efforts in category *(ii)* develop pipelines that accelerate specialized earth observation research, for example by consolidating data-sets and computational resources [34]. These investments have effectively enabled experts participating in category *(i)* to produce custom output faster or at larger spatial scales [54], but do not provide a generalizable system that empowers non-experts to independently and reliably solve arbitrary future tasks.

Here, we are the first to develop, demonstrate, and disseminate a fully general approach that allows almost any individual to deploy SIML rapidly, consistently, and at planet-scale to study essentially any variable detectable from orbit. We demonstrate the simultaneous accessibility and generalizability of Multi-task Observation using Satellite Imagery and Kitchen Sinks ("MOSAIKS") by solving a diverse set of large-scale problems on a personal computer using a single set of features that were computed before the problems were selected. To our knowledge, MOSAIKS is the only unified, rigorously evaluated, publicly available system to reliably and rapidly generalize SIML

at global-scale and high-resolution for new variables without the need for specialized knowledge or computing resources.

Constructing global observations from satellite imagery involves predicting variable y_ℓ at location ℓ using an image $\mathbf{I}_\ell$. For each task, denoted s, it is typically posited that there exists some nonlinear mapping $f(\cdot)$ such that

$$y_\ell^s = f^s(\mathbf{I}_\ell) + \epsilon_\ell^s \tag{4.1}$$

for some reasonably small error ϵ_ℓ^s. Under this assumption, each researcher r tries to find a statistical approximation $f^{s,r}(\cdot)$ from a collection of data using contextual expert knowledge. The challenge is to find a set of candidate functions such that approximating Eq. (4.1) is tractable and prediction quality remains high on newly collected data. Because such problems are difficult, to date, efforts have focused on crafting specific features that serve as basis functions to approximate f^s for a particular task. For example, a researcher studying forests might construct pixel-level features based on prior knowledge of plant physiology or they might use an algorithmic approach, such as a neural-network, to search for a set of features that predict forest conditions. However, we hypothesize that there exists a single set of finite and computable basis functions that is sufficiently general to approximate nearly all well-posed f^s and can be easily applied to global-scale datasets. Finding such a general basis for imagery would allow Eq. (4.1) to be reliably solved non-parametrically, regardless of the task. Under such a general basis transformation, minimizing squared error in Eq. (4.1) using real data becomes a linear regression problem.

We propose an approximate basis of convolutional random kitchen sinks [86, 55] (see Methods and Supplementary Information Section II.3), motivated by their strong theoretical foundation [87] and prior performance encoding genetic sequences [68], classifying photographs [21], and predicting solar flares [55]. Our approach is label-independent and has useful properties for analyzing satellite imagery, such as capturing spatial structure of objects while being invariant to their translation. Furthermore, with a sufficiently large approximate basis and enough data, this approach can theoretically describe any well-behaved f^s in Eq. (4.1) [87] (see Methods). We show that, in practice, despite an implementation with a restricted basis and finite samples, this approach is nonetheless highly effective for approximating Eq. (4.1) across diverse tasks when applied to real imagery. Conceptually, rather than evaluating the content of an image by computing features at test time and applying them to Eq (4.1), as is the task-by-task solution, MOSAIKS "pre-organizes" all images in the sample according to many dimensions of their content (i.e. lifting images to a rich feature space) and then quickly learns and exploits only those dimensions that are useful for a specific task when presented with a novel set of labels (Figure 4.1A).

MOSAIKS provides a general solution to Earth observation problems with the form of Eq. (4.1) using a single, centralized, unsupervised, featurization combined with unlimited, decentralized, supervised, linear regressions that use these features to solve for each task of interest. Specifically, MOSAIKS transforms the daytime image $\mathbf{I}_\ell$ into a single vector of random features $\mathbf{x}(\mathbf{I}_\ell)$. Researcher r then merges their own limited

sample of training labels $y_\ell^{s,r}$ to these tabular data and solves the linear regression

$$y_\ell^{s,r} = \mathbf{x}(\mathbf{I}_\ell)\beta^{s,r} + \epsilon_\ell^{s,r}\,, \tag{4.2}$$

then uses the weights $\hat{\beta}^{s,r}$ to estimate $\hat{y}^{s,r}$ via Eq. (4.2) in new locations where imagery is available but ground truth $y^{s,r}$ is not (Fig. 4.1B, Methods, and Fig. S4). Raw imagery is never analyzed by users. Crucially, the featurization $\mathbf{x}(\cdot)$ is rich and highly descriptive, encoding enough image information such that future researchers can solve Eq. (4.2) for tasks that are unknown at the time of encoding (see Methods). Thus, MOSAIKS transforms the costly problem of finding case-by-case solutions to Eq. (4.1) to the reduced problem of solving a single, pre-determined, linear regression (Eq. (4.2)).

The regression associated with Eq. (4.2) is simple and fast (Methods, Supplementary Information Section II.4), with sampling uncertainty that can be efficiently estimated. Each task studied here is solved using two commands (a merge and a ridge regression) that take minutes to execute on a personal computer (Supplementary Information Section III.2), achieving performance better than a state-of-the-art deep convolutional neural network trained on a high-performance computing cluster (Supplementary Information Section III.1).

Results

We first demonstrate that MOSAIKS generalizes at scale by applying it to many tasks (Fig. 4.2) across the continental United States (US). This allows systematic evaluation of performance in a data-rich environment before extending estimates globally, where ground-truth may be unavailable or unreliable. Using only a single matrix of features $\mathbf{X}$ derived from $\sim$1km$\times$1km (256-by-256 pixels) daytime images spanning the US, we are able to estimate ground-level forest cover ($R^2 = 0.91$), elevation ($R^2 = 0.68$), population density ($R^2 = 0.72$), nighttime lights ($R^2 = 0.84$), income ($R^2 = 0.45$), road length ($R^2 = 0.53$), and house price ($R^2 = 0.47$) in a holdout test sample (Fig. 4.2, Supplementary Information Section II.6 and Table S2). Solving Eq. (4.2) for each task took 7.5 minutes to compute on ten cores (Intel Xeon CPU E5-2630). These results represent all tasks attempted to date (see Methods & Supplementary Information Section III.2). We compare MOSAIKS's performance to a leading image analysis method - the ResNet-18 CNN architecture - using identical imagery and labels, and we find that MOSAIKS provides predictive skill exceeding this computationally expensive alternative for each task (Supplementary Information Section III.1 and Fig. S12). MOSAIKS also outperforms ridge regression models using features extracted from a pre-trained CNN (Supplementary Information Section III.1), an unsupervised featurization technique that often outperforms other unsupervised approaches [17]. These results indicate that MOSAIKS is skillful for a diverse range of possible applications without changing the procedure or features and without task-specific expertise. These results indicate the range of performance that might be reasonably expected in new tasks.

Note that some patterns of variation are not recovered by MOSAIKS, consistent with the hypothesis that some prediction errors are irreducible if key factors are fundamentally impossible to observe from satellite imagery. For example, extremely high

elevations (>3,000m) are not reliably distinguished from high elevations (2,400-3,000m) that appear visually similar; and roughly half the variation in incomes and housing prices is unresolved, presumably because they depend on factors not observable from orbit, such as tax policies or school districts. Thus, we expect a performance ceiling, in the sense that R^2 may be bounded substantially below one, that is specific to each task and cannot be known with certainty.

The dimensionality of MOSAIKS's featurization, and in turn its computational cost and predictive skill, is easily manipulated by dropping or adding additional random features. Cross-validation experiments indicate that increasing the dimensionality of the feature space provides minor gains above $K = 1,000$ features (Fig. 4.3A). A majority of the observable signal with our baseline of $K = 8,192$ features is recovered using only 100 random features (min 81% for income, max 96% for nighttime lights).

Training sets with $N = 8,000$ obtain near-maximum performance relative to that obtained with $N = 64,000$ (Fig. 4.3A), but significant signal is recovered for many outcomes using only $N = 500$ (min 56% for road length, max 87% for forest cover), with the exception of income and housing price tasks, which require larger samples.

A key motivation for using satellites is the ability to collect information in large contiguous areas where labels are not available. To systematically evaluate performance under such conditions, we partition the US in a checkerboard pattern (Fig. 4.3B), training on the "black squares" and testing on the "white squares." Increasing the size of squares (δ) in the checkerboard increases the average distances between train and test observations, simulating increasingly large spatial extrapolations.

For $\delta = 4°$ (444km×341km regions at sample centroid) we find limited loss of performance (R^2 declines $< 10\%$, see Fig. 4.3C and Fig. S9) except for modest declines for income (33%) and road length (38%) and intermediate declines for housing price (52%). Extending extrapolations to $\delta = 16°$ (1778km×1366km) produces limited additional loss in performance (R^2 declines $< 10\%$ further), except for larger losses for income (17%) and elevation (41%) and the collapse of road length performance, possibly due to missing label and data quality (Supplementary Information Section I.1.6 and Figure S1).

MOSAIKS substantially outperforms optimized spatial interpolation of observations, a widely used approach to "fill in" large regions of missing data, across all tasks except elevation and housing price (Fig. 4.3C, grey dashed lines). At its highest performance ($\delta = 0.5°$), spatial interpolation recovers only 25% (nighttime lights) to 80% (forest cover) of the MOSAIKS R^2 except for elevation, where interpolation performs almost perfectly over small ranges ($\delta = 0.5° : R^2 = 0.95$), and housing price, where interpolation slightly outperforms MOSAIKS at small ranges. In both elevation and housing price, interpolation performance converges to that of MOSAIKS over larger distances. Thus, in addition to generalizing across tasks, MOSAIKS generalizes out-of-sample across space with moderate to high fidelity, outperforming spatial interpolation of ground-truth in 5 of 7 tasks.

Having evaluated MOSAIKS systematically in the US, where data sources are relatively diverse and reliable, we demonstrate its ability to scale globally on four of our original tasks for which global labels exist. Using a random sub-sample of locations (train-

ing and validation: N = 444,820, test: N = 111,205; Supplementary Information Section II.10), we construct the first high-resolution planet-scale estimates for the distribution of forest cover ($R^2 = 0.59$), elevation ($R^2 = 0.25$), population density ($R^2 = 0.57$), and nighttime lights ($R^2 = 0.42$) using only a single set of label-independent features ($K = 2048$, Fig. 4.4A). Inconsistent image quality and label reliability, as well as plausibly heterogeneous relationships between imagery and task outcomes across the globe, appear to lower performance relative to the US-only experiments above (Supplementary Information Section II.10). However, increasing training samples and additional optimization will likely improve performance beyond these demonstrated levels (Table S4). Similar to the US-only experiments, predicting extremely high elevations exhibits the largest systematic prediction errors across attempted tasks.

In addition to global scalability, our approach naturally delivers observations that are more finely resolved than the labels in the original training data, thus achieving super-resolution in label predictions. Specifically, MOSAIKS estimates outcomes for sub-regions within images even though image-level labels are only ever used in training (Fig. 4.4B and Fig. S11). Super-resolution results naturally from the linearity of Eq. (4.2) combined with labels being linear combinations of ground-level conditions (Supplementary Information Section II.9 and Fig. S10). Essentially, MOSAIKS estimates the relative contribution of sub-regions within an image to the overall image-level labels. To demonstrate this property, we examine forest cover, our only task where raw ground-truth data is available at substantially finer resolutions than our images. Training only with image-level labels, we nonetheless recover within-image signal (within-image R^2 = 0.31-0.10, see Fig. 4.4C and Supplementary Information Section II.9) when estimating forest cover in 4 to 64 sub-images per image, although signal-to-noise ratios decline at high super-resolutions ($>$ 256 subimages).

Discussion

MOSAIKS aims to complement major prior efforts to both deploy satellite-based instruments globally and harmonize and post-process the massive data sets they retrieve. MOSAIKS allows these Earth observation data to be applied to arbitrary tasks without customization or expert knowledge, while still achieving performance comparable to highly tuned SIML systems designed for specialized tasks (Supplementary Information Section III.1).

Generalization of SIML across tasks dramatically reduces overall costs for a global research ecosystem where imagery is collected and then analyzed to evaluate or monitor a large number of outcomes at planet-scale (Supplementary Information Section III.2). Thus we hope that MOSAIKS will democratize and accelerate access to Earth observation, especially in low-income and data-poor contexts [39, 132], empowering unprecedented progress toward resolving pressing global challenges [140].

Here, we purposefully restrict our analysis to high-resolution tri-band daytime imagery to explore the capabilities of MOSAIKS, but it is straightforward to extended this approach to other remotely sensed and/or gridded products, including hyper-spectral and non-optical data.

Our experiments demonstrate that MOSAIKS is generally more accurate than interpolation from nearby ground-truth data. However, we expect that hybrid approaches which leverage both data sources might improve performance, especially for tasks with important factors unobservable from satellites (e.g. housing price).

The ability of MOSAIKS to match and exceed the performance of a fine-tuned ResNet-18 CNN across all tasks leads us to consider whether concepts from this work can contribute to continued improvement of CNNs applied to SIML. We hypothesize that CNN performance may benefit from incorporating wider, and perhaps shallower, architectures.

We expect that future work may further enhance the performance of MOSAIKS and, in some cases, bespoke task-specific solutions to Eq. (4.1) may achieve higher performance, especially after extensive tuning. Future work should compare MOSAIKS performance to that of task-tailored models such as in [[52, 40, 78]]. To aid in development, bench-marking, and comparisons of SIML approaches, labels used in this study are made publicly available; to our knowledge this represents the largest multi-label benchmark dataset for SIML regression tasks.

Methods

Here we first provide some additional information on our implementation of MOSAIKS and experimental procedures. We then provide more description of the theoretical foundation underlying MOSAIKS. Additional details are contained in the Supplementary Information.

Implementation of MOSAIKS

Each feature $\mathbf{x}_k$ generated by MOSAIKS for an image $\mathbf{I}_\ell$ is created by convolving an $M \times M \times S$ "patch", $\mathbf{P}_k$, across the entire image, where M is the width and height of the patch in units of pixels and S is number of spectral bands. In each step of the convolution, the inner product of the patch and an $M \times M \times S$ sub-image region is taken, and a ReLU activation function with bias $b_k = 1$ is applied. Each patch is a randomly sampled sub-image from the set of training images (Fig. S4). We use patches of width and height $M = 3$ (Fig. S5) and $S = 3$ bands (red, green, and blue). To create a single summary metric for the image-patch pair, these values are then averaged across the entire image, generating the kth feature $\mathbf{x}_k(\mathbf{I}_\ell)$. The dimension of the resulting feature space is equal to K, the number of patches used, and in all of our main analyses we employ $K = 8,192$ (i.e. 2^{13}). Both images and patches are whitened according to a standard image preprocessing procedure before convolution (Supplementary Information Section II.3).

In practice, this one-time featurization can be centrally computed and then distributed to users in tabular form. The featurization described above with $K = 8,912$ features results in a roughly 6 to 1 compression of stored and transmitted imagery data in the cases we study. Notably, storage and computational cost can be traded off with performance by using more or fewer features from each image (Fig. 4.3). Since features are random, there is no natural value for K that is specifically preferable.

Users acquire the centrally calculated feature set, merge a dataset of labels based on geographic location, and then learn *nonlinear* mappings from the original image pixel values to the labels by training a *linear* regression of the labels on the features. We show strong performance across seven different tasks using ridge regression in this last step, although future work may demonstrate that other fitting procedures yield similar or better results for particular tasks.

Experimental procedures

Task selection and data Tasks were selected based on diversity and data availability, with the goal of evaluating the generalizability of MOSAIKS (Supplementary Information Section I.1). Results for all tasks evaluated are reported in the paper. We align image and label data by projecting imagery and label information onto a ∼1km × 1km grid, which was designed to ensure zero spatial overlap between observations (Supplementary Information Sections II.1 and II.2). Images are obtained from the Google Static Maps API (Supplementary Information Section I.2) [33], and labels for the seven tasks are obtained from refs. [[40, 8, 16, 71, 119, 120, 138]]. Details on data are described in Supplementary Information Table S1 and Section I.

US experiments From this grid we sample 20,000 hold-out test cells and 80,000 training and validation cells from within the continental US (Supplementary Information Section II.4). To span meaningful variation in all seven tasks, we generate two of these 100,000-sample data sets according to different sampling methods. First, we sample uniformly at random across space for the forest cover, elevation, and population density, tasks which exhibit rich variation across the US. Second, we sample via a population-weighted scheme for nighttime lights, income, road length, and housing price, tasks for which meaningful variation lies within populated areas of the US. Some sample sizes are slightly reduced due to missing label data ($N = 91,377$ for income, $73,411$ for housing price, and $67,968$ for population density). We model labels whose distribution is approximately log-normal using a log transformation (Supplementary Information Section II.5 and Table S3).

Because fitting a linear model is computationally cheap, relative to many other SIML approaches, it is feasible to conduct numerous sensitivity tests of predictive skill. We present cross-validation results from a random sample, while also systematically evaluating the behavior of the model with respect to: (a) geographic distance between training and testing samples, (b) the dimension K of the feature space, and (c) the size N of the training set (Fig. 4.3, Supplementary Information Sections II.7 and II.8). We also benchmark model performance and computational expense against an 18-layer Residual Network, a common deep network architecture that has been used in satellite based learning tasks [80] (Supplementary Information Sections III.1 and III.2).

Global experiment To demonstrate performance at scale, we apply the same approach used within the data-rich US context to global imagery and labels. We employ a target sample of $N = 1,000,000$, which drops to a realized sample of $N = 556,025$ due to missing imagery and label data outside the US (Fig. 4.4). We generate pre-

dictions for all tasks with globally available labels (forest cover, elevation, population density, and nighttime lights) (Supplementary Information Section II.10).

Super-resolution experiment Predictions at super-resolution (i.e. higher resolution than that of the labels used to train the model), shown in Fig. 4.4B, are generated for forest cover and population density by multiplying the trained ridge regression weights by the un-pooled feature values for each sub-image (Supplementary Information Section II.9). Additional examples of super-resolution performance are shown in Fig. S11. We quantitatively assess super-resolution performance (Fig. 4.4C) using forest cover, as raw forest cover data are available at substantially finer resolution than our common ∼1km x 1km grid. Performance is evaluated by computing the fraction of variance (R^2) within each image that is captured by MOSAIKS, across the entire sample.

Theoretical foundations

MOSAIKS is motivated by the goal of enabling generalizable and skillful SIML predictions. It achieves this by embedding images in a basis that is both descriptive (i.e. models trained using this single basis achieve high skill across diverse labels) and efficient (i.e. such skill is achieved using a relatively low-dimensional basis). The approach for this embedding relies on the theory of "random kitchen sinks" [87], a method for feature generation that enables the linear approximation of arbitrary well-behaved functions. This is akin to the use of polynomial features or discrete Fourier transforms for function approximation generally, such as functions of one dimension. With inputs of high dimension, such as the satellite images we consider, it has been shown experimentally [68, 21, 55] and theoretically [87] that a randomly selected subspace of the basis often performs as well as the entire basis for prediction problems.

Convolutional random kitchen sinks Random kitchen sinks approximate arbitrary functions by creating a finite series of features generated by passing the input variables z through a set of K nonlinear functions $g(z; \Theta_k)$, each paramaterized by draws of a random vector Θ. The realized vectors Θ_k are drawn independently from a pre-specified distributions for each of $k = 1...K$ features. Given an expressive enough function g and infinite K, such a featurization would be a universal function approximator [86]. In our case, such a function g would encode interactions between all subsets of pixels in an image. Unfortunately, for an image of size $256 \times 256 \times 3$, there are $2^{256\times256\times3}$ such subsets. Therefore, the fully-expressive approach is inefficient in generating predictive skill with reasonably concise K because each feature encodes more pixel interactions than are empirically useful.

To adapt random kitchen sinks for satellite imagery, we use convolutional random features, making the simplifying assumption that most information contained within satellite imagery is represented in *local* image structure. Random convolutional features have been shown to provide good predictive performance across a variety of tasks from predicting DNA binding sites [68] and solar flares [55] to clustering photographs [21] (kitchen sinks have also been used in a non-convolutional approach to classify individual pixels of hyper-spectral satellite data [81]). Applied to satellite images, random convolutional features reduce the number of effective parameters in the function by

considering only local spatial relationships between pixels. This results in a highly expressive, yet computationally tractable, model for prediction.

Specifically, we create each Θ_k by extracting a small sub-image patch from a randomly selected image within our image corpus. These patches are selected independently, and in advance, of any of the label data. Each patch is then convolved across the satellite image being featurized, and passed through a pixel-wise ReLU nonlinearity to form an activation map. This convolution captures information from the entire $\mathbb{R}^{256\times256\times3}$ image using only $3 \cdot M^2$ free parameters for each k. Creating and subsequently averaging over the activation map defines our instantiation of the kitchen sinks function $g(z;\Theta_k)$ as $g(\mathbf{I}_\ell;\mathbf{P}_k,b_k) = \mathbf{x}_k(\mathbf{I}_\ell)$, where b_k is a scalar bias term. Our choice of this functional form is guided by both the structural properties of satellite imagery and the nature of common SIML prediction tasks, and it is validated by the performance demonstrated across tasks.

Relevant structural properties of satellite imagery and SIML tasks Three particular properties provide the the motivation for our choice of a convolution and average-pool mapping to define g.

First, we hypothesize that convolutions of small patches will be sufficient to capture nearly all of the relevant spatial information encoded in images because objects of interest (e.g. a car or a tree) tend to be contained in a small sub-region of the image. This is particularly true in satellite imagery, which has a much lower spatial resolution that most natural imagery (Supplementary Information Figure S5).

Second, we expect a single layer of convolutions to perform well because satellite images are taken from a constant perspective (from above the subject) at a constant distance and are (often) orthorectified to remove the effects of image perspective and terrain. Together, these characteristics mean that a given object will tend to appear the same when captured in different images. This allows for MOSAIKS's relatively simple, translation invariant featurization scheme to achieve high performance, and avoids the need for more complex architectures designed to provide robustness to variation in object size and orientation.

Third, we average-pool the convolution outputs because most labels in Earth observation problems can be approximately decomposed into a sum of sub-image characteristics. For example, forest cover is measured by the percent of total image area covered in forest, which can equivalently be measured by averaging the percent forest cover across sub-regions of the image. Labels that are strictly averages, totals, or counts of sub-image values (such as forest cover, road length, population density, elevation, and night lights) will all exhibit this decomposition. While this is not strictly true of all SIML tasks, for example income and average housing price, we demonstrate that MOSAIKS still recovers strong predictive skill on these tasks. This suggests that some components of the observed variance in these labels may still be decomposable in this way, likely because they are well-approximated by functions of sums of observable objects.

Additional interpretations The full MOSAIKS platform, encompassing both featurization and linear prediction, bears similarity to a few related approaches. Namely, it can be interpreted as a computationally feasible approximation of kernel ridge re-

gression for a fully convolutional kernel or, alternatively, as a two-layer CNN with an incredibly wide hidden layer generated with untrained filters. A discussion of these interpretations and how they can help to understand MOSAIKS's predictive skill can be found in Supplementary Information Section II.3.

Figure 4.1: A generalized approach to combining satellite imagery with machine learning (SIML) without users handling images. MOSAIKS is designed to solve an unlimited number of tasks at planet-scale quickly, recognizing that every satellite image may be evaluated according to every potential task. Centralized calculations allow MOSAIKS to "pre-organize" all images based on large number of non-parametric random features ("kitchen sinks") describing the content of each image. Later, when a task is determined by a user, MOSAIKS simply reorders all images based on the features that are relevant to that task. (A) For example, nine images from the US sample are ordered based on two independent tasks, population density and forest cover, both of which have distinct identifying features that are observable in each image. (B) Schematic of the MOSAIKS process. N images are each transformed using random convolutional features (Methods and Supplementary Information Section II.3) into a compressed and highly descriptive K-dimensional feature vector before labels are known. Once features are computed, they can be stored in tabular form (matrix X) and used for unlimited tasks without recomputation. Users interested in a new task (s) merge their own labels (y^s) to features for training in locations where labels are known. Here, *User 1* has forest cover labels for locations $p+1$ to N and *User 2* has population density labels for locations 1 to q. Each user then solves a single linear regression for β^s in $\mathbb{R}^K$. Linear prediction using β^s and MOSAIKS features $\mathbf{X}$ then generates SIML estimates for label values at all locations. Generalizability allows different users to solve different tasks using an identical procedure and the same table of features—differing only in the user-supplied label data for training. Each task can be solved by a user on a desktop computer in minutes without users ever manipulating the imagery.

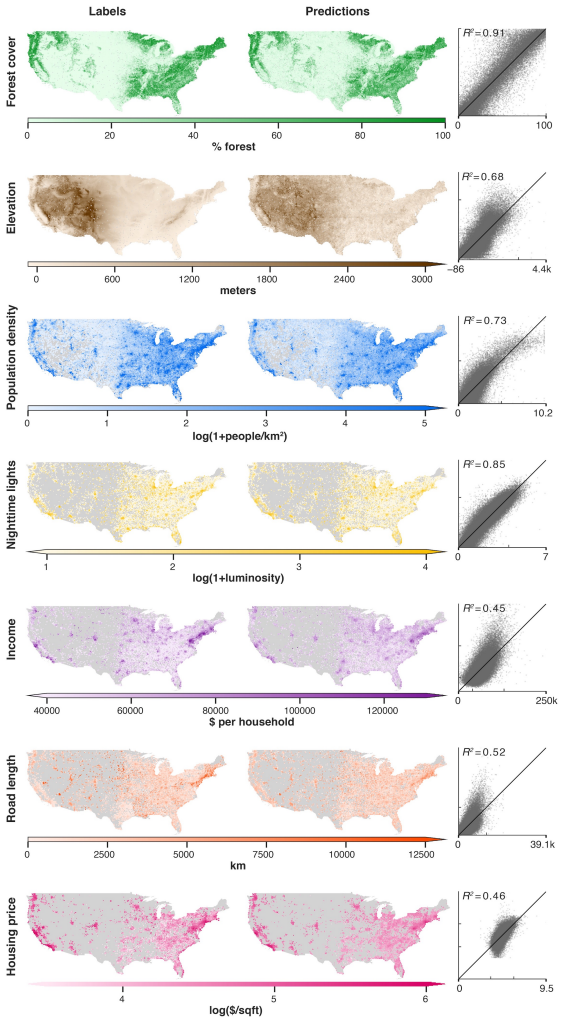


Figure 4.2: High resolution prediction of many tasks across the continental US using daytime images processed once, before tasks were chosen. 100,000 daytime images were each converted to 8,192 features and stored. Seven tasks were then selected based on coverage and diversity to evaluate performance. Eq. (4.2) is solved for each task using the same procedure. Left maps: 80,000 observations used for training and validation, aggregated up to 20km×20km cells for display. Right maps: concatenated validation set estimates from 5-fold cross-validation for the same 80,000 grid cells (observations are never used to generate their own prediction), identically aggregated for display. Scatters: Validation set estimates (vertical axis) vs. "ground-truth" (horizontal axis); each point is a ∼1km×1km grid cell. Black line is at 45°. Test set and validation set performance are essentially identical (Table S2), validation set values are shown for display purposes only since there are more observations. The top three tasks are uniformly sampled across space, bottom four tasks are sampled using population weights (Supplementary Information Section II.1); grey areas did not generate a sample in the experiment.

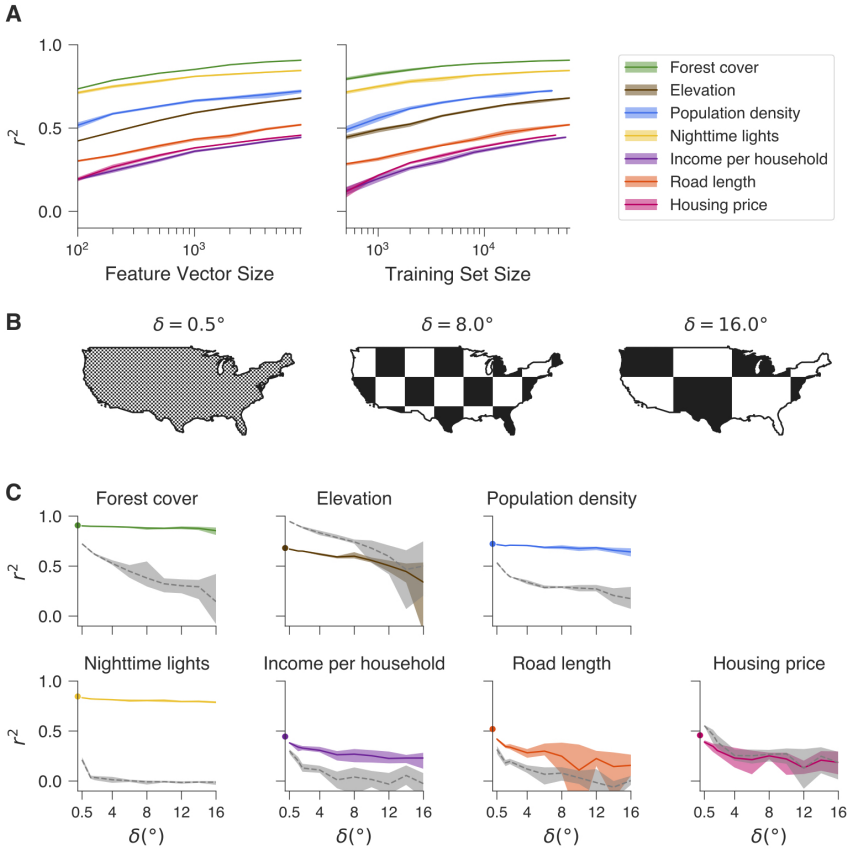


Figure 4.3: Fidelity achieved for smaller K, N, and over large contiguous regions with no ground truth data. (A) Validation set R^2 performance for all seven tasks while varying the number of random features K and holding $N = 64,000$ (left) and while varying N and holding $K = 8,192$ (right). Shaded bands indicate the range of predictive skill across 5 folds. Lines indicate mean skill. (B-C) Evaluation of performance over regions of increasing size that that are excluded from training sample. (B) Data is split in half using a "checkerboard" partition, where the width and height of each square is δ (measured in degrees). Example partitions with $\delta = 0.5°$, 8°, 16°. For a given δ, training occurs using data sampled from "black squares" and performance is evaluated in "white squares." (C) Colored lines are average performance of MOSAIKS in the US across δ values for each task. Benchmark performance from Fig. 4.2 are indicated as circles at $\delta = 0$. Grey dashed lines indicate corresponding performance using only spatial interpolation with an optimized radial basis function kernel instead of MOSAIKS (Supplementary Information Section II.8). To moderate the influence of the exact placement of square edges, training and test sets are resampled four times for each δ with the checkerboard position re-initialized using offset vertices (see Supplementary Information Section II.8 and Fig. S9). The range of outcomes are plotted as colored or grey bands.

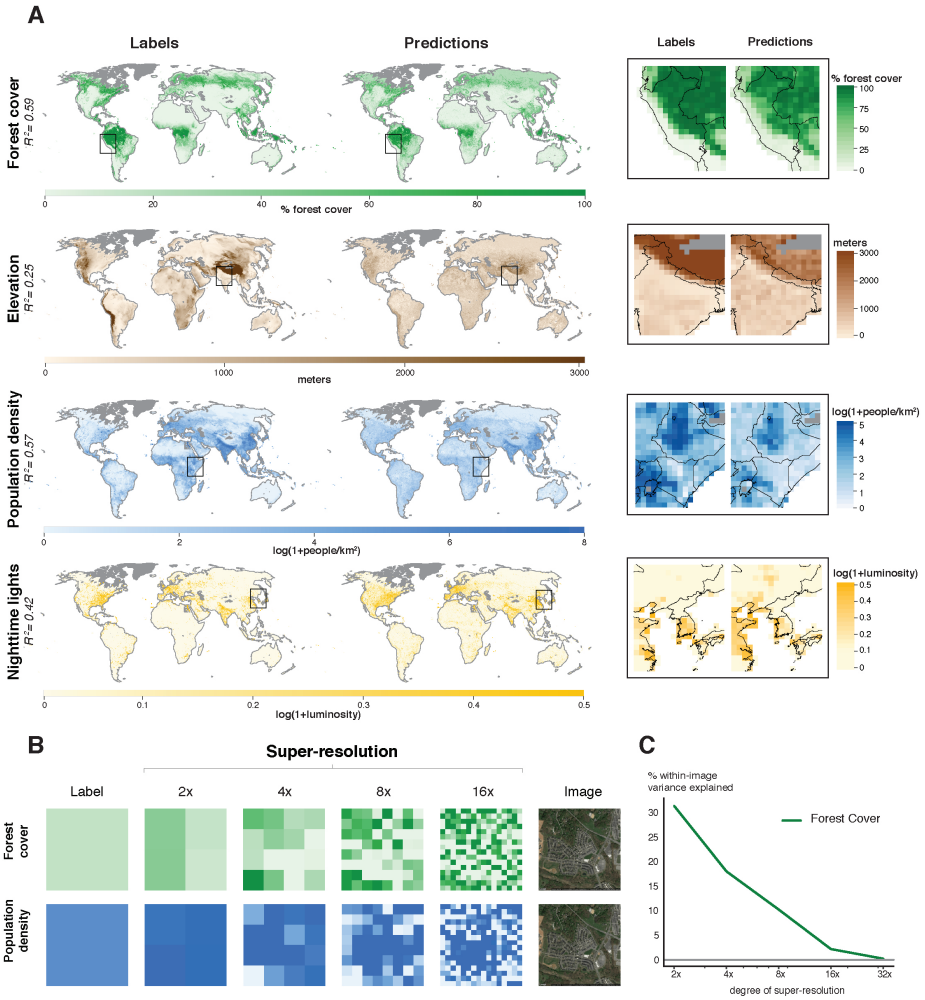


Figure 4.4: Scaling MOSAIKS to observe multiple variables simultaneously across the planet and achieving super-resolution. (A) Training data (left maps) and estimates using a single featurization of daytime imagery (right maps). Insets (far right) marked by black squares in global maps. Training sample is a uniform random sampling of 1,000,000 grid cells, 556,025 for which imagery were available and could be matched to task labels. Out-of-sample predictions are constructed using 5-fold cross-validation. For display purposes only, maps depict ∼50km × 50km average values (ground truth and predictions at ∼1km × 1km). (B-C) Both labels and features in MOSAIKS are linear combinations of sub-image ground-level conditions, allowing β^s to be applied to imagery of any spatial extent (Supplementary Information Section II.9). Thus, the system achieves super-resolution by generating meaningful estimates at spatial resolutions finer than the original labels used for training. (B) Example super-resolution estimates at 2×, 4×, 8×, and 16× label resolution (See Fig. S11 for additional examples). (C) Systematic evaluation of within-image R^2 across the entire sample recovered in the forest cover task (US only; Supplementary Information Section II.9).

Supplementary Information

The primary goal of our analysis is to develop, evaluate and contextualize the performance of MOSAIKS. In the following three supplementary sections we describe the data used in this evaluation, the experiments conducted, and how MOSAIKS compares to other approaches in the literature. We also describe the intuition behind and the mechanics of MOSAIKS's algorithms in greater detail.

I Data

This section describes the datasets we use to construct our ground truth labels across all seven of our tasks: forest cover, elevation, population density, nighttime lights, income, road length, and housing price. In addition, we describe the imagery used in the analysis. In Section II.2 we detail our method for linking the labeled data for each outcome to the imagery (Fig. S3).

In evaluating the ability of MOSAIKS to generalize, we are interested in its ability to recover different types of variables, including: (i) variables that are averages of sub-image properties, (ii) variables that not directly observable through daytime imagery but are a function of visible objects in the image, such as nighttime lights, and (iii) variables that are an underlying factor that determines what material appears in the image, such as elevation. Labels may also be a combinations of (i)-(iii), such as housing price or household income. An advantage of MOSAIKS is that it solves all these cases without any alteration of method. In the main text, we use the the same set of image features to predict all seven outcomes and, in principle, this set of features can be used to predict an unlimited number of outcomes (Section II.3, so long as the outcomes and the images are aligned as described in Section II.2).

For each task, we obtain an up-to-date and geographically complete publicly available datasource to match with the images. Most of these data are based on measurements from 2010 - 2015, though our data on population density draws from sources that date back as far as 2005 in order to achieve global coverage. Our imagery data, from the Google Static Maps API (Section I.2), was mostly acquired in 2018, though in some cases images may be a few years older.

Task	Units	Native resolution	Data source
Forest cover	% forest cover	∼30m × 30m	[40]
Elevation	meters	∼611.5m × 611.5m	[8]
Population density	people per sq. km.	∼1km × 1km	[16]
Nighttime lights	radiance	∼500m × 500m	[71]
Income	USD per household	census block group	[119]
Road length	km	polyline	[120]
Housing price	USD per sq. ft.	geocoded point data	[138]

Extended Data Table S1: Data sources for all tasks. Note that for all raster data sets (forest cover, elevation, population density, and nighttime lights) stated resolutions apply to grid cells located at the equator; raster size in Euclidean distance will vary with latitude.

I.1 Labels

Tasks were chosen to represent outcomes of classes (i)-(iii) above subject to the condition that high resolution and up-to-date label data is available across the US. Below we describe these data sources. See Section II.2 and Fig. S3 for a description of how we assign raw label data to images.

I.1.1 Forest cover To measure forest cover, we use globally comprehensive raster data from [[40]], which is designed to accurately measure forest cover in 2010. This dataset is commonly used to measure forest cover when ground-based measurements are not available [6, 15]. Forest in these data is defined as vegetation greater than 5m in height, and measurements of forest cover are given at a raw resolution of roughly 30m by 30m. These estimates of annual maximum forest cover are derived from a model based on Landsat imagery captured during the growing season. Specifically, the authors train a pixel-level bagged decision tree using three types of features: "(i) reflectance values representing maximum, minimum and selected percentile values (10, 25, 50, 75 and 90% percentiles); (ii) mean reflectance values for observations between selected percentiles (for the max-10%, 10-25%, 25-50%, 50-75%, 75-90%, 90%-max, min-max, 10-90%, and 25-75% intervals); and (iii) slope of linear regression of band reflectance value versus image date." These estimates of forest cover were derived using different spectral bands than we observe in our imagery, and using information about how surface reflectance changes over the growing season, which we did not observe. This gives us confidence that we are indeed learning to map visual, static, high-resolution imagery to forest cover, rather than simply recovering the model used in [[40]].[1]

I.1.2 Elevation We use data on elevation provided by Mapzen, and accessed via the Amazon Web Services (AWS) Terrain Tile service. These Mapzen terrain tiles provide global elevation coverage in raster format. The underlying data behind the Mapzen tiles comes from the Shuttle Radar Topography Mission (SRTM) at NASA's Jet Propulsion Laboratory (JPL), in addition to other open data projects.

These data can be accessed through AWS at different zoom levels, which range from 1 to 14 and, along with latitude, determine the resolution of the resulting raster. To align with the resolution of our satellite imagery, we use zoom level 8, which leads to a raw resolution of 611.5 meters at the equator.[2]

I.1.3 Population density We use data on population density from the Gridded Population of the World (GPW) dataset [16]. The GPW data estimates population on a global 30 arc-second (roughly 1 km at the equator) grid using population census tables and geographic boundaries. It compiles, grids, and temporally extrapolates population data from 13.5 million administrative units. It draws primarily from the 2010 Population and Housing Censuses, which collected data between 2005 and 2014. GPW data in the US comes from the 2010 census.[3]

[1] These data can be accessed at: `https://landcover.usgs.gov/glc/TreeCoverDescriptionAndDownloads.php`.

[2] We accessed these data via the R function `get_aws_terrain` from the `elevatr` package. Code and documentation can be found here: `https://www.github.com/jhollist/elevatr`.

[3] These data can be accessed at `http://sedac.ciesin.columbia.edu/data/collection/gpw-v4`

I.1.4 Nighttime lights We use luminosity data generated from nighttime satellite imagery, which is provided by the Earth Observations Group at the National Oceanic and Atmospheric Administration (NOAA) and the National Geogphysical Data Center (NGDC). The values we use are Version 1.3 annual composites representing the average radiance captured from satellite images taken at night by the Visible Infrared Imaging Radiometer Suite (VIIRS). We use values from 2015, the most recent annual composite available.

This composite is created after the Day/Night VIIRS band is filtered to remove the effects of stray light, lightening, lunar illumination, lights from aurora, fires, boats, and background light. Cloud cover is removed using the VIIRS Cloud Mask product. These values are provided across the globe from a latitude of 75N to 65S at a resolution of 15 arc-seconds. The radiance units are nW cm^{-2} sr^{-1} (nanowatts per square centimeter per steradian).

Like forest cover, these labels are themselves derived from satellite imagery. However, because they capture luminosity at night, while our satellite imagery is taken during the day, the labels for luminosity and the imagery used to predict luminosity represent independent data sources. Our ability to predict nighttime lights depends on how well objects visible during the day are indicative of light emissions at night.[4]

I.1.5 Income We use the American Community Survey (ACS) 5-year estimates of median annual household income in 2015. These data are publicly available at the census block group level, of which there are 211,267 in the US, including Puerto Rico. On average, block groups are around 38 km^2, though block groups are smaller in more densely populated areas.[5]

I.1.6 Road length We use road network data from the United States Geological Survey (USGS) National Transportation Dataset, which is based on TIGER/Line data provided by US Census Bureau in 2016. Shapefiles for each state provide the road locations and types, including highways, local neighborhood roads, rural roads, city streets, unpaved dirt trails, ramps, service drives, and private roads. Road types are indicated by a 5-digit code Feature Class Code which is assigned by the Census Bureau.[6] The variable we predict is road length (in kilometers), which is computed as the total length of all types of roads that are recorded in a given grid cell.

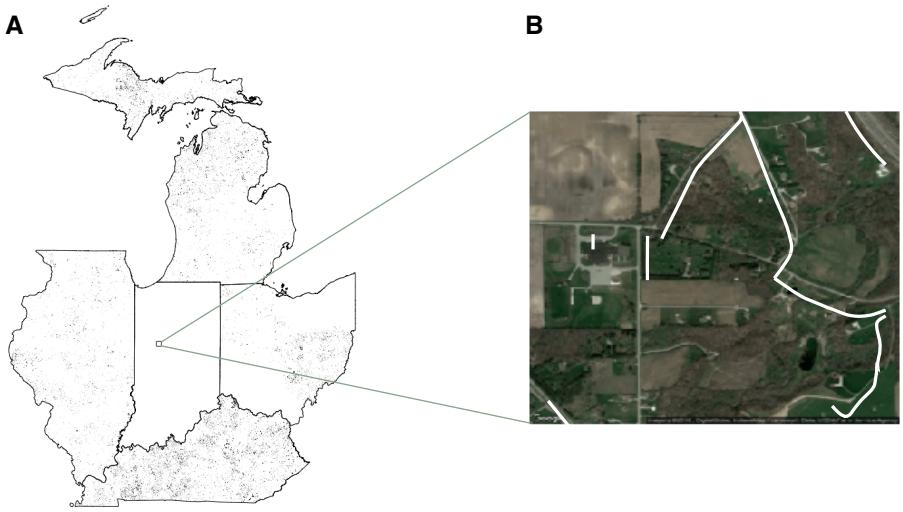
The Census Bureau database is created and corrected via a combination of partner supplied data, aerial images, and fieldwork. The spatial accuracy of linear features of roads and coordinates vary by source materials used. The accuracy also differs by region, causing cases in which some regions lack recordings of certain road types, the most common one being private roads and dirt trails. For example, private roads are rarely recorded in Indiana and some regions in Ohio (Fig. S1A), despite satellite images that suggest they are present (Fig S1B).[7]

[4]These data can be accessed at `https://www.ngdc.noaa.gov/eog/viirs/download_dnb_composites.html#NTL_2015`.

[5]These data are accessible using the `acs` package in R [32], table number B19013.

[6]`https://www.census.gov/geo/reference/mtfcc.html`

[7]The data can be accessed at: `https://prd-tnm.s3.amazonaws.com/index.html?prefix=`

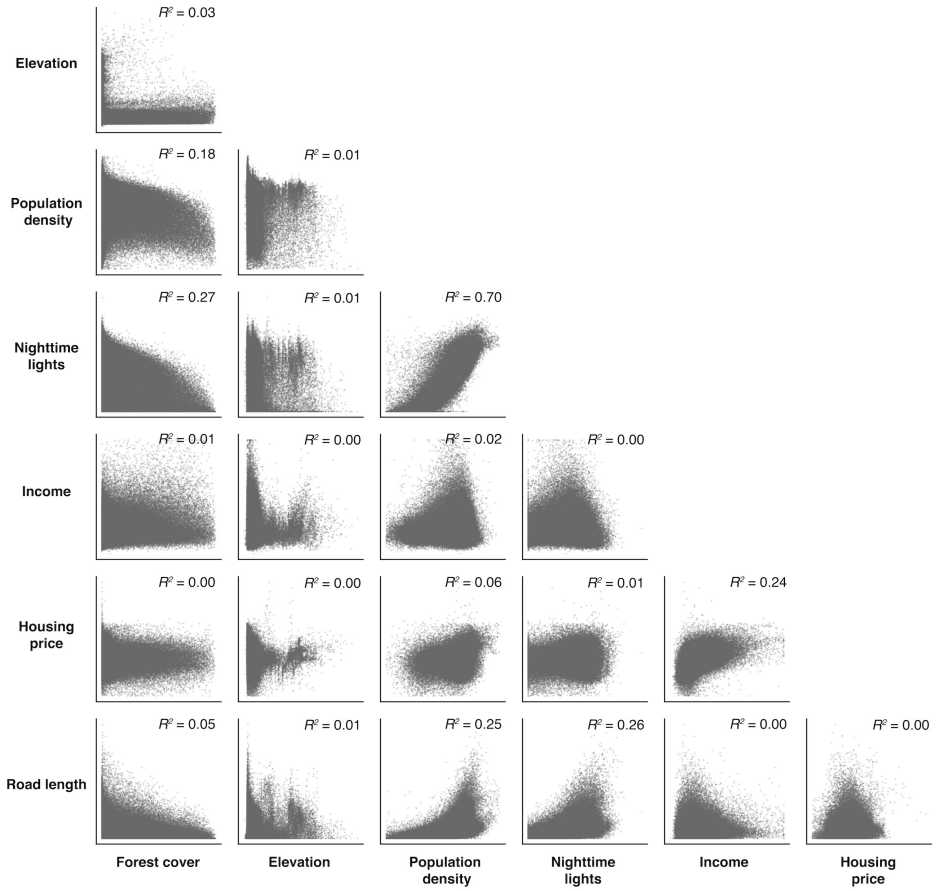


Extended Data Figure S1: Quality of ground truth road data varies by region. (A) Private roads in the northern Midwest recorded in the USGS National Transportation Dataset. The conspicuous lack of recorded private roads in Indiana and sections of Ohio suggests that road data quality in certain regions may be lacking. (B) Overlaying recorded roads of all types (shown in white) over a single satellite image in Indiana, demonstrates that some roads that are easily visible from satellite imagery are missing in the available data that we use to construct labels.

I.1.7 Housing price We estimate housing price per square foot using sale price and assessed square footage values for residential buildings, obtained through the Zillow Transaction and Assessment Database (ZTRAX). This dataset aggregates transaction and assessment data across the United States, combining reported values from states and counties with widely varying regulations and standards. Thus, significant data cleaning is required. Furthermore, because some states do not require mandatory disclosure of the sale price, we currently have limited data for the following states: Idaho, Indiana, Kansas, Mississippi, Missouri, Montana, New Mexico, North Dakota, South Dakota, Texas, Utah, and Wyoming. To address data quality issues, we develop a quality assurance and quality control (QA/QC) approach that is based on approaches employed in previous work [69, 31, 116] but adapted for our case.

ZTRAX contains data on the majority of buildings in the United States, initially comprising 374 million detailed records of transactions across more than 2,750 counties. The data is organized into two components - *transaction data* and *assessment data*. These two datasets are linked, allowing us to merge the latest sale price of a property to the latest assessment data. To minimize the effect of nation-wide trends in housing price that would be unobservable from our cross-sectional satellite imagery, we limit our dataset to sales occurring in 2010 or later. Further, we restrict our analysis to buildings coded as "residential" or "residential income - multi-family" and drop any

`StagedProducts/Tran/Shape/`



Extended Data Figure S2: Correlation of labels across tasks. Each figure shows a scatter plot of labeled outcomes for one of our seven tasks against another. All points come from a population-weighted random sampling of grid cells (as described in Section II.1) across the US. Scatters and R^2 values are shown across approximately 100,000 grid cell labels, depending on the data availability for each task.

sale that was coded as an intra-family transfer. To obtain a square footage value, we follow the example in Zillow Research's GitHub repository [139] and take the maximum reported square footage for a given improvement, and then sum over all improvements on a given property.

To reduce the number of potentially miscoded outliers at the bottom end of the distribution of sale price and property size, we drop any remaining sales that fall under \$10,000 USD, any properties that fall under 100 sq. ft., and any \$/sq. ft. values under \$10. To address outliers on the high end of the distribution, we take this restricted sample and further cut our dataset at the 99th percentile of \$/sq. ft. by state. Afterwards, we select the most recent recorded sale price for each property (divided by the most recent assessed square footage) to comprise our final dataset of housing price per square foot.

I.1.8 Correlation of outcomes across tasks The seven tasks described above were selected in order to evaluate the performance of MOSAIKS across many diverse contexts. Figure S2 evaluates the extent to which this was achieved, by plotting label values against one another. A few of the labels are moderately correlated, most notably population density and nighttime lights, but in general there is substantial orthogonal variation across these seven tasks.

I.2 Imagery

We use satellite imagery from Google Static Maps API [33], zoom level 16 (see Fig. 4.1A for examples). This gives roughly 1km × 1km images which are 640 × 640 pixels across and 3 dimensions deep (red, green, and blue spectral bands). We coarsen these images to 256 × 256 × 3 prior to featurizing, meaning that our models are trained on images with roughly 4m resolution. These images can be composites of several satellite images – sources include the Landsat, Sentinel, SPOT, Pleiades, WorldView and QuickBird satellites.[8] Prior to downloading, images were geo-rectified and pre-processed to remove cloud occlusions.[9]

II Methods

This section describes the methods that we use to define samples (Section II.1), to construct labels (Section II.2), and to construct features (Section II.3) for each image. It then describes how we separate data for training and evaluation (Section II.4), train models (Section II.5), test predictive skill (Section II.6), test sensitivity to the dataset size (Section II.7) and test model extrapolation performance (Section II.8). Next, we describe tests of model performance at sub-label or "super" resolution as well as at the global scale (Sections II.9 and II.10).

II.1 Grid definition and sampling strategy

II.1.1 Grid definition: To evaluate the generalizability of MOSAIKS performance across tasks we need a standardized unit of observation to link raw labels for all tasks and imagery. To do this, we construct a single global grid onto which we project both satellite imagery and labeled data. We design the grid to match our source of satellite imagery to ensure adjacent images do not overlap. Each element of the grid, i.e. each "grid cell," was designed to be a square in physical space. Because the earth is a sphere, the angular extent of grid cells changes across latitudes.[10]

II.1.2 Sampling strategy: For our primary experiment in the continental US we subsample sets of 100,000 observations, roughly 1.25% of the grid cells in the continental US, using two distinct sampling strategies.[11] First, we sample uniformly-at-random

[8] In some cases aerial photography is also integrated into images.

[9] More information is available at: `https://developers.google.com/maps/documentation/maps-static/dev-guide`.

[10] For the continental US (spanning 25 to 50 degrees latitude and -125 to -66 longitude), the grid cells are 0.0138 degrees in width (1.39 km) at the southern edge of the grid, and 0.0138 degrees in width (0.98 km) at the northern edge of the grid. The grid cells are 0.012 degrees in height (1.39 km) at the southern edge of the grid, and 0.0089 degrees in height (.98 km) at the northern edge of the grid.

[11] We discard marine grid cells, but do not discard grid cells that are composed only of lakes or smaller inland bodies of water.

(UAR) from all grid cells within the continental US. This sampling strategy is most appropriate for tasks like forest cover, where there is meaningful variation in most regions of the country. Second, we implement a population-weighted (POP) sampling strategy. To generate this sample, each grid cell is weighted by population density values taken from Version 4 of the Gridded Population of the World dataset, which provides a raster of population density estimates for the year 2015.[12] This weighted sampling strategy is most applicable to tasks like housing price, where the most meaningful variation lies in more populated regions of the US. We use the UAR grid when sampling population density to avoid any issues that might arise from sampling a task using the same variable as sampling weights. In both the UAR and POP samples, we randomly sample just once; all results in the paper are displayed using the same two subsets of the full grid. Note that these sub-sampled grid cells, by construction, are each covered by exactly one satellite image without having to process data over the entire US.

In our main results, we use the UAR sample for the forest cover, elevation, and population density tasks. We use the POP sample for nighttime lights, income, road length, and housing price. See Section II.10 for a discussion of how we extend this grid and sampling procedure to the global scale.

II.2 Assigning labeled data to sampled imagery

To assign labels to each grid cell, we spatially overlay our raw labeled data and our custom grid. The native format and spatial resolution of the labeled data vary across the tasks studied, necessitating different aggregation or disaggregation procedures for each task. Here, we describe the approach taken in each task (Fig. S3).

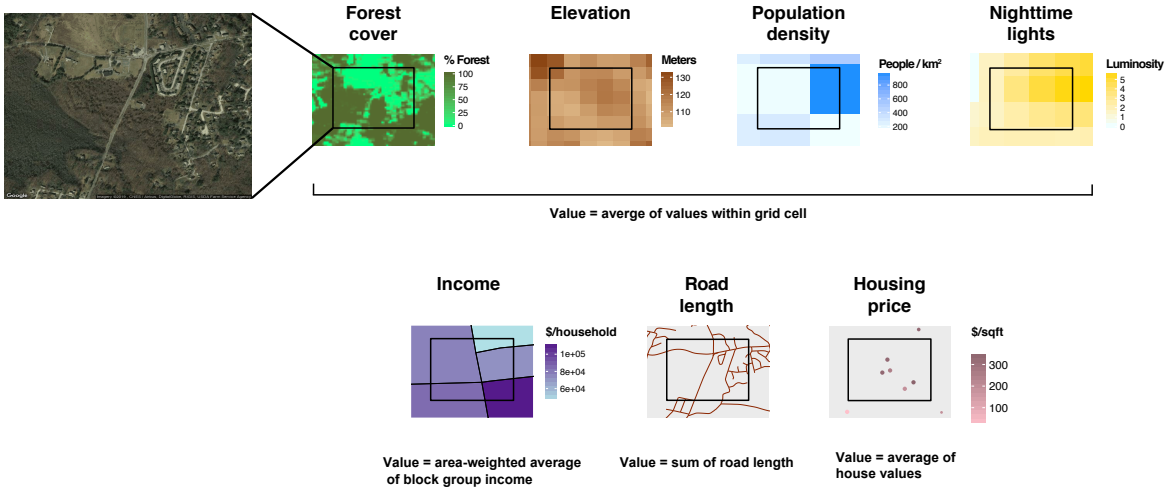
The raw forest cover, elevation, population density and nighttime lights data are provided natively as rasters with higher spatial resolution than our custom grid. For these tasks, we perform aggregation by calculating the mean of all labeled pixels with centroids that fall within the imagery grid cell. The resulting labels indicate mean forest cover, mean elevation, mean population density, and mean nighttime lights across the image grid cell.

Our road length data are provided as high-resolution spatial line segments. To aggregate these data to the image grid cell, we calculate the sum of road length segments within each image. The resulting labels indicate the total length of recorded roads that fall within an image grid cell.

Our housing price data are available as individual geocoded house sales. We aggregate these geocoded prices to the image grid cell by taking the average housing price per square foot across all sale prices that fall within the extent of the image. The resulting labels indicate the average housing price per square foot across all observed houses within a grid cell.

Our income data are provided at the block-group level (see Section I.1 for details). In some parts of the U.S., these block-groups are larger in total area than our image grid cells. However, in other regions, block-groups are smaller than our image grid cells. To treat both cases consistently, we aggregate incomes to the grid cell level by

[12]These data are available at `http://sedac.ciesin.columbia.edu/data/collection/gpw-v4/sets/bro`



Extended Data Figure S3: Calculation of grid cell labels from raw data. We calculate labels by spatially overlaying our grid cells and raw labeled data. We calculate labels as the average of raw label values that fall within the grid cell, except for roads where we calculate the label as the sum of road length within the grid cell.

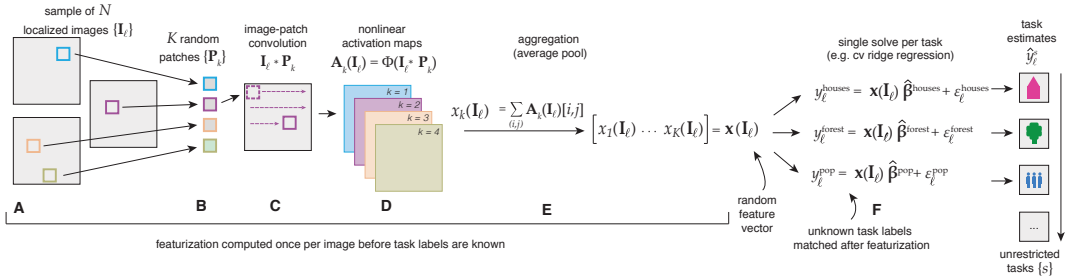
taking the weighted average of block-group incomes, where the weights are the area of intersection between the image grid cell and the block-group polygons. These weights are normalized to unity for each grid cell. The resulting labels indicate the area-weighted average median income across the grid cell.

Future users of a production-scale version of MOSAIKS would employ label data of arbitrary format and resolution. The above approaches provide guidelines for how to match various forms of label data to the pre-computed image feature grid, but other methods may be used. In the simplest case, for example, sparse point data could be directly matched to the nearest grid cell centroid.

II.3 Featurization of satellite imagery

II.3.1 Notation In our context, the input variable z is a set of satellite images $\mathbf{I}$, each corresponding to a physical location, ℓ. We use brackets to denote indexing into images, with colons denoting sub-regions of images (e.g. $\mathbf{I}_\ell[i,j]$ is the $(i,j)^{th}$ pixel of image $\mathbf{I}_\ell$, $\mathbf{I}_\ell[i:i+M, j:j+M]$ is the square sub-image of size $M \times M$ starting at pixel (i,j).) Because images have a third dimension (spectral bands), a colon $\mathbf{I}_\ell[i,j,:]$ denotes all bands at pixel (i,j). Indexing into non-image objects is denoted with subscripts (e.g. the k^{th} element of vector $\mathbf{x}$ is denoted as $\mathbf{x}_k$ and the k^{th} patch in a set of patches $\mathbf{P}$ is denoted as $\mathbf{P}_k$). We denote inner products with angular brackets $\langle \cdot, \cdot \rangle$ and the convolution operator with $*$.

II.3.2 Connection to the kitchen sinks framework The *random kitchen sink* featurization used in MOSAIKS relies on a nonlinear mapping $g(z; \Theta_k)$, where z is an input variable and Θ_k is a randomly drawn vector. Here, we describe the implementation details of this featurization in the context of satellite imagery. Connecting our implementation and notation to the framework of random kitchen sinks, the random



Extended Data Figure S4: MOSAIKS process from featurization to multi-task prediction. Given a large sample of N satellite images (A), a random sample of K patches (B) are drawn. (C) These K random patches P_k are convolved over each image $\mathbf{I}_\ell$ and (D) passed through a nonlinear function $\phi(\cdot)$ to generate K activation maps. (E) Pixel-specific activations are pooled across each image to generate one set of $N \times K$ features that are stored and distributed to all users. (F) The same random feature vector x is used in cross-validated ridge regression across many distinct tasks, after labeled and geo-referenced data y_ℓ is matched to features from each image $\mathbf{I}_\ell$ (as shown in Fig. 1B of the main text). (G) Models trained via ridge regression can be used to generate predictions across unrestricted tasks for any location with satellite imagery.

variables Θ_k are instantiated as the values of a random patch $\mathbf{P}_k$ and the bias b_k. The input variable z is an image $\mathbf{I}_\ell$, and $g(z;\Theta_k)$ represents the convolution of the patch over the image, followed by addition of the bias b_k and application of a element-wise ReLU function and an average pool, as described in the Methods of the main article and detailed below.

II.3.3 Methodological Details Fig. S4 depicts our featurization process. As described in Section I.1 and II.1, we begin with two sets (uniform and population-weighted samples) of $N = 100,000$ satellite images, each of which measures $640 \times 640 \times 3$ pixels (the third dimension represents the visible red, green, and blue spectral bands). We then coarsen the images to $256 \times 256 \times 3$ pixels to reduce computation. Next, we draw $K/2 = 4,096$ small sub-image "patches" of size $M \times M \times 3$ uniformly at random from the 80,000 images that comprise our training and validation set, and calculate the negative of each patch to get another $4,096$ patches (Fig. S4A, S4B). Our chosen specification sets $M = 3$, so that each patch $\mathbf{P}_k$ is of dimension $3 \times 3 \times 3$ (see Fig. S5 for performance in experiments using different patch sizes).

We then "whiten" each patch by zero components analysis (ZCA), a common pre-processing routine in image processing [58]. ZCA whitening pre-multiplies each patch by a transformation such that the resulting empirical covariance matrix of the whitened patches is the identity matrix. We then convolve each patch $\mathbf{P}_k$ over each of the N images (Fig. S4C) to obtain a set of $254 \times 254 \times 1$ pixel matrices for each image $\mathbf{I}_\ell$[13].During the convolutions each $3 \times 3 \times 3$ sub-image $\mathbf{I}_\ell[i : i + M, j : j + M, :]$ is also

[13]To improve efficiency of the featurization process, our implementation calculates the inner product of patch and image only for the original $K/2$ patches. We then create an additional $K/2$ values equal to the negative of each of the original inner products.

whitened according to the same whitening matrix as is applied to the patches.[14] We then apply a pixel-wise nonlinearity operator Φ to each resulting matrix to obtain K *nonlinear* activation maps $\mathbf{A}_k(\mathbf{I}_\ell) = \Phi(\mathbf{P}_k * \mathbf{I}_\ell + \mathbf{b}_k)$ for each image $\mathbf{I}_\ell$ (Fig. S4D) so that the $(i,j)^{th}$ pixel of the k^{th} activation map is defined as

$$\mathbf{A}_k(\mathbf{I}_\ell)[i,j] = \Phi(\langle \mathbf{I}_\ell[i:i+M, j:j+M, :], \mathbf{P}_k \rangle + b_k),$$

where b_k is a bias term from the constant bias matrix $\mathbf{b}_k$, in which every element is equal to $b_k = 1$. We use $\Phi(\mathbf{I}_\ell; \mathbf{P}_k, \mathbf{b}_k) = \text{ReLU}(\mathbf{P}_k * \mathbf{I}_\ell + \mathbf{b}_k) := \max\{\mathbf{P}_k * \mathbf{I}_\ell + \mathbf{b}_k, 0\}$ as the nonlinear operator. We then aggregate across the image by taking the average of the nonlinear activation maps (Fig. S4E). The combination of the nonlinear operator $\Phi(\cdot)$ and average pooling composes the function $g(\cdot)$ above, and creates a scalar value for each patch k and image ℓ pair:

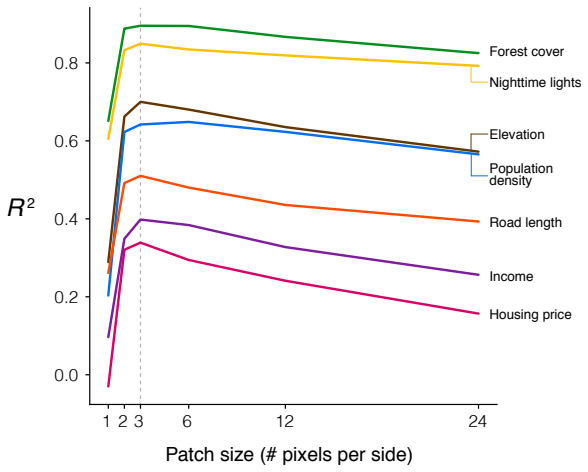
$$\mathbf{x}_k(\mathbf{I}_\ell) = \frac{1}{254^2} \sum_{i=1}^{254} \sum_{j=1}^{254} \mathbf{A}_k(\mathbf{I}_\ell)[i,j] \tag{4.3}$$

Stacking these scalars across all K patches provides the resulting K-dimensional feature vector, $\mathbf{x}(\mathbf{I}_\ell) := \begin{bmatrix} \mathbf{x}_1(\mathbf{I}_\ell) & \mathbf{x}_2(\mathbf{I}_\ell) & \dots & \mathbf{x}_K(\mathbf{I}_\ell) \end{bmatrix} \in \mathbb{R}^K$. This featurization thus embeds the original image $\mathbf{I}_\ell$ into a K-dimensional feature space, which can then be mapped to many different outcomes using task-specific models (s) implemented by researchers (r): $y_\ell^{s,r} = \mathbf{x}(\mathbf{I}_\ell)\beta^{s,r} + \epsilon_\ell^{s,r}$, as in Eq. (4.2) and illustrated in Fig. S4F. Although Eq. (4.2) is linear in these features, it may express a function that is highly nonlinear with respect to $\mathbf{I}_\ell$ because these features themselves are nonlinear with respect to the images.

II.3.4 Patch size and sampling We approximate the idealized complete convolutional basis, which contains features for all patch sizes, with the simpler truncated basis where we use only a single patch size. Throughout our main analysis, we use a $3 \times 3 \times 3$ patch size for $\mathbf{P}_k$. While larger patches may, in principle, enable the detection of image features with larger spatial structure, we find that, in practice, patch size $M = 3$ performs best across all seven tasks (Fig. S5). This finding suggests that most information contained within satellite imagery of this resolution can be represented by local-level image structure, and that the inclusion of "non-local" relationships reduces the efficiency of the function approximator by introducing more degrees of freedom. This empirical finding is consistent with previous applications of kitchen sink features [106].

We draw patches randomly from the empirical distribution of $M \times M \times 3$ patches from our training data set of satellite images. Drawing patches from the empirical distribution, rather than generating them randomly, allows us to sample efficiently from the distribution of sub-images we will encounter in the sample. This patch selection process is almost identical to the filter selection methods described in [[20, 89, 4]]. It may be valuable for future research to explore whether MOSAIKS performance and

[14] In practice, we apply the whitening operator as a right multiplication to the original 8192×27 whitened patch matrix in order to reduce computation.



Extended Data Figure S5: Performance by patch size. Featurization in MOSAIKS relies on convolving an $M \times M \times 3$ patch $\mathbf{P}_k$ across satellite images. M indicates the width in pixels of each sub-image patch, and the third dimension indexes the 3 spectral bands used throughout the analysis in this paper (an analogous approach can be applied to hyperspectral data). This figure shows, for each task, test set R^2 for patch sizes $M = 1, 2, 3, 6, 12$ and 24, using $K = 1,024$ patches for each M. The dotted gray line indicates the benchmark model used throughout the paper, with $M = 3$.

computational efficiency could be improved through patch selection algorithms. For example, one goal in selecting patches-based features is to promote relative sparsity in the resulting patch-based features, as in [[97]]. However, any attempt to tailor patch selection or featurization to a particular task of interest requires sacrificing the generalizability of this task-agnostic featurization. It remains an open question whether a non-randomly selected set of basis patches could potentially achieve similar (or greater) performance than what we present here when applied to arbitrary new tasks.

II.3.5 Alternative interpretations relating MOSAIKS to kernels and CNNs

The main article and methods section describes how MOSAIKS's convolutional random features enable nonparametric approximations of nonlinear functions through an embedding in a rich basis that expresses local spatial relationships. We believe that the description in the main article is the simplest and most direct explanation for the design of MOSAIKS and why it achieves high performance, but there are alternative ways to imagine a design process and implementation that would also lead to MOSAIKS as a final product. Here, we provide two of these alternative interpretations of the approach, the first relating to kernel methods and the second relating to convolutional neural networks. While these are not our preferred interpretations, we believe they provide useful lenses to consider why MOSAIKS works and they may be helpful to researchers thinking about related problems.

First, one could interpret the design of MOSAIKS *as if* we were attempting to design a computationally tractable approximation to implementing a ridge regression using a convolutional kernel and the kernel trick. Under this interpretation, one could arrive at

the same design of MOSAIKS using the following logic: (i) Design a kernel that allows us to describe the "similarity" of every image to every other image in the sample. (ii) For any new task, we want to use a kernel regression to predict the unobserved labels of new images based on their similarity to all other images—specifically, predicted labels would be a weighted sum of all observed labels using weights determined by this kernel-based measure of image similarity, i.e. the kernel trick. (iii) Unfortunately, calculating such a kernel exactly would be computationally intractable on a data set as large as the one we use, so instead use convolutional kitchen sinks (i.e. the featurization in MOSAIKS) to approximate the desired kernel regression. This last step follows from prior work demonstrating two concepts. First, random features can approximate the lifted feature space induced by well-known kernels [85] as the number of random features increases. Second, convolutions of random patches drawn from joint Gaussian distributions has been proven to approximate, in the limit, a kernel in which every sub-image from one image is compared with every sub-image from another using an arc-cosine distance function [55]. Thus, convolutions with random patches should, in the the limit, approximate a kernel that compares every sub-image with every other sub-image in the sample. However, because our distribution of patches is drawn from training imagery, rather than from Gaussian distributions, there is not an analytical expression that is known for the kernel being approximated by MOSAIKS in the limit.

The above logic would arrive at a design essentially the same as MOSAIKS, although it is not our preferred motivation or interpretation of why MOSAIKS works because it is a more complicated rationale than is needed. Ref. [[87]] showed that the existence of an associated kernel is not necessary for performance using kitchen sinks. Rather, it is simply the embedding of an input in a descriptive basis that provides the predictive skill, the insight that motivates our preferred—and we think simpler—interpretation presented in the main text. Nevertheless, the interpretation of MOSAIKS in the context of kernels motivates one way to understand the mechanism through which MOSAIKS achieves predictive skill at low computational cost. Namely, it enables the approximation of a nonparametric kernel regression, using some (unknown) fully convolutional kernel that is sufficiently rich to represent meaningful similarity between images but costly enough to prohibit a direct application of the kernel trick.

An additional way to contextualize MOSAIKS is in terms of its computational elements. In particular, MOSAIKS uses image convolutions and nonlinear activation operations common to convolutional neural networks (CNNs) [5]. Indeed, MOSAIKS is mathematically identical to the architecture one would arrive at if one designed a very shallow and very wide CNN without using backpropogation and instead using random filters. Specifically, MOSAIKS could be viewed as a two-layer CNN that has an 8,192-neuron wide hidden layer with untrained weights that are randomly initialized by drawing from sub-images in the sample, and that uses an average-pool over the entire image. In contrast to the conventional CNN approach of optimizing weights (via backpropogation), the random initialization with no subsequent optimization significantly reduces training time and avoids numerical challenges associated with non-convex optimization procedures (such as vanishing gradients). Thus, in the main text, we do not frame MOSAIKS as a CNN because MOSAIKS does not exploit the primary benefits of

a deep CNN, since MOSAIKS lacks intermediate layers and does not implement back-propogation. Nonetheless, some readers may find this description more intuitive, and, as mentioned in the main article, we believe that the high performance of MOSAIKS might motivate the design of CNN architectures that share some of these computational elements.

Because deep CNNs are a state-of-the-art tool for SIML tasks, we provide further comparisons of MOSAIKS performance and cost relative to this benchmark in Sections III.1 and III.2, respectively.

II.4 Data separation practices and cross-validation

We split our data into a 20% holdout test sample and an 80% training and validation sample. Within the training and validation sample, we perform 5-fold cross validation in our primary analysis, splitting the training and validation sample into 5 sets of 80% training data (64% of full sample) and 20% validation data (16% of full sample), such that the validation sets are disjoint.

II.4.1 Creating the holdout test set Before any of the label data are touched, we remove a hold-out test set that is chosen uniformly at random from the entire sample, consisting of 20% of the original data. The analysis and diagnostic procedures that follow use only the remaining 80% of the observations. The held-out test set is only used once, for the purposes of comparison to the validation set performance in Table S2. It is important to keep these data untouched until this point to ensure that our final performance results do not suffer from over-fitting.

II.4.2 Tuning hyperparameters We choose the optimal λ in Eq. (4.5) for each outcome through 5-fold cross-validation over the training and validation sample. Specifically, λ is chosen to maximize average performance (R^2) across 5 folds, from a list of candidate values.[15] For tasks with the same sampling scheme (i.e. UAR versus population-weighted sampling), the folds are consistent across tasks, so that each of the five folds comprises the same set of locations across the tasks.

II.4.3 Using cross-validation to measure model robustness In addition to being a principled way of selecting hyperparameters, cross-validation gives us a notion of how robust our model is to changes in the training and validation samples. Since each of the 5 validation sets is disjoint and randomly selected, the empirical spread of performance across folds gives us a notion of the variability of our model when applied to new data sets from the same distribution. Understanding this variation is one way of understanding the performance of our model; it gives us a notion of variance of aggregated performance (e.g. R^2 over the entire sample, for a given set of hyperparameters). A useful aspect of MOSAIKS's low computational cost of model training, however, is that it enables researchers to calculate the variance of individual predictions by bootstrapping.

[15] We choose these candidate values so as to ensure the chosen optimal λ is not the minimum or maximum of all λs supplied.

II.5 Training and testing the model

In our primary model (results shown in Fig. 4.2 of the main text) we solve for grid cell labels as a linear function of random convolutional features using a ridge regression model and a cross-validation procedure. To obtain training and validation sets, we follow the data separation practices outlined in Sec. II.4, and drop any observations with missing values. The resulting combined training and validation set sizes are N = 80,000 for forest cover, 80,000 for elevation, 54,375 for population density, 80,000 for nighttime lights, 73,102 for income, 80,000 for road length, and 58,729 for housing price.

Population density, nighttime lights, and housing price have label distributions that are approximately log-normal (Fig. S6), so we take a log transformation of the labels. We add 1 before logging to avoid dropping labels with an initial value of zero (see Section II.6.2 for performance in logs and levels for all tasks).[16]

With these labels and features in hand, we regress each outcome y_ℓ^s for each task s on features $\mathbf{x}_\ell$ as follows:

$$y_\ell^s = \mathbf{x}(\mathbf{I}_\ell)\beta^s + \epsilon_\ell^s \tag{4.4}$$

which is the same as Eq. (4.2) from the main text. We solve for β^s by minimizing the sum of squared errors plus an l_2 regularization term:

$$\min_{\beta^s} \frac{1}{2}||y_\ell^s - \mathbf{x}(\mathbf{I}_\ell)\beta^s||_2^2 + \frac{\lambda^s}{2}||\beta^s||_2^2 \tag{4.5}$$

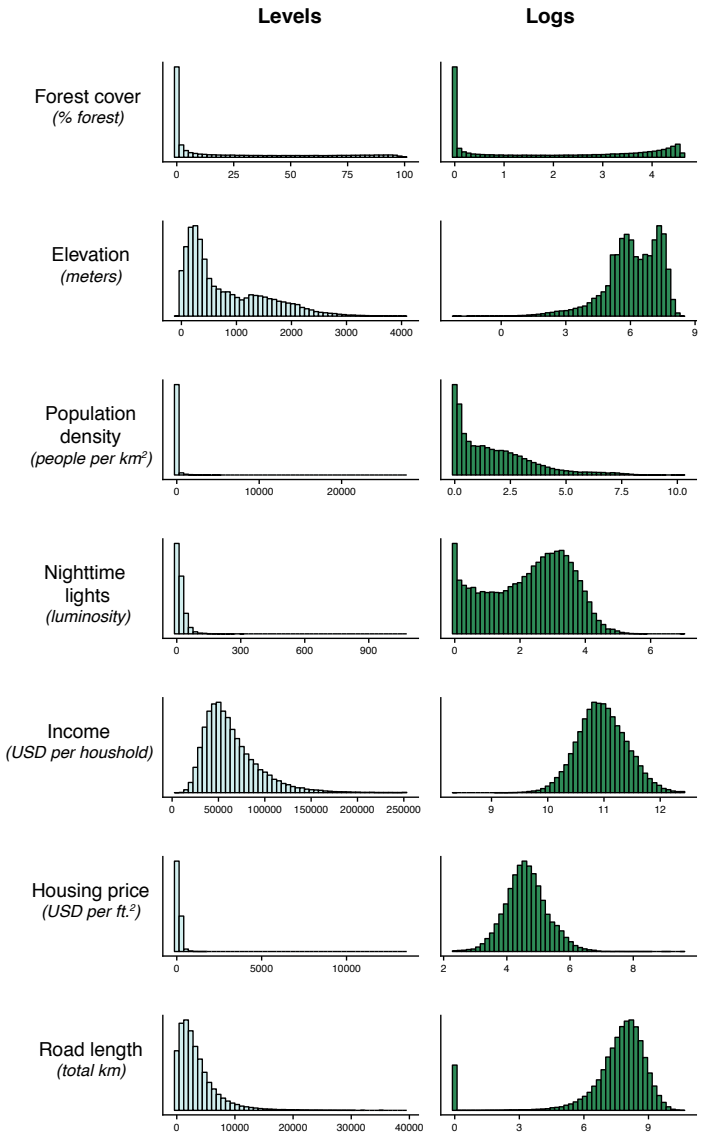
We use ridge regression across all outcomes to demonstrate the generalizability of using a single set of image features across many simple regression models. Further, this standardized methodology facilitates comparison of performance and sensitivity across tasks. We note that other modeling choices could potentially improve fit (e.g. using a hurdle model for zero-inflated outcome distributions such as road length); we leave such task-specific explorations for future research.

In visual display of results and calculation of performance metrics such as R^2, we clip our predictions for each task at the minimum and maximum values observed in the labeled data.

The resulting weights (i.e. regression coefficients) $\hat{\beta}^s$ obtained from estimation of Eq. (4.4) indicate, along with the variance of the features, which features k (derived from random patch $\mathbf{P}_k$) capture meaningful information for prediction in each task. Fig. S7 demonstrates that the recovered weights are stable across cross-validation folds within a task. The first two columns show standardized weights that are estimated from disjoint training and validation splits for the same task.[17] Values corresponding to each axis are the regression weights estimated when the corresponding fold composes the

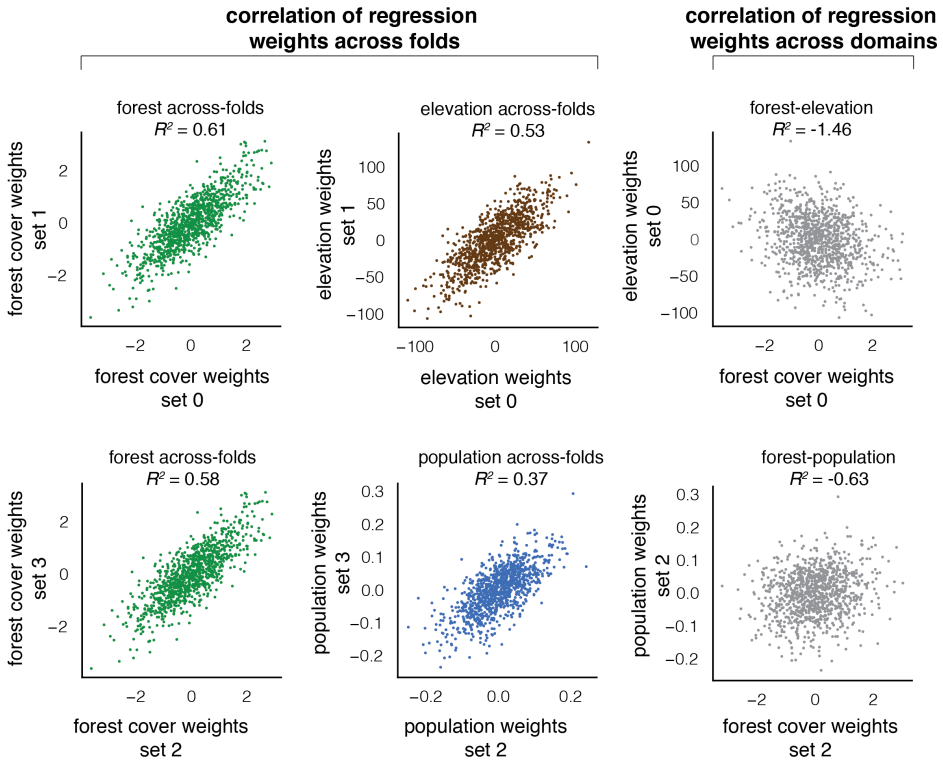
[16] Since housing price per square foot is always positive, for that variable we use just a log transformation.

[17] For consistency across comparisons, R^2 is calculated on standardized regression weights, which have been demeaned and divided by their standard deviations. The number of random features is set to $K = 1,024$ for visual display purposes.



Extended Data Figure S6: Distribution of outcome variables in levels and logs. Histograms show the distribution of each outcome variable over all sampled image grid cells (approximately 100,000 observations, depending on data availability). Forest cover, elevation, and population density are sampled uniform at random across the continental US, while all other variables are randomly sampled with population weighting. The first column shows the distribution in levels, and the second in logs. For elevation, population density, nighttime lights, and road length, logs were taken after adding 1 to the raw values, given the propensity of zero values in these outcomes.

training set. High R^2 values indicate a strong correlation between regression weights from *different folds within a single task* (forest cover, elevation, and population density are shown), demonstrating that similar linear combinations of features are selected by the regression model, even when the sample of training images changes. This suggests



Extended Data Figure S7: Regression weights across folds within a task vs. across tasks within a fold. All scatterplots indicate regression weights for forest cover, elevation and/or population density. Each point depicts estimated coefficent values for the kth feature (β_k^s) when trained on either different samples or different labels. In the across-fold examples (first two columns), we learn weights for disjoint training and validation splits for the same task via cross-validation in which one fold acts as the training set and the other as the validation set. Values corresponding to each axis are the regression weights when that fold is the training set (e.g. the top left scatter shows $\{\beta_k^{forest1}, \beta_k^{forest0}\}$), and indicate a strong correlation across regression weights from different folds. In the across-task examples (last column), regression weights are shown for the same training and validation sets for two distinct tasks (e.g. the top right scatter shows $\{\beta_k^{elevation0}, \beta_k^{forest0}\}$). We see that there is virtually no correlation in regression weights across tasks, demonstrating that predictions across tasks lie in orthogonal subspaces of the feature space. Across all examples here, we set the number of random features to $K = 1,024$. For consistency across comparisons, R^2 is calculated on standardized regression weights, which have been demeaned and divided by their standard deviations.

that specific sets of patches consistently contain valuable information in predicting outcomes for a specific task. However, different combinations of patches are useful for different tasks, and we find no correlation in the weights recovered *between tasks*. For example, in the last column of the figure, we show that regression weights that are

recovered for forest cover and elevation (top right) are essentially orthogonal as are regression weights recovered for forest cover and population (lower right). In these two plots, regression weights are shown for the same training and validation sets, but for two distinct tasks. Sets of features that are relevant for prediction in one task appear to be irrelevant for another, as there is virtually no correlation in regression weights.

II.5.1 Intuition The consistency of weights recovered in MOSAIKS across folds within a task, and the orthogonality of weights recovered within a fold but across tasks, provides some intuition for why MOSAIKS provides consistent results and generalizes across a very large (potentially infite) number of potential tasks. The rich featurization $\mathbf{x}(\mathbf{I}_\ell)$ locates image $\mathbf{I}_\ell$ in a very high-dimensional (K-dimensional) feature space. Solving for β^s in Eq. (4.4) then identifies the K-dimensional vector β^s that points in the direction of most rapid ascent (the gradient vector) for labels y^s, when the position of images $\mathbf{x}(\mathbf{I}_\ell)$ are projected onto this vector. Because the feature space is so large — our baseline model has an 8,192-dimensional feature space — there are a vast number of orthogonal gradient vectors that can be drawn through this space along which images can be organized for different tasks. The left and center panels of Fig. S7 illustrate that similar K-dimensional gradient vectors β^s are selected when solving for the same task but using different samples (each point depicts an element of the vector β^s). The right panels shows that for different tasks, the gradient vectors are orthogonal and point in completely unrelated directions in the feature space. This orthogonality means that predictions $\hat{y}^s$ for different tasks will be independent of one another, even though both are constructed as linear combinations of the same set of features.

II.6 Primary model test set performance, robustness to functional form, and spatial distribution of errors

Here, we describe how we test for overfitting to the training and validation set in our primary model, test for primary model performance robustness to alternative functional forms, and characterize the spatial distribution of primary model error.

II.6.1 Performance in a holdout test set To test for overfitting, we evaluate the performance of our primary model on a randomly sampled 20% holdout set. These data were never used for model selection and were only touched at the end of our analysis to check for overfitting. To conduct this test, for each outcome, we use cross-validation within the training set to determine the outcome-specific optimal λ. We then retrain the model on the full training set using this optimal λ, and evaluate this model on the holdout test set. We find that performance in the test set is nearly identical to that of the validation set (Table S2), which indicates that our models were not overfit to the data. For some performance metrics, such as the maps in the main text, we present validation set performance (instead of the test set) because the sample is larger and the performance is unchanged.

II.6.2 Robustness of model to alternative functional forms Throughout the main text, we report primary model performance in each task from a model estimated with labels that are either logged (e.g. population density), or in levels (e.g. forest cover). The decision regarding functional form for each task was made based on the

Task	Cross-validation R^2	Test set R^2
Forest cover	0.91	0.91
Elevation	0.68	0.68
Population density	0.73	0.72
Nighttime lights	0.85	0.84
Income	0.45	0.45
Road length	0.52	0.53
Housing price	0.46	0.47

Extended Data Table S2: Model performance in the hold out test set. For each outcome, we use 5-fold cross-validation within the training/validation set using 80% of our labeled data to optimally select task-specific hyperparameters in ridge regression (i.e. λ). We then retrain each model on the full training set using this optimal λ. Performance on the validation set (column 1) is compared to that of the held out test set (column 2).

underlying distribution of labels across our image grid cells. Many outcomes, such as housing prices, display exceptionally skewed distributions that approximate log-normality (see Fig. S6). For these outcomes, we take the natural log of the image grid cell values in model training and testing. Table S3 shows model performance for all tasks under both the levels and logs functional forms.[18] Tasks with highly skewed distributions, such as population density, housing price per square foot, and nighttime lights have substantially higher performance (R^2 increases by 10-64%) after being logged. Tasks whose labels display much less skew in levels, such as road length, income, and elevation show small to modestly reduced performance (4-21%) when their outcomes are modeled in logs.

II.6.3 Spatial distribution of errors Fig. S8 shows the distribution of errors over space, for the model predictions presented in Fig. 4.2. The model systematically over-predicts low values and under-predicts high values across all tasks. This is likely due to our choice of ridge regression, which favors predictions that tend toward the mean due to the ℓ_2 penalty. The structured correlation of errors across space suggests that there is substantial room for model improvement, potentially from including task specific knowledge. For example, our models of housing price and elevation could likely, respectively, be improved by adding in information about school districts –to address clustering of house price error in parts of big cities – or location – to help identify large areas of high elevation such as the Rocky Mountains. We recognize that there exists substantial room for task-specific model performance, which we leave for future research. Further, discontinuities in the error structure over political boundaries can help identify inconsistency in label quality. For example, the sharp increase in road

[18]In tasks where negative values or zeros are present (e.g. forest cover, elevation, and nighttime lights), we drop negative values and add one to zero values before taking logs for this test.

Task	Log model R^2	Levels model R^2
Forest cover	0.89	**0.91**
Elevation	0.59	**0.68**
Population density	**0.73**	0.52
Nighttime lights	**0.85**	0.77
Income	0.43	**0.45**
Road length	0.41	**0.52**
Housing price	**0.46**	0.28

Extended Data Table S3: Model performance across tasks and functional forms. All R^2 values indicate performance using the optimal hyperparameter λ after 5-fold cross-validation. In the log model, the outcome variable is defined as the natural logarithm of the original labeled data (e.g. natural log of the average forest cover over an image gridcell). In the levels model, the outcome variable is simply the level of the aggregated labeled data, as defined in Section II.2. Values in bold are reported in the main text.

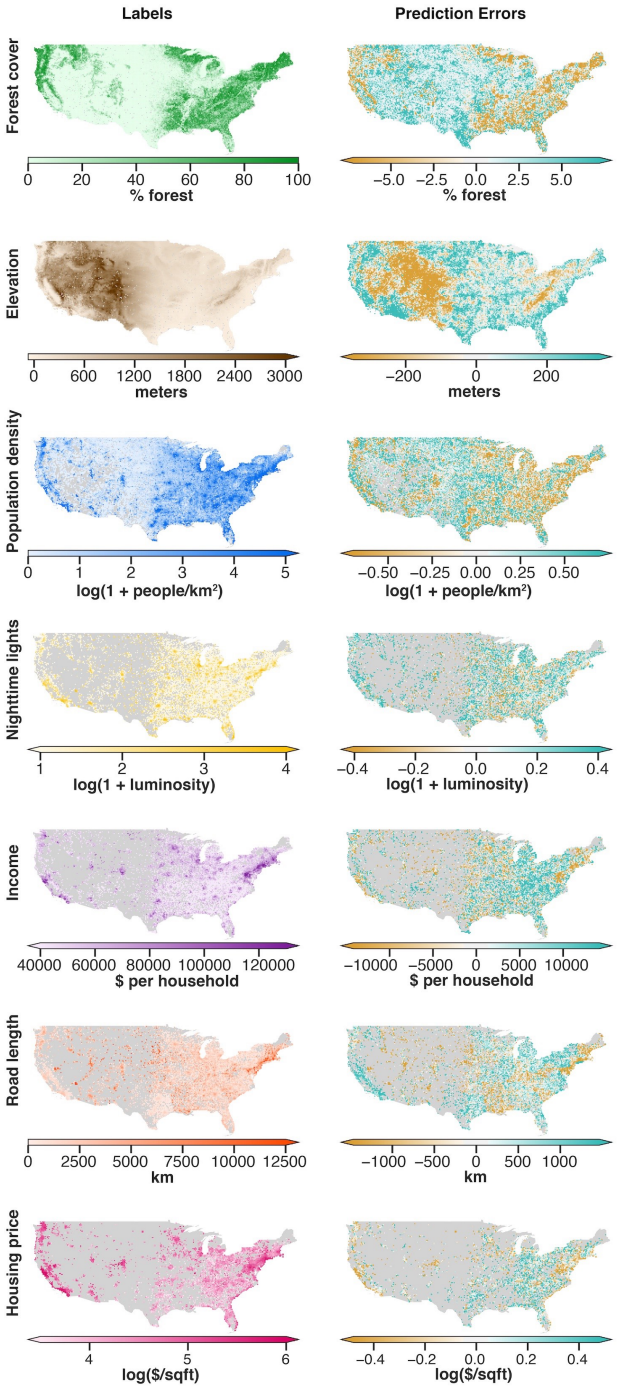
length prediction error moving across the border from Louisiana to Texas suggests that the raw data labeling in these two states may differ methodologically, which introduces error into the label, and in turn, the model.

II.7 Altering the number of features and training set size

To better understand factors that could improve the primary model, we test the sensitivity of its performance to the number of features and the training set size (results shown in Fig. 4.3 in the main text). Understanding the returns to additional features and observations enables better optimization of model performance given cost constraints.

Since features in MOSAIKS are generated randomly, there is no theoretical reason to select a specific number of features. To test the sensitivity of the primary model performance to the number of features, we train a model identically to our primary specification (Section II.5) except that we vary the number of features across the values $\{100, 200, 500, 1000, 2000, 4096, 8192\}$ (Fig. 4.3A). For each set of features and each task, we conduct 5-fold cross-validation to recover the optimal hyperparameter λ.

Notably, using only 100 features recovers a substantial amount of the variation across tasks. Of the tasks, the least variation is recovered for income (R^2 using 100 features is 81% of R^2 using 8,192 features) and the most variation is retained in nighttime lights (R^2 using 100 features is 96% of R^2 using 8,192 features). This suggests that in computation or memory-limited settings, fewer features could be used with only minor losses in performance. On the other hand, even with 8,192 features, performance does not fully flatten out (on a logarithmic scale). This suggests that performance could be improved further by increasing the number of features past $K = 8,192$. At the limit of our testing, a doubling of K from $4,096$ to $8,192$ led to a largest performance increase



Extended Data Figure S8: Labels and prediction errors over space for each task. Left maps: 80,000 observations used for training and validation, aggregated up to 20km x 20km cells for display. Right maps: prediction errors from concatenated validation set estimates from 5-fold cross-validation for the same 80,000 grid cells, identically aggregated for display.

of 0.026 R^2 for income and a smallest of 0.010 R^2 for forest cover.

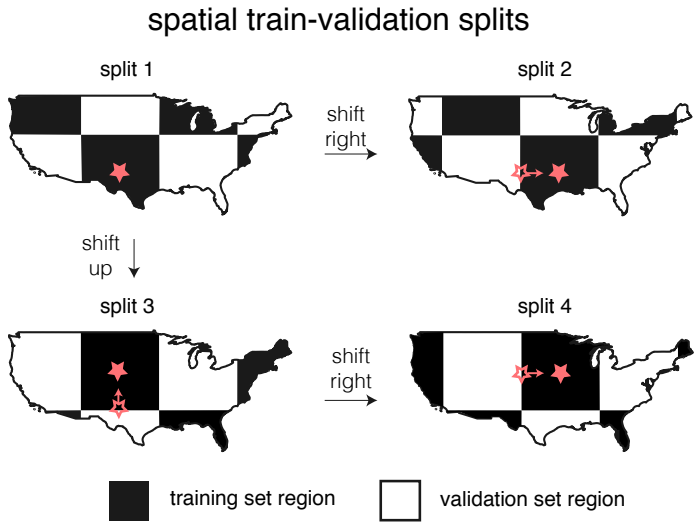
To test the sensitivity of primary model performance to the number of training samples, we train a model identical to our primary specification (with 8,192 features) except with a varying size of training set (from 500 to 64,000 images) (Fig. 4.3A).[19] In cases where the training set has fewer than 64,000 total observations due to missing data (e.g. population density, income, road length and housing price), we use the full training data set to construct our largest training sample.

Similarly to increasing the number of features, increasing the training set size increases model performance with diminishing marginal returns. Notably, models trained on only 500 observations recover at minimum 56% (road length) of performance relative to $N = 64,000$ and at maximum 87% (forest cover), excluding income and housing price, which require larger samples to attain performance. This suggests that, for all but the most difficult SIML tasks, MOSAIKS may be useful even when label collection is very costly. For the tasks with the best R^2 performance (forest cover, nighttime lights), performance plateaus out as the number of training observations approaches $64,000$. However, for the remaining five tasks, these results show that more training data could substantially increase performance further. The range of performance gain from increasing $N = 32,000$ to $64,000$ is bounded below by forest cover (.005 R^2) and above by road length (.027 R^2).

II.8 Testing generalizability across space and comparison to kernel-based interpolation.

To understand the ability of our model to predict outcomes in large contiguous regions with no ground truth, we design an experiment where we evaluate models using training and validation sets that are increasingly far away from each other in space. Specifically, we iteratively create a grid over the US with a side length of δ degrees and then use this grid to divide the training and validation dataset ($N = 80,000$) into spatially disjoint sets of roughly equal size. We create these disjoint sets by assigning observations that lie in every other box within the grid to the train set and test set, respectively, creating a checkerboard pattern with the train set and test set, as shown in Fig. 4.3B. We vary the width δ of each square in the grid range across the values of $\{0.5, 1.5, 2, 4, 6, 8, 10, 12, 14, 16\}$ degrees (roughly 40 to 1400 km) in sequential runs of the experiment. As δ increases, validation set observations become on average farther away from the training set points. This distance makes prediction on the validation set more difficult, because observations in the validation set are now likely to be less similar to those in the training set. We learn the model on the training set using ridge regression. To assess the stability of this performance, we offset the checkerboard and re-run the above analysis four times – once in the original location and then three more times – shifting the grid up, right, and both up and right by half the width of the grid (see Fig. S9). The ℓ_2 regularization term, λ, is selected to maximize average performance in the four validation sets, as we would select in in a standard cross-validation procedure.

[19]The same per-fold validation sets are used for each iteration of this analysis as well as for the primary analysis and for the test of model performance sensitivity to the number of features.



Extended Data Figure S9: Illustration of the procedure to systematically shift train and validation sets in space when assessing the performance of MOSAIKSover regions with no ground-truth data. To assess the ability of MOSAIKS to generate meaningful predictions when extrapolating across large spatial distances, we conduct a "checkerboard" experiment (Section II.8, Fig. 3B-C of the main text) in which the training set ("black squares") and validation set ("white squares") are separated by increasingly large distances. The length of a square in each experiment is δ, measured in degrees. This figure demonstrates the four different train/validation splits that are created by shifting a given spatial checkerboard (split 1) by $\delta/2$ to the right (split 2), $\delta/2$ up (split 3), and both simultaneously (split 4).

The performance plotted in Fig. 4.3C is the performance on the the resulting validation sets. We find that across most tasks, performance degrades only slightly as the distance between training observations and testing observations increases. This suggests that MOSAIKS is indeed learning image-label mappings that transfer across spatial regions.

II.8.1 Comparison of MOSAIKS to kernel-based spatial interpolation In these experiments we demonstrate that MOSAIKS outperforms spatial interpolation (or extrapolation, depending on geometry) – a commonly used simple technique to fill in missing data (Fig. 4.3C). This suggests that MOSAIKS, and SIML generally, exploits the spectral and structural content of information within an image to generate predictions at national scale that extend beyond what can be captured by geographic location alone.

We compare MOSAIKSto kernel-based spatial interpolation using a Gaussian Radial Basis Function (RBF) kernel, a simple and general widely used approach. In this approach, the value for a point in the validation set at location $\ell_v \in \mathbb{R}^2$ is predicted to

be a weighted sum of the values of all the points in the training set ℓ_t, as follows:

$$\hat{y}_v^s = \frac{\sum_{\ell_t \in [\text{Train}]} y_t^s w(\ell_t, \ell_v)}{\sum_{\ell_t \in [\text{Train}]} w(\ell_t, \ell_v)}; \qquad w(\ell_t, \ell_v) = e^{-\frac{1}{2\sigma^2}\|\ell_t - \ell_v\|^2}$$

Here, w is the weight assigned to each observation in the training set based on kernel values that are indexed to distance, such that w decreases as the distance between the point being predicted and the point in the training set increases. We select σ – the parameter that determines the rate at which w degrades with distance – to maximize average performance on the validation set across all four spatially-offset runs, similar to how we tune λ in the spatial extrapolation experiment described above. The optimal value of the bandwidth parameter σ will depend on the task at hand, as well as the average distance from points in the validation set to points in the training set. To ensure comparability, spatial interpolation based predictions and performance are computed for the exact same samples as used for MOSAIKS in each checkerboard partition.

II.9 Super-resolution

As discussed in the Methods section of the main text, the featurization method in MOSAIKS exploits the fact that many image-level outcomes of interest are linearly decomposable across sub-image regions. This is done by creating image-level features that are averages of statistics from all sub-image regions. Because these features are ultimately used in linear regression (i.e. Eq. (4.2)), a natural property of this approach is that weights estimated from Eq. (4.2) can be used not only to generate predictions of outcome variables at the image-scale, but also at the scale of any sub-image region. As satellite imagery are available at increasingly high spatial resolution, this "super-resolution" property is both practical and powerful, enabling researchers to generate novel predictions at higher resolution than available ground truth data.

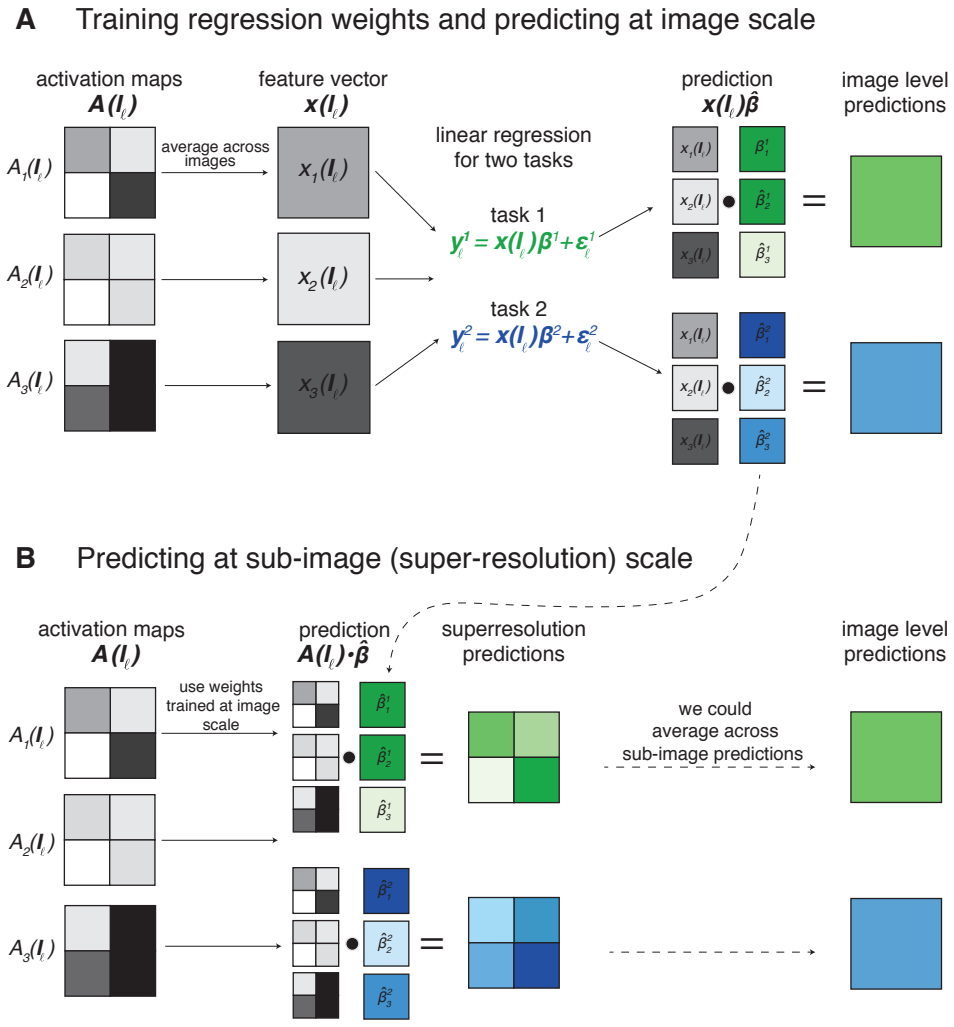
This section gives mathematical justification for a simple method to use MOSAIKS to predict outcomes of interest at a finer resolution than available labeled data. We display the super-resolution properties of MOSAIKS visually, and quantitatively document the empirical performance of this super-resolution approach.

Why MOSAIKS naturally achieves super-resolution for label predictions

Given an image-label pair $\{\mathbf{I}_\ell, y_\ell^s\}$, the goal of super-resolution is to resolve which sub-regions of the image $\mathbf{I}_\ell$ contribute to high or low values of y_ℓ^s. Recall that for image $\mathbf{I}_\ell$, feature vector $\mathbf{x}(\mathbf{I}_\ell)$ is a K dimensional vector, where each scalar element $\mathbf{x}_k(\mathbf{I}_\ell)$ of $\mathbf{x}(\mathbf{I}_\ell)$ is an average across the pixels of the image of the values obtained by convolving sub-regions of the image with patch $\mathbf{P}_k$. As in Section II.3, denote by $\mathbf{X}$ the full random feature matrix in $\mathbb{R}^{N \times K}$, so that $\mathbf{X}_{\ell k}$ denotes the k^{th} element of the feature vector describing image $\mathbf{I}_\ell$. By Eq. (4.3), we can decompose the feature elements as:

$$\mathbf{X}_{\ell k} := \mathbf{x}_k(\mathbf{I}_\ell) = \frac{1}{254^2} \sum_{i=1}^{254} \sum_{j=1}^{254} \mathbf{A}_k(\mathbf{I}_\ell)[i, j]$$

where $\mathbf{A}_k$ is the activation map associated with patch $\mathbf{P}_k$. Since we're using a linear model to form predicted values, we can trace these values back to subregions of the



Extended Data Figure S10: Illustration of the procedure to construct predictions at image resolution and super-resolution. Panel A illustrates the standard MOSAIKS prediction pipeline. After convolution with random patches, nonlinear activation maps $\mathbf{A}_k(\mathbf{I}_\ell)$ are averaged across images to construct a set of image-level features $\mathbf{x}_k(\mathbf{I}_\ell)$ used in linear regression to generate predictions at image-scale (Section II.3). Panel B illustrates how the weights trained using labels and features at image-scale in panel A can be used to generate predictions at resolutions higher than the images and labeled data, achieving predictions at super-resolution. The scalar product of the entire activation map $\mathbf{A}_k(\mathbf{I}_\ell)$ and the estimated weights vector $\hat{\beta}$ generates super-resolution predictions at any desired sub-image scale larger than pixel-level. The last column of panel B illustrates the fact that super-resolution predictions, when averaged across an image, are identical to predictions generated from the standard process in panel A.

original image. When we perform a linear regression for task s, the resulting regression weights are a vector $\hat{\beta}^s \in \mathbb{R}^K$ such that the scalar $\hat{\beta}^s_k$ describes the relative weight of feature k in the image-resolution label predictions. The prediction of outcome s using image $\mathbf{I}_\ell$ thus decomposes as:

$$\begin{aligned}
\hat{y}^s_\ell &= \mathbf{X}_\ell \hat{\beta}^s \\
&= \sum_{k=1}^{K} \mathbf{X}_{\ell k} \cdot \hat{\beta}^s_k \\
&= \sum_{k=1}^{K} \left(\frac{1}{254^2} \sum_{i=1}^{254} \sum_{j=1}^{254} \mathbf{A}_k(\mathbf{I}_\ell)[i,j] \right) \cdot \hat{\beta}^s_k \\
&= \frac{1}{254^2} \sum_{i=1}^{254} \sum_{j=1}^{254} \underbrace{\left(\sum_{k=1}^{K} \hat{\beta}^s_k \cdot (\mathbf{A}_k(\mathbf{I}_\ell)[i,j]) \right)}_{\text{super-resolution prediction}}
\end{aligned}$$

where the third line follows from substituting $\mathbf{X}_{\ell k}$ according to Eq. (4.3). Therefore, we can associate with each pixel indexed by (i, j) a predicted super-resolution value:

$$\hat{y}^s_{\ell,(i,j)} = \sum_{k=1}^{K} \hat{\beta}^s_k \cdot (\mathbf{A}_k(\mathbf{I}_\ell)[i,j]) \tag{4.6}$$

which is that pixel's predicted label value, and thus its contribution to the overall predicted image-level label value $\hat{y}_\ell$ for $\mathbf{I}_\ell$. These pixel-level predictions can be average-pooled to larger sub-image scales as shown in Fig. 4.4B. If averaged over the entire image, the standard full-image prediction $\hat{y}^s_\ell$ is recovered. The procedure to construct super-resolution predictions, and a comparison of it to the precedure to construct impage-level predictions, is illustrated in Fig. S10.

Fig. S11 demonstrates empirical performance of Eq. (4.6) using ten examples of this approach at super-resolutions on both the forest cover and population density outcomes. The ten images were randomly selected from the union of observations with forest cover $> 10\%$ and population density > 100 people/km^2 to ensure that all images considered had a non-negligible value for each variable.[20]

In our formulation, super-resolution predictions are easily estimable during featurization. Consider again the per-pixel contributions of Eq. (4.6). An alternative way to express this is

$$\hat{y}^s_{\ell,(i,j)} = \left(\sum_{k=1}^{K} \hat{\beta}^s_k \cdot \mathbf{A}_k(\mathbf{I}_\ell) \right)[i,j]$$

That is, super-resolution estimates are just a linear combination of the activation maps $\mathbf{A}_k(\mathbf{I}_\ell)$ weighted by $\hat{\beta}^s_k$ (see Fig. S10). Every time we featurize a new image $\mathbf{I}'_\ell$, we

[20] To ensure that weights decomposed as a sum, as in Eq. (4.6), we used level values (i.e. not log-transformed) for population density labels in Fig. S11.

must perform the step of computing the K activation maps $\{\mathbf{A}_k(\mathbf{I}'_\ell)\}_{k=1}^K$ (Fig. S4 D). Therefore, if we already have a suitable regression weight vector $\hat{\beta}^s$ for task s, for any new images $\mathbf{I}'_\ell$ that we featurize, we can compute the super-resolution predictions $\sum_{k=1}^K \hat{\beta}_k^s \cdot \mathbf{A}_k(\mathbf{I}_\ell)$ as weighted combinations of the activation maps at negligible additional cost, prior to pooling, in the existing featurization pipeline.

Evaluating super-resolution performance

To systematically evaluate the ability of MOSAIKS to accurately predict outcomes at super-resolution, we evaluate the within-image label variation that MOSAIKS's super-resolution predictions accurately explain. We use forest cover for this test because the raw label resolution is substantially finer than the grid cell used to construct labels (see Section II.1 and Fig. S3), so we are able to attach "true" labels to super-resolution predictions within each image. In our main analysis, we construct grid cell forest cover labels by averaging fine-resolution raw forest cover data (see Section II.2). Here we leverage the fine resolution of the raw data to compare super-resolution performance of a model trained on aggregated labels but tested on high-resolution raw forest cover data.

Specifically, we take a randomly drawn subset of $N = 16,000$ grid cells from the U.S. UAR grid as our super-resolution dataset. We solve for the regression weights $\hat{\beta}^s$ using ridge regression (with $\lambda = 1e3$[21] Note that these weights $\hat{\beta}^s$ are trained on aggregate labels averaged over the grid cell as shown in Fig. S3). We then use the super-resolution prediction technique described above (Eq. 4.6) to get 254×254 pixel-level predictions as in Fig. S10B, using the weights derived from image-level labels, as in panel A of the same figure. These pixel-level predictions can then be aggregated to any super-resolution, where increasing aggregation (lower super-resolution) reduces noise in the predictions at the cost of lower resolution.

We assess the performance of super-resolution at a variety of scales by calculating the percent of the variance of the raw within-image forest cover labels that can be explained by the super-resolution predictions at each scale. For example, to assess the performance of $2\times$ super-resolution predictions, we average predictions from the 254×254 super-resolution predictions by quadrants, resulting in four predicted values (twice the original resolution). We perform the same per-quadrant average for the raw fine-resolution forest cover labels. We demean both the within-image predictions and labels to eliminate any across-image variation, thereby focusing this test on the ability of the predictions to explain residual within-image variation. We then concatenate these within-image predictions and labels across the $N = 16,000$ images, so that the resulting R^2 value reported is the percent of super-resolution label variance explained by super-resolution predictions, across $64,000 = 16,000 \cdot 2^2$ label-prediction pairs.

The resulting performance of super-resolution predictions at different scales is shown in Fig. 4.4C for width scales of $2\times, 4\times, 8\times, 16\times$, and $32\times$. A width of size w results in w^2 predictions per image. We go up to $w = 32$ because the native width of the forest cover labels ($\sim$ 30m) is just under 1/32 the width of the original image ($\sim$ 1km).

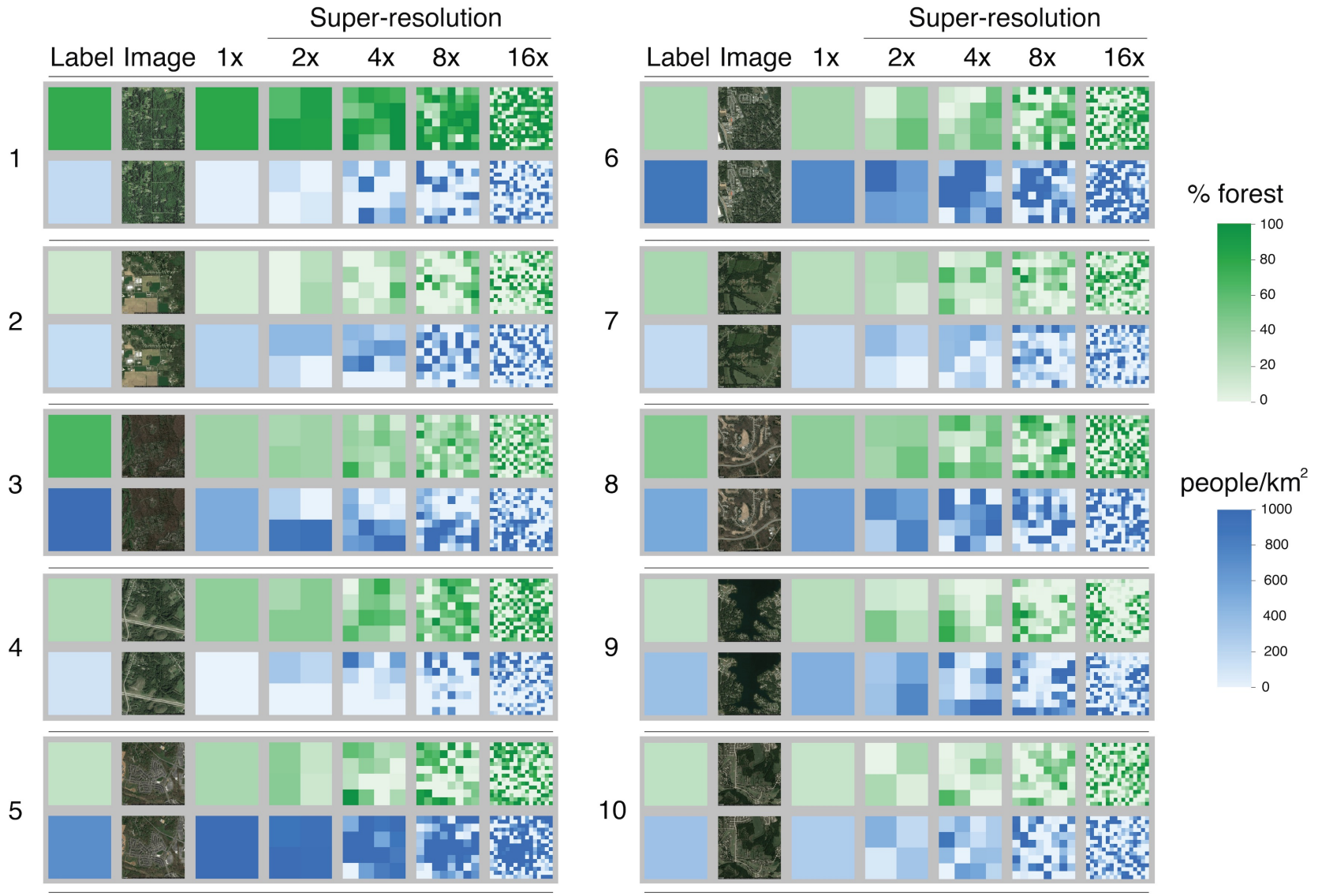
[21] While setting λ to zero would recover the per-image predictions exactly, we found that having a higher regularization value resulted in more stable per-pixel value attributions.

Super-resolution predictions are trained only on the aggregate label at the image-level, as in Eq. (4.2). Nonetheless, as Fig. 4.4C shows, we are able to explain over 31% of the within-image label variations at 2× super-resolution, and over 18% of the variation using 4× super-resolution grids. This performance degrades for super-resolution widths $\geq 16\times$, which are closer to the (hidden) native resolution of the labels themselves.

Comparisons to other within-image prediction algorithms The most similar approach in the literature to MOSAIKS's super-resolution predictions are methods specifically designed for pixel-level classification, or *semantic labelling* of satellite imagery [26, 123]. However, these approaches make use of sub-image labels for training, as opposed to our setting, where only one label per image (per task) is provided. Such semantic labelling approaches tend to use a downsample-then-upsample approach inspired by auto-encoders [121] to learn lower-dimensional latent representations which are then up sampled to image-size prediction maps from which per-pixel classifications can be made. The upsampling procedure introduces more parameters to be tuned during model training, as well as additional computational cost in producing predictions. We again contrast this complex machinery with the simplicity of MOSAIKS 's approach, which calculates super-resolution predictions as a weighted sum of activation maps.

Conditions where super-resolution is most easily interpretable The linear decomposition of Eq. (4.6) holds when using labels that represent the average or sum of values within a grid cell, such as forest cover, elevation, population density, nighttime lights, income, or road length. However, it does not hold exactly when values are transformed nonlinearly after aggregation (e.g. $\log(\sum y) \neq \sum \log(y)$).[22] In these cases, the interpretation of super-resolution estimates requires care. Another case in which the interpretation of the sub-image predictions is difficult is when an image-level characteristic is not directly the sum of sub-image parcels. For instance, when predicting mean housing price in a grid cell, a manicured park might contribute to a higher value, yet that component of the image does not, in itself, have any associated housing price. In this case, we we would interpret the sub-image predictions as "contributions to grid cell mean housing price" rather than the more natural interpretation as simply "a finer resolution prediction of housing price."

[22] This issue could be addressed – in the case of logged variables – if one obtained a geometric mean image-level outcome rather than an arithmetic mean.



Extended Data Figure S11: Super-resolution performance across ten randomly selected images. Each set of images indicate the image-level labels (column 1), the image itself (column 2) and predicted outcomes from MOSAIKS at increasing levels of super-resolution (columns 3-7). These ten examples were selected uniformly at random from images in which our labels indicated at least 10% total forest cover and at least 100 people/km^2.

II.10 Global model

For our global analysis, we create a global grid, composed of roughly 420 million cells just over 1km^2 in size, using an identical structure to that described in Section II.1 for the US. To obtain observations for our global analysis, we sub-sample 1,000,000 cells from this grid, sampling UAR from non-marine grid cells. These relatively sparse sampling of global data is simply due to the cost of obtaining imagery data.

One of the difficulties in sub-sampling from the global grid is that there are many grid cells where no Google imagery is available (there are negligibly few missing images in the US grid). After discarding grid cells with missing imagery from our original sample of 1,000,000 observations, we are left with $N = 556,025$ valid observations which we use to train/validate (80%, $N = 444,820$) and test (20% $N = 111,205$) the model.

When generating features ($K = 2,048$) for our global model, we conduct featurization as described in Section II.3. Note that we use patches drawn randomly from the *global* sample of images, not just from within the US.

II.10.1 Performance of the global model in the continental US In Figure 4.4 of the main text, we demonstrate the ability of MOSAIKS to scale globally in four of our tasks where global labels exist (forest cover, nighttime lights, population density, and elevation). We show in the main text how MOSAIKS performs when trained on this relatively sparse sample of global images and labels. However, for researchers focused on a particular region of the world, a model trained with more densely sampled data in that region (as in our US analysis) is likely to perform much better than the sparsely-sampled globally-trained model shown in the main text.

To demonstrate this, we contrast the performance of the globally-trained model within the continental US only, with that of our primary specification that was trained and tested in the US using a much more densely sampled set of labels and a sampling scheme well-suited to each task (i.e. nighttime lights sampling is population weighted). Results are shown in Table S4). The globally-trained model was trained on just $N = 18,414$ US observations for all four globally-available tasks. This contrasts with $N = 80,000$ for forest cover, elevation, nighttime lights, and road length, and 54,375 for population density, 73,102 for income, and 58,729 for housing price in the US-only model. Table S4 makes clear that while the global model has substantial explanatory power within the sparsely-sampled US, there are gains from focusing model training and data sampling to the region of interest for a particular application.

III Comparisons to other models

Here, we compare the performance and computational cost of MOSAIKS to other approaches in the literature.

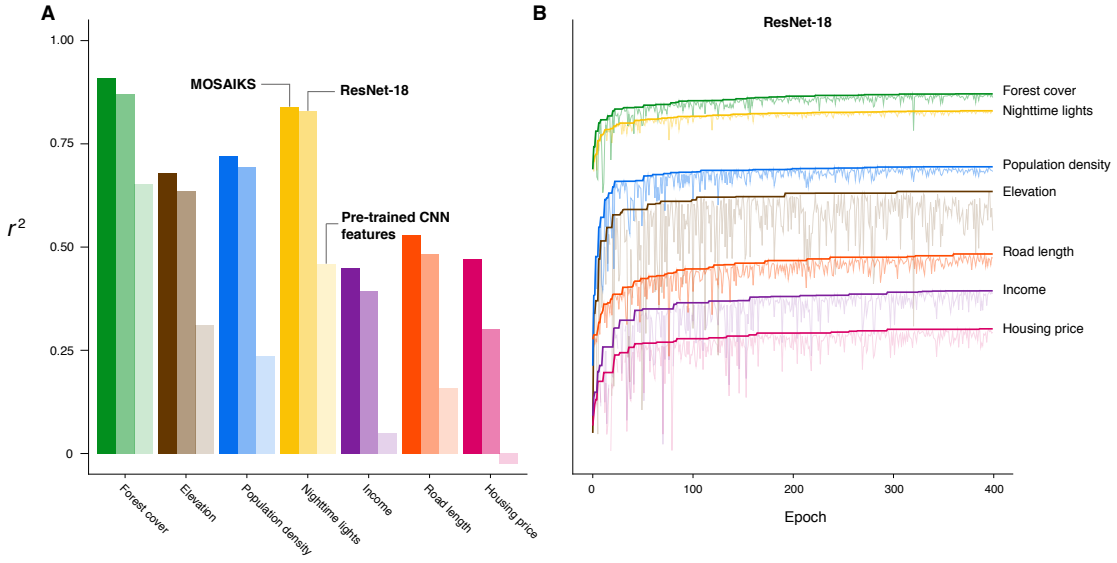
III.1 Benchmarking performance

Convolutional neural networks (CNNs) have become the default "gold standard" in many image recognition tasks [59], and are increasingly used in remote sensing applications [52, 123, 29, 80, 91, 79, 53, 137, 49, 63]. Simultaneously, alternative generalizable and efficient pipelines have been developed incorporating unsupervised featurization

Task	**US model** N=80,000* K=8,192	**US model** N=18,414 K=8,192	**US model** N=80,000* K=2,048	**US model** N=18,414 K=2,048	**Global model in US** N^{US}=18,414 K=2,048
Forest cover (R^2)	0.91	0.90	0.88	0.88	0.67
Elevation (R^2)	0.68	0.64	0.63	0.60	0.24
Population density (R^2)	0.73	0.72	0.69	0.69	0.48
Nighttime lights (R^2)	0.85	0.83	0.82	0.81	0.66

Extended Data Table S4: Model performance in the continental US using a model trained within the US versus one trained globally with sparse US data sampling. The main text shows performance with a global model trained on substantially less densely sampled observations ($N^{\text{Global}} = 556,025$; $N^{\text{Global in US}} = 18,414$) and number of features ($K = 2,048$) than in the model trained within the continental US ($N = 80,000$, $K = 8,192$). This table compares performance across models trained and tested within the US (columns 1 – 4) with the model trained on a global sampled and tested on the US (column 5). Column 1 uses the full sample of observations and number of features shown in the main text. Column 2 degrades the sample by limiting N, retraining the model within the US using the same number of observations as fall within the US in the sample used to train the global model. Column 3 degrades the sample by limiting K, retraining the model within the US using the same number of image features as used to train the global model. Column 4 degrades the sample by limiting both N and K, retraining the model within the US using the number of observations and features used to train the global model. Column 5 shows the performance of the global model within the continental US. Each column displays R^2 values indicating performance using the optimal hyperparameters after 5-fold cross-validation. Note that the global model relies on substantially poorer data quality than the model trained within the US and that nighttime lights is sampled with population weights in the US model but not in the global model. *N=80,000 for all tasks except population density, where $N = 73,102$.

and/or a classification or regression algorithm [17, 18, 79, 50, 81]. MOSAIKS is low-cost and generalizable like these latter models but – unlike these other models – it offers performance for regression problems similar to that of leading CNN architectures. Here we quantitatively assess the performance of MOSAIKS relative to (a) a CNN trained end-to-end with the outcomes of interest and (b) a similarly cheap, unsupervised featurization used in place of random convolutional features in the MOSAIKS infrastructure. For (b), we use the features generated by the last hidden layer of a pre-trained variant of the CNN (trained on natural imagery). This common approach is unsupervised in that the weights of the CNN are not trained using the labels of the outcome of interest, and such an approach has been shown to have better predictive performance than many other unsupervised featurization algorithms (e.g. GIST, SIFT, Bag of Visual Words) on satellite image tasks [17].



Extended Data Figure S12: Comparison of test accuracy between MOSAIKS, ResNet, and a regression model using features from a pre-trained CNN. Panel A shows task-specific performance of MOSAIKS (dark bars), in contrast to: an 18-layer variant of the ResNet Architecture (ResNet-18) trained end-to-end for each task (middle bars); and an unsupervised featurization that uses the last hidden layer of a 152-layer ResNet variant that was pre-trained using natural imagery in combination with ridge regression (lightest bars). Panel B shows the performance of ResNet-18 by task and training epoch, demonstrating that all tasks reached an asymptote after 400 epochs. Dark lines indicate the cumulative maximum performance by epoch, while light lines indicate epoch-specific performance.

III.1.1 Comparison to a deep convolutional neural network and an alternative unsupervised featurization First, we compare the performance of MOSAIKS to that of a tuned Residual Network (ResNet)[43] – a common, versatile deep network architecture used in recent satellite-based learning tasks [80]. We train this network *end to end* to predict outcomes in all seven tasks across the continental US, using as input the same imagery as that used by MOSAIKS.

We train an 18-layer variant of the ResNet Architecture using stochastic gradient descent to minimize the mean squared error (MSE) between the predictions and labels with an initial learning rate of $1e-4$ and momentum parameter of 0.9, training the model for 400 epochs, at which point performance reaches an asymptote (Fig. S12B). We employ a standard train/test split of 80%/20%, matching our approach when evaluating MOSAIKS.

Second, we compare MOSAIKS performance to a similarly cheap, unsupervised featurization generated by the last hidden layer of a pre-trained variant of the CNN used above, trained on natural imagery. To execute this comparison, we use the features from the last layer of a 152-layer variant of the ResNet Architecture, and then run ridge regression on these features for each task, as is done in MOSAIKS.

Fig. S12A demonstrates that MOSAIKS (dark bars) performs better or on par with ResNet (middle bars) across all seven tasks, while providing substantially greater performance than ridge regression run on features from the pre-trained CNN (lightest bars).

III.1.2 Interpretation of test accuracy comparisons Note that the performance of these models represents a reasonable lower bound on potential performance; some task-specific enhancements could be used to improve predictive power for each of these methods. For example, more layers could be added to ResNet or alternative architectures could be tested for specific tasks. In the case of MOSAIKS and the pre-trained ResNet features, more flexible regression models could have been used to estimate a variant of Eq. (4.2), such as increasing K, using a nonlinear model, or leveraging a hurdle model in tasks with a large number of zero observations. While these task-specific changes may marginally improve performance of any of these approaches, prior research on similar image recognition tasks suggests further gains for the ResNet are likely to be minimal [141]. While the similarity of performance in Fig. S12 is perhaps surprising, it is also encouraging for further research. This comparison suggests that wide, shallow networks using local-level features (analogous to random convolutional features) are as descriptive as more complex, highly optimized CNN architectures for satellite remote sensing, across many tasks.

We note that the similarity in performance of MOSAIKS and ResNet across tasks (Fig. S12) is consistent with a hypothesis that both approaches are reaching the limit of information that is provided by satellite imagery for predicting the outcomes we test. A human prediction baseline has not been established but could provide insight on whether there is substantial room for improvement in skill for each of these tasks, although we suspect many of these tasks will be difficult for nonexpert humans (e.g. nightlights or house prices).

III.2 Comparing costs

In practice, high computational costs can limit the use of SIML methods – especially when resources are scarce, such as in government agencies of low-income countries [39] or research teams and NGOs with limited budgets. Specifically designed to address this challenge, MOSAIKS scales across many research tasks by decoupling featurization from task selection, model-fitting, and prediction. The computationally costly step of featurization is done centrally on a fast computer with a graphics processing unit (GPU); individual practitioners need only download the pre-computed features, merge on labels for the task they select, and run regressions. Because features are created and stored by a central entity, the research community makes use of a cached set of computations, reducing the overall computational burden of widespread SIML and any external social costs generated by these computations [109]. Additionally, this decoupling of task-agnostic computations from task-specific computations allows practitioners to run more diagnostic analyses on their tasks, such as those presented in Fig. 4.3 of the main text.

From the perspective of a user who can access pre-computed MOSAIKS features to train and validate a new task, we find that MOSAIKS is 1080× faster than a state-of-

the-art neural net architecture (ResNet) (Table S5). Moreover, the ResNet does not achieve better predictive performance on the tasks we have studied (Fig. S12). From the perspective of the entire computational ecosystem, which bears the cost of image featurization in addition to model training and testing, we find that MOSAIKS is 32× faster than the ResNet. The times in Table S5 reflect our wall-clock time on a single Amazon EC2 instance and with seven domains specified in advance, so that the time costs are similar to that of introducing a single new domain *ex post*.[23] These ecosystem-wide costs of featurization per task continuously decline as MOSAIKS becomes more widely adopted, because features can be cached centrally and distributed to multiple users who are training and/or testing SIML in common locations.

We considered only one CNN architecture, which we chose because of its use in previous remote sensing applications [52]. We did not attempt to innovate in neural net architectural design or algorithms. While one could pursue targeted innovations in neural networks for remote sensing, such as in [[137]], we emphasize that our method is currently three orders of magnitude faster for the user than off-the-shelf fine-tuned CNN methods (Table S5), does not require a GPU for prediction,[24] and achieves the same or better prediction performance (Fig. S12). There is recent work that aims to train networks to learn a "common representation" that can generalize across tasks, but this is a subject of ongoing research [99], requires the tasks to be known in advance, and has yet to be demonstrated or evaluated at scale.

	ResNet	**MOSAIKS**
Component	Time	Time
Training set featurization ($N = 80$k)	**∼ 2.7 days**	∼ 1.2 hours
Model training		**∼ 3.5 minutes**
Holdout set featurization ($N = 20$k)	**∼ 40 seconds**	∼18 minutes
Holdout set prediction		**0.1 seconds**
Total cost to ecosystem	**∼ 2.7 days**	∼ 1.6 hours
Total cost to user	**∼ 2.7 days**	**∼ 3.6 minutes**

Extended Data Table S5: Wall-clock times of components of MOSAIKS compared with a fine-tuned CNN. Bold times are those that a practitioner using each method would incur. MOSAIKS times use the full $K = 8,192$ features. All operations were run for seven domains on an Amazon EC2 p3.xlarge instance with a Tesla V100 GPU and 16GB of onboard RAM. Cost of computation on this machine is roughly $\$3/hr$.

[23]The computational cost of training and testing models on 7 domains known *a priori* is similar to the cost of training and testing one new domain because if domains are known *a priori* one can solve the regressions and train the CNN jointly for all domains.

[24]Table S5 shows wall-clock times for ResNet and MOSAIKS on a GPU. However, future users of MOSAIKS are more likely to train and test their models on a standard laptop. We find that the total cost to the user of training and testing a new task on a standard laptop is approximately 6 minutes, less than 2× the value shown for a GPU.

Bibliography

[1] Berkeley Earth Daily Land Data. Accessed 01/01/2017.

[2] Cape Grim Greenhouse Gas data. Accessed 03/16/2018.

[3] Mauna Loa CO2 monthly mean data. Accessed 03/16/2018.

[4] A. Agarwal, S. M. Kakade, N. Karampatziakis, L. Song, and G. Valiant. Least squares revisited: Scalable approaches for multi-class prediction. *arXiv preprint arXiv:1310.1949*, 2013.

[5] M. Alber, P.-J. Kindermans, K. Schütt, K.-R. Müller, and F. Sha. An Empirical Study on The Properties of Random Bases for Kernel Methods. *Advances in Neural Information Processing Systems 30*, (1):2763–2774, 2017.

[6] R. Alkama and A. Cescatti. Biophysical climate impacts of recent changes in global forest cover. *Science*, 351(6273):600–604, feb 2016.

[7] P. B. Alton. Reduced carbon sequestration in terrestrial ecosystems under overcast skies compared to clear skies. *Agricultural and Forest Meteorology*, 148(10):1641–1653, 2008.

[8] Amazon Web Services. Terrain Tiles, 2018.

[9] V. H. Blackman. The Compound Interest Law and Plant Growth. *Annals of Botany*, XXXIII(CXXXI):353–360, 1919.

[10] E. Blanc and W. Schlenker. The Use of Panel Models in Assessments of Climate Impacts on Agriculture. *Review of Environmental Economics and Policy*, (October):258–279, 2017.

[11] O. Boucher, P. Randall, C. Artaxo, G. Bretherton, P. Feingold, V.-M. Forester, Y. Kerminen, H. Kondo, U. Liao, P. Lohmann, S. Rasch, S. Satheesh, B. Sherwood, and Stevens. 2013: Clouds and Aerosols. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 571–657, 2013.

[12] M. Burke and K. Emerick. Adaptation to Climate Change: Evidence from US Agriculture. *American Economic Journal: Economic Policy*, 8(3):106–140, 2016.

[13] J. Burney and V. Ramanathan. Recent climate and air pollution impacts on India agriculture. *Proceedings of the National Academy of Sciences*, 111(46):16319–16324, 2014.

[14] T. A. Carleton and S. M. Hsiang. Social and economic impacts of climate. *Science*, 353(6304), 2016.

[15] K. M. Carlson, R. Heilmayr, H. K. Gibbs, P. Noojipady, D. N. Burns, D. C. Morton, N. F. Walker, G. D. Paoli, and C. Kremen. Effect of oil palm sustainability certification on deforestation and fire in Indonesia. *Proceedings of the National Academy of Sciences of the United States of America*, 115(1):121–126, jan 2018.

[16] Center for International Earth Science Information Network (CIESIN). Gridded Population of the World, Version 4, 2016.

[17] G. Cheng, J. Han, and X. Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, oct 2017.

[18] A. M. Cheriyadat. Unsupervised feature learning for aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 52(1):439–451, 2014.

[19] E. Chuvieco. *Earth observation of global change: The role of satellite remote sensing in monitoring the global environment*. 2008.

[20] A. Coates, A. Arbor, and A. Y. Ng. An Analysis of Single-Layer Networks in Unsupervised Feature Learning. *International Conference on Articial Intelligence and Statistics*, pages 215–223, 2011.

[21] A. Coates and A. Y. Ng. 2014 10 16 - Learning Feature Representations with K-means. 2012.

[22] A. Colin Cameron, J. B. Gelbach, and D. L. Miller. Robust inference with multiway clustering. *Journal of Business and Economic Statistics*, 29(2):238–249, 2011.

[23] P. J. Crutzen. Albedo enhancement by stratospheric sulfur injections: A contribution to resolve a policy dilemma? *Climatic Change*, 77(3-4):211–219, 2006.

[24] M. Déqué. Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: Model results and statistical correction according to observed values. *Global and Planetary Change*, 57(1-2):16–26, 2007.

[25] E. G. Dutton and J. R. Christy. Solar radiative forcing at selected locations and evidence for global lower tropospheric cooling following the eruptions of El Chichón and Pinatubo. *Geophysical Research Letters*, 19(23):2313–2316, 1992.

[26] O. Firat, G. Can, and F. T. Y. Vural. Representation learning for contextual object and region detection in remote sensing. In *2014 22nd International Conference on Pattern Recognition*, pages 3708–3713. IEEE, 2014.

[27] Food and Agriculture Organization of the United Nations. Accessed 01/01/2016. FAOSTAT.

[28] J. Ge, J. Su, Q. Fu, T. Ackerman, and J. Huang. Dust aerosol forward scattering effects on ground-based aerosol optical depth retrievals. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 112(2):310–319, 2011.

[29] M. Gechter, N. Tsivanidis, R. Cooper, J. Eaton, P. Fajgelbaum, J. Hersh, C.-T. Hsieh, D. Keniston, K. Krishna, A. Pethe, V. Phatak, D. Ray, S. Redding, H. Chen, I. Choi, J. Costantini, D. He, S. Jacob, P. Lin, A. Lopez, R. Pandit, P. Sanghvi, C. Suhr, and S. Thube. The Welfare Consequences of Formalizing Developing Country Cities: Evidence from the Mumbai Mills Redevelopment *. 2018.

[30] R. Gelaro, W. McCarty, M. J. Suárez, R. Todling, A. Molod, L. Takacs, C. A. Randles, A. Darmenov, M. G. Bosilovich, R. Reichle, K. Wargan, L. Coy, R. Cullather, C. Draper, S. Akella, V. Buchard, A. Conaty, A. M. da Silva, W. Gu, G. K. Kim, R. Koster, R. Lucchesi, D. Merkova, J. E. Nielsen, G. Partyka, S. Pawson, W. Putman, M. Rienecker, S. D. Schubert, M. Sienkiewicz, and B. Zhao. The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). *Journal of Climate*, 30(14):5419–5454, 2017.

[31] M. Gindelsky, J. Moulton, and S. Wentland. Valuing Housing Services in the Era of Big Data: A User Cost Approach Leveraging Zillow Microdata. *Forthcoming in NBER-CRIW*, Volume on, 2019.

[32] E. H. Glenn. acs: Download, Manipulate, and Present American Community Survey and Decennial Data from the US Census, 2019.

[33] Google Developers. Maps Static API.

[34] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27, 2017.

[35] R. Greenwald, M. H. Bergin, J. Xu, D. Cohan, G. Hoogenboom, and W. L. Chameides. The influence of aerosols on crop production: A study using the CERES crop model. *Agricultural Systems*, 89(2-3):390–413, 2006.

[36] L. Gu, D. D. Baldocchi, S. C. Wofsy, J. W. Munger, J. J. Michalsky, S. P. Urbanski, and T. a. Boden. Response of a deciduous forest to the Mount Pinatubo eruption: enhanced photosynthesis. *Science*, 299(5615):2035–2038, 2003.

[37] R. Gupta, E. Somanathan, and S. Dey. Global warming and local air pollution have reduced wheat yields in India. *Climatic Change*, 2016.

[38] R. Gupta, E. Somanathan, and S. Dey. Global warming and local air pollution have reduced wheat yields in India. *Climatic Change*, 140(3-4):593–604, 2017.

[39] B. Haack and R. Ryerson. Improving remote sensing research and education in developing countries: Approaches and recommendations. *International Journal of Applied Earth Observation and Geoinformation*, 45:77–83, mar 2016.

[40] M. C. Hansen, P. V. Potapov, R. Moore, M. Hancher, S. a. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, a. Kommareddy, A. Egorov, L. Chini, C. O. Justice, J. R. G. Townshend, P. Patapov, R. Moore, M. Hancher, S. a. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommaredy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend. High-Resolution Global Maps of. *Science*, 342(November):850–854, 2013.

[41] F. Harrell. *Regression Modeling Strategies*. Springer-Verlag, New York, 1st edition, 2001.

[42] T. Hayasaka, N. Iwasaka, G. Hashida, I. Takizawa, and M. Tanaka. Changes in stratospheric aerosols and solar insolation due to Mt. Pinatubo eruption as observed over the western Pacific. *Geophysical Research Letters*, 21(12):1137–1140, 1994.

[43] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[44] M. Hess, P. Koepke, and I. Schult. Optical properties of aerosols and clouds. *Bulletin of the American Meteorological Society*, 79(5):831–844, 1998.

[45] S. Hsiang and R. E. Kopp. An Economist's Guide to Climate Change Science. *Journal of Economic Perspectives*, 32(4):3–32, 2018.

[46] S. M. Hsiang. Climate Econometrics. *Annual Review of Resource Economics*, (July):1–33, 2016.

[47] S. M. Hsiang, D. Lobell, and M. Roberts. Climate Change and Crop Choice : Evidence from Australia , Brazil , China , Europe and the United States. 2015.

[48] S. M. Hsiang and K. C. Meng. Tropical economics. *American Economic Review*, 105(5):257–261, 2015.

[49] W. Hu, J. H. Patel, Z.-A. Robert, P. Novosad, S. Asher, Z. Tang, M. Burke, D. Lobell, and S. Ermon. Mapping Missing Population in Rural India: A Deep Learning Approach with Satellite Imagery. In *Conference on Artificial Intelligence, Ethics, and Society*, Honolulu, HI, 2019.

[50] J. Inglada, A. Vincent, M. Arias, B. Tardy, D. Morin, and I. Rodes. Operational high resolution land cover map production at the country scale using satellite image time series. *Remote Sensing*, 9(1):95, 2017.

[51] ISCCP. International Satellite Cloud Climatology Project (ISCCP) D1. Accessed 07/02/2016.

[52] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

[53] N. Jean, S. Wang, A. Samar, G. Azzari, D. Lobell, and S. Ermon. Tile2Vec: Unsupervised representation learning for spatially distributed data. 2018.

[54] Z. Jin, G. Azzari, C. You, S. Di Tommaso, S. Aston, M. Burke, and D. B. Lobell. Smallholder maize area and yield mapping at national scales with Google Earth Engine. *Remote Sensing of Environment*, 228(March):115–128, 2019.

[55] E. Jonas, M. Bobra, V. Shankar, J. Todd Hoeksema, and B. Recht. Flare Prediction Using Photospheric and Coronal Image Data. *Solar Physics*, 293(3):1–22, 2018.

[56] K. D. Kanniah, J. Beringer, P. North, and L. Hutley. Control of atmospheric particles on diffuse radiation and terrestrial plant productivity: A review. *Progress in Physical Geography*, 36(2):209–237, 2012.

[57] B. Kravitz, A. Robock, O. Boucher, H. Schmidt, K. E. Taylor, G. Stenchikov, and M. Schulz. The Geoengineering Model Intercomparison Project (GeoMIP). *Atmospheric Science Letters*, 12(2):162–167, 2011.

[58] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[59] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[60] T. Li, E. Heuvelink, T. A. Dueck, J. Janse, G. Gort, and L. F. Marcelis. Enhancement of crop photosynthesis by diffuse light: Quantifying the contributing factors. *Annals of Botany*, 114(1):145–156, 2014.

[61] D. B. Lobell and C. B. Field. Estimation of the carbon dioxide (CO2) fertilization effect using growth rate anomalies of CO2 and crop yields since 1961. *Global Change Biology*, 14(1):39–45, 2008.

[62] D. G. Macmartin, B. Kravitz, J. C. S. Long, and P. J. Rasch. Geoengineeringwith stratospheric aerosols: What do we not know after a decade of research? *Earth's Future*, 4(11):1–6, 2016.

[63] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez. Convolutional Neural Networks for Large-Scale Remote-Sensing Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2):645–657, feb 2017.

[64] A. E. Maxwell, T. A. Warner, and F. Fang. Implementation of machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing*, 39(9):2784–2817, 2018.

[65] J. M. McGrath, A. M. Betzelberger, S. Wang, E. Shook, X.-G. Zhu, S. P. Long, and E. A. Ainsworth. An analysis of ozone damage to historical maize and soybean yields in the United States. *Proceedings of the National Academy of Sciences of the United States of America*, 112(46):14390–5, 2015.

[66] L. M. Mercado, N. Bellouin, S. Sitch, O. Boucher, C. Huntingford, M. Wild, and P. M. Cox. Impact of changes in diffuse radiation on the global land carbon sink. *Nature*, 458(7241):1014–1017, 2009.

[67] C. Monfreda, N. Ramankutty, and J. A. Foley. Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. *Global Biogeochemical Cycles*, 22(1):1–19, 2008.

[68] A. Morrow, V. Shankar, D. Petersohn, A. Joseph, B. Recht, and N. Yosef. Convolutional Kitchen Sinks for Transcription Factor Binding Site Prediction. (Iid), 2017.

[69] J. Moulton and S. Wentland. Monetary Policy and the Housing Market. In *Annual Meeting of the American Economic Association*, Philadelphia, PA, 2018.

[70] U. Niemeier, H. Schmidt, K. Alterskjær, and J. E. Kristjánsson. Solar irradiance reduction via climate engineering: Impact of different techniques on the energy balance and the hydrological cycle. *Journal of Geophysical Research Atmospheres*, 118(21):11905–11917, 2013.

[71] NOAA National Centers for Environmental Information. Version 1 VIIRS Day/Night Band Nighttime Lights, 2019.

[72] J. R. Norris, R. J. Allen, A. T. Evan, M. D. Zelinka, C. W. O'Dell, and S. A. Klein. Evidence for climate change in the satellite cloud record. *Nature*, 536(7614):72–75, 2016.

[73] J. R. Norris and A. T. Evan. Empirical removal of artifacts from the ISCCP and PATMOS-x satellite cloud records. *Journal of Atmospheric and Oceanic Technology*, 32(4):691–702, 2015.

[74] Ocean Studies Board. Climate Intervention: Reflecting Sunlight to Cool the Earth. *National Academies Press*, 2015.

[75] A. J. Oliphant, D. Dragoni, B. Deng, C. S. B. Grimmond, H. P. Schmid, and S. L. Scott. The role of sky conditions on gross primary production in a mixed deciduous forest. *Agricultural and Forest Meteorology*, 151(7):781–791, 2011.

[76] R. K. Pachauri, M. R. Allen, V. R. Barros, J. Broome, W. Cramer, R. Christ, J. A. Church, L. Clarke, Q. Dahe, P. Dasgupta, and Others. *Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change*. IPCC, 2014.

[77] G. Papaioannou, N. Papanikolaou, and D. Retalis. Theoretical and Applied Climatology Relationships of Photosynthetically Active Radiation and Shortwave Irradiance. *Theoretical and Applied Climatology*, 27:23–27, 1993.

[78] J. F. Pekel, A. Cottam, N. Gorelick, and A. S. Belward. High-resolution mapping of global surface water and its long-term changes. *Nature*, 540(7633):418–422, 2016.

[79] O. A. Penatti, K. Nogueira, and J. A. Dos Santos. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 44–51, 2015.

[80] A. Perez, C. Yeh, G. Azzari, M. Burke, D. Lobell, and S. Ermon. Poverty Prediction with Public Landsat 7 Satellite Imagery and Machine Learning. (Nips), 2017.

[81] A. Pérez-Suay, J. Amorós-López, L. Gómez-Chova, V. Laparra, J. Muñoz-Marí, and G. Camps-Valls. Randomized kernels for large scale Earth observation applications. *Remote Sensing of Environment*, 202:54–63, dec 2017.

[82] J. Pongratz, D. B. Lobell, L. Cao, and K. Caldeira. Crop yields in a geoengineered climate. *Nature Climate Change*, 2(2):101–105, 2012.

[83] J. R. Porter, L. Xie, A. J. Challinor, K. Cochrane, S. M. Howden, M. M. Iqbal, D. B. Lobell, and M. I. Travasso. Food security and food production systems. *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 485–533, 2014.

[84] J. Proctor, S. Hsiang, J. Burney, M. Burke, and W. Schlenker. Estimating global agricultural effects of geoengineering using volcanic eruptions. *Nature*, 560(7719):480–483, 2018.

[85] A. Rahimi and B. Recht. Random Features for Large-Scale Kernel Machines. In *Advances in Neural Information Processing Systems*, 2007.

[86] A. Rahimi and B. Recht. Uniform approximation of functions with random bases. In *46th Annual Allerton Conference on Communication, Control, and Computing*, pages 555–561. IEEE, sep 2008.

[87] A. Rahimi and B. Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Nips*, 1(1):1313–1320, 2008.

[88] N. Ramankutty, A. T. Evan, C. Monfreda, and J. A. Foley. Farming the planet: 1. Geographic distribution of global agricultural lands in the year 2000. *Global Biogeochemical Cycles*, 22(1):1–19, 2008.

[89] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do imagenet classifiers generalize to imagenet? *arXiv preprint arXiv:1902.10811*, 2019.

[90] M. J. Roberts, W. Schlenker, and J. Eyer. Agronomic weather measures in econometric models of crop yield with implications for climate change. *American Journal of Agricultural Economics*, 95(2):236–243, 2013.

[91] C. Robinson and F. Hohman. A Deep Learning Approach for Population Estimation from Satellite Imagery. *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*, 2017.

[92] A. Robock. Volcanic Eruptions and Climate. *Reviews of Geophysics*, 38(2):191–219, 2000.

[93] A. Robock, D. G. MacMartin, R. Duren, and M. W. Christensen. Studying geoengineering with natural and anthropogenic analogs. *Climatic Change*, 121(3):445–458, 2013.

[94] A. Robock, A. Marquardt, B. Kravitz, and G. Stenchikov. Benefits, risks, and costs of stratospheric geoengineering. *Geophysical Research Letters*, 36(19):1–9, 2009.

[95] M. L. Roderick and G. D. Farquhar. Geoengineering: Hazy, cool and well fed? *Nature Climate Change*, 2(2):76–77, 2012.

[96] M. L. Roderick, G. D. Farquhar, S. L. Berry, and I. R. Noble. On the direct effect of clouds and atmospheric particles on the productivity and structure of vegetation. *Oecologia*, 129(1):21–30, 2001.

[97] A. Romero, C. Gatta, and G. Camps-Valls. Unsupervised deep feature extraction for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 54(3):1349–1362, 2016.

[98] W. B. Rossow, A. Walker, V. Golea, A. Inamdar, and B. Hankins. International Satellite Cloud Climatology Project Climate Data Record, H-Series [HGG], 2016.

[99] S. Ruder. An Overview of Multi-Task Learning in Deep Neural Networks. *arXiv*, jun 2017.

[100] W. J. Sacks, D. Deryng, J. A. Foley, and N. Ramankutty. Crop planting dates: An analysis of global patterns. *Global Ecology and Biogeography*, 19(5):607–620, 2010.

[101] M. Sato, J. E. Hansen, M. P. McCormick, and J. B. Pollack. Straotpheric Aerosol Optical Depths 1850 - 1990. *Journal of Geophysical Research*, 98(D12):389–416, 1993.

[102] L. D. Schiferl and C. L. Heald. Particulate matter air pollution may offset ozone damage to global crop production. *Atmospheric Chemistry and Physics*, 18(8):5953–5966, 2018.

[103] W. Schlenker and D. B. Lobell. Robust negative impacts of climate change on African agriculture. *Environmental Research Letters*, 5(1):14010, 2010.

[104] W. Schlenker and M. J. Roberts. Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. *Proceedings of the National Academy of Sciences*, 106(37):15594–15598, 2009.

[105] J. Sheffield, G. Goteti, and E. F. Wood. {D}evelopment of a 50-{Y}ear {H}igh-{R}esolution {G}lobal {D}ataset of {M}eteorological {F}orcings for {L}and {S}urface {M}odeling. *J. Climate*, 19(13):3088–3111, 2006.

[106] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, San Diego, CA, sep 2015.

[107] A. R. Stine and P. Huybers. Arctic tree rings as recorders of variations in light availability. *Nature Communications*, 5(May):1–8, 2014.

[108] S. Strada, N. Unger, and X. Yue. Observed aerosol-induced radiative effect on plant productivity in the eastern United States. *Atmospheric Environment*, 122:463–476, 2015.

[109] E. Strubell, A. Ganesh, and A. McCallum. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*, 2019.

[110] Q. Tang, P. G. Hess, B. Brown-Steiner, and D. E. Kinnison. Tropospheric ozone decrease due to the Mount Pinatubo eruption: Reduced stratospheric influx. *Geophysical Research Letters*, 40(20):5553–5558, 2013.

[111] K. E. Taylor, R. J. Stouffer, and G. A. Meehl. An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, 93(4):485–498, 2012.

[112] K. E. Taylor, R. J. Stouffer, and G. A. Meehl. An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, 93(4):485–498, 2012.

[113] L. Thomason and T. Peter. Stratospheric Processes and their Role in Climate Assessment of Stratospheric Aerosol Properties. Technical Report No. 4, SPARC, 2006.

[114] M. Tollenaar, J. Fridgen, P. Tyagi, P. W. S. Jr, and S. Kumudini. The contribution of solar brightening to the US maize yield trend. *Nature Climate Change*, 7(March), 2017.

[115] K. E. Trenberth and A. Dai. Effects of Mount Pinatubo volcanic eruption on the hydrological cycle as an analog of geoengineering. *Geophysical Research Letters*, 34(15):1–5, 2007.

[116] Union of Concerned Scientists. Underwater: Rising Seas, Chronic Floods, and the Implications for US Coastal Real Estate. Technical report, 2018.

[117] Union of Concerned Scientists. UCS Satellite Database. jan 2019.

[118] D. W. Urban, M. J. Roberts, W. Schlenker, and D. B. Lobell. The effects of extremely wet planting conditions on maize and soybean yields. *Climatic Change*, 130(2):247–260, 2015.

[119] U.S. Census Bureau. 2015 American Community Survey 5-Year Estimates, Table B19013.

[120] U.S. Census Bureau. TIGER/Line Geodatabases, 2016.

[121] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(Dec):3371–3408, 2010.

[122] A. Vogelmann, T. Ackerman, and R. Turco. Enhancements in biologically effective ultraviolet radiation following volcanic eruptions. *Nature*, 359:47–49, 1992.

[123] M. Volpi and D. Tuia. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2):881–893, 2017.

[124] M. Webb, C. Senior, S. Bony, and J. J. Morcrette. Combining ERBE and ISCCP data to assess clouds in the Hadley Centre, ECMWF and LMD atmospheric climate models. *Climate Dynamics*, 17(12):905–922, 2001.

[125] M. Wild, H. Gilgen, A. Roesch, A. Ohmura, C. N. Long, E. G. Dutton, B. Forgan, A. Kallis, V. Russak, and A. Tsvetkov. From Dimming to Brightening: Decadal Changes in Solar Radiation at Earth's Surface. *Science*, 308(5723):847–850, 2005.

[126] M. Wild, A. Roesch, and C. Ammann. Global dimming and brightening - evidence and agricultural implications. *CAB Reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources*, 7(003):1–7, 2012.

[127] C. Willmott and K. Matsurra. Terrestrial Air Temperature and Precipitation: Monthly and Annual Time Series (1950 - 1999). Accessed 01/01/2016. 2001.

[128] J. M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA, 2002.

[129] World Meteorological Organization. World Radiation Data Centre. Accessed 01/08/2015.

[130] L. Xia, J. P. Nowack, S. Tilmes, and A. Robock. Impacts of stratospheric sulfate geoengineering on tropospheric ozone. *Atmospheric Chemistry and Physics*, 17(19):11913–11928, 2017.

[131] L. Xia, A. Robock, J. Cole, C. L. Curry, D. Ji, A. Jones, B. Kravitz, J. C. Moore, H. Muri, U. Niemeier, B. Singh, and S. Tilmes. Solar radiation management impacts on agriculture in China: A case study in the Geoengineering Model Intercomparison Project (GeoMIP). *Journal of Geophysical Research: Atmospheres*, 119:8695–8711, 2014.

[132] L. Yu, L. Liang, J. Wang, Y. Zhao, Q. Cheng, L. Hu, S. Liu, L. Yu, X. Wang, P. Zhu, X. Li, Y. Xu, C. Li, W. Fu, X. Li, W. Li, C. Liu, N. Cong, H. Zhang, F. Sun, X. Bi, Q. Xin, D. Li, D. Yan, Z. Zhu, M. F. Goodchild, and P. Gong. Meta-discoveries from a synthesis of satellite-based land-cover mapping research. *International Journal of Remote Sensing*, 35(13):4573–4588, 2014.

[133] M. D. Zelinka, T. Andrews, P. M. Forster, and K. E. Taylor. Quantifying components of aerosol-cloud-radiation interactions in climate models. *Journal of Geophysical Research*, 119(12):7599–7615, 2014.

[134] M. D. Zelinka, S. A. Klein, K. E. Taylor, T. Andrews, M. J. Webb, J. M. Gregory, and P. M. Forster. Contributions of different cloud types to feedbacks and rapid adjustments in CMIP5. *Journal of Climate*, 26(14):5007–5027, 2013.

[135] B. C. Zhang, J. J. Cao, Y. F. Bai, S. J. Yang, L. Hu, and Z. G. Ning. Effects of cloudiness on carbon dioxide exchange over an irrigated maize cropland in northwestern China. *Biogeosciences Discussions*, 8(1):1669–1691, 2011.

[136] P. Zhang, J. Zhang, and M. Chen. Economic impacts of climate change on agriculture: The importance of additional climatic variables other than temperature and precipitation. *Journal of Environmental Economics and Management*, 83:8–31, 2017.

[137] Y. Zhong, F. Fei, Y. Liu, B. Zhao, H. Jiao, and L. Zhang. SatCNN: satellite image dataset classification using agile convolutional neural networks. *Remote Sensing Letters*, 8(2):136–145, 2017.

[138] Zillow. ZTRAX: Zillow Transaction and Assessor Dataset, 2017.

[139] Zillow Research. zillow-research/ztrax.

[140] A. Zolli. After big data: The coming age of "big indicators". *Stanford Social Innovation Review*, January 2018.

[141] B. Zoph and Q. V. Le. Neural Architecture Search with Reinforcement Learning. 2016.