DETECTING THE EFFECT OF DUST AND OTHER CLIMATE VARIABLES ON CROP YIELDS USING DIAGNOSTIC STATISTICAL CROP MODELS

A Dissertation in
Meteorology
by
Alexis L Hoffman

© 2018 Alexis L Hoffman

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2018
The dissertation of Alexis L Hoffman was reviewed and approved* by the following:

Chris E Forest  
Associate Professor of Climate Dynamics  
Dissertation Advisor  
Chair of Committee

Armen R Kemanian  
Associate Professor of Plant Science

Marcelo Chamecki  
Associate Professor of Atmospheric Science

Gregory Jenkins  
Professor of Meteorology

Natalie Mahowald  
Professor of Earth and Atmospheric Science  
Special Member

David J Stensrud  
Professor of Meteorology  
Department Head

*Signatures are on file in the Graduate School.
Food security and agriculture productivity assessments require a strong understanding of how climate and other drivers influence regional crop yields. While the effects of temperature, precipitation, and carbon dioxide are relatively well-understood, the effect of dust on crop yields has yet to be thoroughly investigated. This line of inquiry is warranted because many areas of the world with frequent dust storms and high dust loadings are often food insecure, and because wind erosion is prevalent in the High Plains of the United States, a major crop-producing area. Existing research suggests that the effect of dust on yields should be largely negative, but until now this has not been investigated on a regional scale. A major hindrance to understanding the effect of dust on crop yields is insufficient data and inadequate methods of analysis. In this dissertation, we developed data and analysis methods for three distinct projects to determine whether dust affects regional crop yields.

In the first project, we validated the use of random forest, a machine learning technique, as a diagnostic crop model that can be used to assess the impact of individual climate predictors on yields. Because we motivated this research with food security concerns, we analyzed climate signals in the crop yield record of sub-Saharan Africa from 1962-2014. From this work, we determined that random forest could function as a statistical crop model, but the data quality and resolution inhibited the ability to detect the effect of dust on yields in this area of the world.

In our second line of inquiry, we shifted the focus to the central region of the United States for its high quality and high resolution data, as well as its importance as a major crop-producing region of the world. Because these data had higher temporal resolution, we could explore individual phases of the growing season. We developed crop-specific algorithms to compute the planting date, establishment phase, critical window, and grain filling phase to investigate yield responses to phase-specific climate predictors. Using these data, the random forest identified distinct phase-specific responses for important climate predictors. Finally, we computed dust metrics from three different data sources and merged them with climate and yield data in the central region of the United States to estimate the
impact of dust on yields. Over the entire central US region, we found that including dust as a predictor in each crop model did not improve yield predictions for the region as a whole. However, when crop models were applied to individual states, we found several instances in which dust weakly reduced yields. Although these state-specific results were encouraging, we presented them cautiously because the yield responses could be an artifact of either partitioning the data or a true yield response that is obscured when data was spatially aggregated. While the results were largely inconclusive, we have advanced the capabilities of statistical crop modeling, developed data sets that can be used to move the science forward, and revealed new questions that merit further research.
# Table of Contents

List of Figures ix
List of Tables xvii
Acknowledgments xx

## Chapter 1
### Introduction 1
  1.1 Purpose ......................................................... 1
  1.2 Motivation ..................................................... 2
    1.2.1 Existing research on climate and crop yields ............. 3
    1.2.1.1 Crop Model: Process-based crop models ............... 3
    1.2.1.2 Crop Model: Statistical crop models ................. 4
  1.3 Scientific justification of research question ................... 5
    1.3.1 Dust ......................................................... 5
    1.3.1.1 General Characteristics ............................... 6
    1.3.2 Aerosol and erosion effect on plants ..................... 7
    1.3.2.1 Aerosols & plants ...................................... 7
    1.3.2.2 Soil erosion & plants ................................ 10
  1.4 Evolution of research agenda .................................. 11

## Chapter 2
### Analysis of climate signals in the crop yield record of Sub-Saharan Africa 15
  2.1 Abstract .................................................... 15
  2.2 Introduction ................................................ 16
  2.3 Materials & Methods ......................................... 20
    2.3.1 Data .................................................... 20
    2.3.1.1 Crop yield ........................................... 20
Chapter 3

Effects of historical weather on US yields since 1980

3.1 Introduction ................................................. 77
3.2 Data ......................................................... 80
Chapter 4
Detecting the effect of dust events on crop yields in the central US

4.1 Introduction ....................................................... 111
4.2 Data & Methods .................................................. 113
  4.2.1 Yield and irrigation data .................................. 114
  4.2.2 Climate data ................................................ 114
  4.2.3 Dust event from MODIS data ............................. 116
  4.2.4 Station data ................................................ 118
    4.2.4.1 Global Historical Climatology Network - Daily ... 118
Chapter 4

4.2.4.2 Integrated Surface Data Hourly Global 120
4.2.4.3 Dust event identification: station data 121
4.2.5 Statistical crop model 122
4.2.6 Growing season phases 122
  4.2.6.1 Methods to compute growth phases 122
  4.2.6.2 Data used to compute growth phases 125
4.2.7 Experimental Design 125
4.3 Results 128
  4.3.0.1 Sensitivity to RF parameters 129
  4.3.0.2 Phase-specific analysis 131
  4.3.0.3 Residual analysis 132
  4.3.0.4 Spatial analysis 132
4.3.1 AOD effect on yields 135
4.4 Discussion 139
  4.4.1 Potential yield impact of dust events 139
  4.4.2 Inconclusive results 140
  4.4.3 Potential sources of error in methodology or data 142
4.5 Conclusion: 145
4.6 Supplement 147
  4.6.1 Characteristics of dust storms in this region 147
  4.6.2 MODIS Methodology 148
  4.6.3 Station Methodology 148
  4.6.4 Experimental set up 148

Chapter 5

Summary and Conclusion 167
5.1 Future work 169

Bibliography 172
List of Figures

1.1 Comparison of the (a) least developed countries (Food and Agriculture Organization of the United Nations, 2014) and (b) total AOD from monthly MODIS-Aqua data from 2002-2014. Red countries in (a) denote the least developed countries, and warmer colors represent deeper optical depths in (b). Note that AOD, not dust optical depth is plotted in this figure, so biomass burning, air pollution, and mineral aerosols contribute to AOD. ...................................................... 2

1.2 Flow chart describing knowledge gap. .......................................................... 3

2.1 The 35 countries from the four regions of interest used in this study. Countries with greater than 10% undernourishment rates are denoted by solid colors, all other countries are hatched (FAO, IFAD, and WFP, 2015). .............................................................. 19

2.2 (a) Average annual yield for maize, groundnut, and sorghum from 1962 to 2014 over all countries. (b) Partial dependence plot for time. Maize, groundnut, and sorghum are denoted by green, orange, and black lines, respectively. Partial dependence plots graph the independent variable against the model outcome, after accounting for the average effect of other independent variables in the model . 26

2.3 Modeled yield is plotted against observed yield for (a,c,e) the simple linear model (identical to the model included in Row 3 of Table 2.3) and (b,d,f) the Random Forest model (identical to the model included in Row 4 of Table 2.3) for (a,b) maize, (c,d) groundnut, and (e,f) sorghum. The data used to train the model are designated by black diamonds, while the test data is designated by red circles. The R² and mean squared errors for the training and test sets are included in the top left corner of panel in blue, black, and red, respectively. .............................................................. 32
2.4 Partial dependence plots for (a) maximum temperature (°C), (b) precipitation (mm), and (c) vapor pressure deficit (hPa). The y-axis represents the average deviation in yields caused by a given variable (kg/ha). We plot data between the 2.5-97.5th percentiles of the climate data to focus on the more robust signals, though several small, exaggerated features remain.

2.5 Feature contribution (FC) plot for groundnut. Panel titles designate which variable is being plot along the x-axis: (a) maximum temperature, (b) accumulated precipitation, (c) vapor pressure deficit, (d) time, and (e) country index. Panel titles also include the R² (leave-one-out goodness of fit) of the average FC line (denoted in orange). The color gradient is applied in all panels along the maximum temperature axis changing from green-blue-red with increasing temperature. The country indices are: Mali (1), Niger (2), Chad (3), Sudan (4), Burkina Faso (5), Senegal (6), The Gambia (7), Guinea Bissau (8), Guinea (9), Sierra Leone (10), Cote D’Ivoire (11), Ghana (12), Togo (13), Benin (14), Nigeria (15), Somalia (16), Kenya (17), Tanzania (18), Uganda (19), Ethiopia (20), Rwanda (21), Burundi (22), Zambia (23), Malawi (24), Mozambique (25), Zimbabwe (26), Botswana (27), Namibia (28), South Africa (29), Lesotho (30), Swaziland (31). Groundnut yield data for Lesotho is missing, therefore the country index for Swaziland is 30 for this crop only (Table S2.1).

2.6 Dual (or 3D) partial dependence plot for (a) maize and (b) sorghum yield, with maximum temperature (tmx) and precipitation (pre) as the predictor variables. Maximum temperature (tmx) and precipitation (pre) are plotted in the horizontal, xy-plane, and the average modeled yield on the vertical, z-axis. Color gradient aligns with average modeled yield.

S2.1 Crop area map for (a) maize, (b) groundnut, and (c) sorghum based on data from (Monfreda et al., 2008). The countries included in the analysis are outlined in black. The scale is logarithmic and ranges from white to red as crop area fraction increases.

S2.2 Average harvest day for (a) maize, (b) groundnut, and (c) sorghum based on data from (Sacks et al., 2010). The countries included in the analysis are outlined in black. The scale is logarithmic and ranges from white to red as crop area fraction increases.
S2.3 Sample distributions for data used in regression analyses: (a) annual yields (b)-(k) climate variables during the growing season. The bin sizes for each variable are 25 kg/ha, 0.1°C, 0.1°C, 0.1°C, 10 mm, 0.1 hPa, 2.5 mm/month, 0.5%, 0.025 (unitless), 0.075 hPa, 0.025 (unitless), for yield, \( \text{tmp}, \text{tmn}, \text{tmx}, \text{pre}, \text{vap}, \text{ETo}, \text{cld}, \text{SPI}, \text{vpd}, \text{and aridity} \), respectively.

S2.4 Same as Figure S2.3 but for groundnut.

S2.5 Same as Figure S2.3 but for sorghum.

S2.6 Heat map visualization of the correlation matrix computed for all candidate predictors listed in Table 2.1.

S2.7 Solid black line plots OOB error rate as a function of the number of trees for a model with all predictors. Dashed red line represents the same function for a model with 6 variables: \( \text{time}, \text{country}, \text{tmx}, \text{pre}, \text{vpd}, \text{and cld} \). Dashed blue line represents the same function using the 5 variables in the final model: \( \text{time}, \text{country}, \text{tmx}, \text{pre}, \text{and vpd} \). Dashed green line represents the same function using only four variables: \( \text{time}, \text{country}, \text{tmx}, \text{and pre} \). Vertical grey line delineates \( n_{\text{tree}} = 300 \).

S2.8 Prediction performance of Random Forest model with systematic reduction of predictors for maize, groundnut, and sorghum, respectively, using k-fold cross validation (k=10). This Random Forest model uses \( n_{\text{tree}}=300 \) and \( m_{\text{try}}=3 \).

S2.9 Dependence of groundnut on country index for data used in the regression. Yields which occurred when maximum temperatures were greater than 30°C are denoted by red dots. The horizontal blue segments represent the mean groundnut yields for SAF/WAF and EAF/SAF.

S2.10 Feature contribution plot for maize with the color gradient applied along the maximum temperature axis. A consistent y-axis is maintained across each panel, which limits the ability to distinguish features in the climate variable plots.

S2.11 Feature contribution plot for sorghum with the color gradient applied along the country index axis. The color scale identifies each country.
S2.12 Raw partial dependence plot for maximum temperature (not deviations) for maize, groundnut, and sorghum, respectively. Similar to Figure 2.3, data within the innermost 95th percentile are marked by a green "x". The red line represents the linear trend line through all plotted values; the fit represents the R$^2$ for each trend line. The red number in the bottom left of each panel represents the percent change relative to the average yield caused by one degree of maximum temperature increase.

S2.13 Composite standardized partial dependence plots for maize, groundnut, and sorghum. A composite model is based on combining the data for all three crops, so that panel data are specified on the crop-country-year level. The 5th and 95th percentile bounds are designated by vertical grey lines. If data is missing for a given interval (i.e. if data exists for maize, but not groundnut and sorghum at 23°C), the average is computed based on quantities which are present. The thick, red line represents a trend line based on the data between the 5th and 95th percentiles to highlight the most robust trends in the yield responses to climate.

S2.14 Residuals from linear models (a,c,e) and residuals from Random Forest models (b,d,f). Residuals are plotted as a function of the modeled yields and a color gradient is applied along the observed yield axis (not plotted). This gradient indicates increasing yields as colors shift from blue-green-red.

S2.15 Total fertilizer consumption from all countries in the analysis from 1961-2014. Units of consumption are 10,000 metric tonnes of nutrient. Black and red lines represent nitrogen and phosphorous fertilizer consumption. The dashed lines represent the total when South Africa is omitted.

S2.16 Average nitrogen fertilizer consumption of each country between 1961-2014.

S2.17 Average yield for maize, groundnut, and sorghum between 1962 and 2014 for all countries in the final regression model. (b) Partial dependence plot for time for the model used in the manuscript. (c) Partial dependence plot for time with nitrogen fertilizer included as a predictor. Yield response to time should be interpreted as response to technology. As in the manuscript, maize, groundnut, and sorghum are denoted by green, orange, and black lines, respectively.
S2.18 Partial dependence plots for (a) maximum temperature (°C), (b) precipitation (mm), and (c) vapor pressure deficit (hPa) for the model which includes consumption of nitrogen fertilizer. The average deviation in yields caused by a given variable (kg/ha) are plotted against each variable. Data between the innermost 95th percentiles are plotted to focus on the robust signals.

S2.19 Partial dependence plots for (a) maximum temperature (°C), (b) standardized precipitation index (unitless), and (c) vapor pressure deficit (hPa) in a model where SPI is used in lieu of accumulated precipitation. Data between the innermost 95th percentiles are plotted to focus on the robust signals.

S2.20 Time dependence of maximum temperature for maize. Red line denotes the best fit line through the data using a simple linear regression. In the bottom right of each panel, the first blue number represents the average change in maximum temperature from 1962-2014 computed with the slope of the trend line. The second blue number in parentheses represents one standard deviation of the year-to-year variability in maximum temperature.

S2.21 Same figure as S2.20, but for groundnut.

S2.22 Same figure as S2.20, but for sorghum.

3.1 Average estimated planting dates for corn. Dark colors represent earlier in the year, while light colors represent later. We have plotted a planting date for every county in which at least one year reported corn yields that meets the irrigation criteria.

3.2 Predicted compared to actual yields for each crop.

3.3 Partial dependence plot for phase-specific accumulated precipitation (PRCP) and vapor pressure deficit (VPD) for each growth phase, including the growing season. Colors represent the growth phase: establishment (blue), critical window (green), grain filling (red), growing season (black). Responses from the 2.5-97.5th quartiles are plotted.

3.4 Same as above, but for maximum daily temperature (TMAX) and minimum daily temperature (TMIN).

3.5 Partial dependence plot for phase-specific extreme degree days (EDD) accumulated for each growth phase, including the growing season. Colors represent the growth phase: establishment (blue), critical window (green), grain filling (red), growing season (black). Responses from the 2.5-97.5th quartiles are plotted.
3.6 Optimal temperature and precipitation ranges are plotted for corn in 2016 (top) and 2012 (bottom). Counties which lie within the optimum TMAX and TMIN ranges are highlighted in red and blue, respectively. Counties which experienced more than 400 mm of precipitation are colored grey. Counties that experienced optimal growing conditions are therefore darkest (dark purple).

S3.1 Dual partial dependence plot for TMAX and PRCP for sorghum.
S3.2 3-dimensional partial dependence plot for corn with vapor pressure deficit averaged over the critical window and accumulated precipitation over the growing season. VPD-CW and PRCP-GS are plotted in the horizontal, xy-plane, and the average modeled yield on the vertical, z-axis. Color gradient aligns with average modeled yield.

S3.3 Segment plot depicting average growing season for sorghum. If a state has no data, then sorghum is either not planted there, or is planted in counties with harvest area irrigation over 25%. Blue represents the establishment phase, green represents the critical window, red represents the grain filling phase, while black represents a two-week drying period (not included in this study).

S3.4 Same as Figure S3.3, but for corn.
S3.5 Same as Figure S3.3 and S3.4, but for soybean.

4.1 Region of interest plotting corn yield from 2005 in kg/ha. States without county outlines are not included in the analysis. Counties with missing yield data (white) are due to missing or irrigated yields.

4.2 Station locations for GHCND data (blue plus signs) and ISD data (red diamonds).

4.3 Segment plot depicting average growing season for winter wheat. Blue represents the establishment phase, green represents the critical window, red represents the grain filling phase, while black represents a two-week drying period (not included in this study).

4.4 Log plot of the total number of dust events during the growing seasons of corn from 2002 to 2016. (top left) MODIS (top right) GHCND - WT07 reports (bottom left) ISD - NWS Dust Storm Warning (bottom right) ISD - NWS Blowing Dust Advisory.

4.5 Experimental structure used to check for a detectable effect of dust events on crop yields.
4.6 Variable importance plot for corn using WT07 (GHCND) as the dust metric. Variables above the dashed line are considered important, as they are in the top third of the variables. The model $R^2$ is also included in the bottom left. Note that the exact order of the variables is not necessarily informative, changing different parameters of the model (e.g. $m_{try}$) will shift the metrics around slightly. \textbdseries{} \textbf{136}

4.7 Partial dependence plot for dust event frequency (derived from MODIS) in the phase identified by the model as being most important in the state-by-state analysis. Y-axis represents average yield in kg/ha. All values are plotted (not the innermost 95\textsuperscript{th} percentile). \textbf{137}

4.8 Partial dependence plot for AOD for each growth phases, including the growing season. Colors represent the growth phase: establishment (blue), critical window (green), grain filling (red), growing season (black). Responses from the 2.5- 97.5\textsuperscript{th} quartiles are plotted. These crop models include the MODIS dust metric, but the plots for AOD do not change if the metric is changed or removed. \textbf{138}

4.9 Average number of dust events per day over the region of interest and time series. Vertical grey lines are included every 50 days. Black line represents the centered 30-day rolling average. This represents the annual cycle of dust events. Colored segments from the top down denote the average growing season over the region for sorghum, corn, soybean, and winter wheat, respectively. The first (blue) segment denotes the establishment phase, the second (green) denotes the critical window, the third (red) denotes grain filling, and the final short black segment represents drying, which can occur before or after harvest. \textbf{144}

S4.1 Maximum temperature averaged from 1980 to 2016 for MetData, GHCND, and ISD. Scale units are °C. \textbf{147}

S4.2 State-averaged maximum temperature for MetData (blue), GHCND (red), and ISD (green) for Texas, Wisconsin, Iowa, and North Dakota. Temperatures are °C. \textbf{149}

S4.3 Reporting inconsistencies of WT07. Red line (left y-axis) represents the number of codes issued each year over the region of interest used in this study. Black diamonds (right y-axis) represent the number of stations reporting WT07 each year in the United States. \textbf{150}

S4.4 Seasonality of Deep Blue AOD at 550 nm over the region of interest (1631 counties). Grey polygon bounds the 25\textsuperscript{th} to 75\textsuperscript{th} percentiles. Thick black line denotes 50\textsuperscript{th} percentile. Thick green line denotes 90\textsuperscript{th} percentile, while red points indicate the maximum AOD. \textbf{151}
S4.5 Heat map visualization of the correlation matrix computed for all candidate predictors included in basis models for sorghum. . . . . . 152
S4.6 Heat map visualization of the correlation matrix computed for all candidate predictors included in basis models for corn. . . . . . 153
S4.7 Heat map visualization of the correlation matrix computed for all candidate predictors included in basis models for soybean. . . . . . 154
S4.8 Heat map visualization of the correlation matrix computed for all candidate predictors included in basis models for wheat. . . . . . 155
S4.9 Heat map visualization of the correlation matrix computed for residuals of the fundamental variable model for sorghum against all dust predictors and AOD. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 156
S4.10 Heat map visualization of the correlation matrix computed for residuals of the fundamental variable model for corn against all dust predictors and AOD. . . . . . . . . . . . . . . . . . . . . . . . . 157
S4.11 Heat map visualization of the correlation matrix computed for residuals of the fundamental variable model for soybean against all dust predictors and AOD. . . . . . . . . . . . . . . . . . . . . . . . . . 158
S4.12 Heat map visualization of the correlation matrix computed for residuals of the fundamental variable model for winter wheat against all dust predictors and AOD. . . . . . . . . . . . . . . . . . . . . . . . . . 159
S4.13 Illustration of the instability of partial dependence plots when variable importance is extremely low. The regression in each panel uses a different seed for the RF. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 164
List of Tables

2.1 Abbreviations and descriptions of variables included in the dataset. Variable abbreviations are used when referring to modeled responses, while variable names are generally reserved for discussing general plant responses. ................................................................. 22

2.2 Variable importance for each crop and each of the variables in the final regression model. The number of observations used in each regression is provided beneath each crop. Values in the table represent percent increase in MSE as a result of slightly permuting a particular variable. ................................................................. 25

2.3 Model comparison using $R^2$ (RMSE in kg/ha is reported in parentheses) of the baseline, linear, and Random Forest models in this study. The variables for the Random Forest model are the same as Equation 2.1. We also include the average weather model from Schlenker and Lobell (2010) for comparison with a similar study (where $\bar{h}_{it}$, $p_{it}$, and $c_i$ represent the mean temperature, total precipitation, and country fixed effects, respectively). The data used in this study are not the same as in Schlenker and Lobell (2010), so these models cannot be compared directly. ................................................................. 28

S2.1 Country indices included in manuscript. Asterisks denote a special case: groundnut yields for Lesotho do not exist, so the index for Swaziland becomes 30 in that instance. ................................................................. 60

S2.2 Variable importance metric for all variables initially considered in the analysis. ................................................................. 62

3.1 Potential climate predictors used in this study. Fundamental variables used in the final models are denoted with an asterisk. Throughout the text, each variable may be referenced for an individual phase, in which case the variable will appear as below, but appended with a dash and phase abbreviation: ES, CW, GF, and GS for establishment, critical window, grain filling, and growing season, respectively. 82
3.2 Phase-specific thermal times, base temperature, and planting thresholds for sorghum, corn, and soybean. Asterisk denotes that a frost restriction is in place. The unit °C d represents a degree day.

3.3 Variable importance for the fundamental variable models for sorghum, corn, and soybean. Numbers beneath crop names denote R² values for each model. Variable importance was calculated using the raw, unscaled permutation accuracy importance measure (Strobl et al., 2007). Unscaled data should only be compared within single columns (models). Higher values for raw permutation importance measure indicate more important variables. Average rank of each variable-phase combination is included in the rightmost column; higher ranks denote more important variables.

3.4 Optimal grain filling temperature ranges, and optimum temperature for both TMAX and TMIN during grain filling (derived from partial dependence information plotted in Figure 3.4).

3.5 Comparison of explanatory power (R²) between crop models that only include fundamental variables and those that include all variables from Table 3.1. Models with all predictors also include two variables measuring cold temperature strain during grain filling, while both models include year as a predictor.

4.1 Station-level daily variables extracted from GHCND database. Asterisks denote a variable computed from the GHCND station data. These data were computed for each day and each station, which were then aggregated to form county-level statistics.

4.2 Station-level daily variables extracted from the mandatory variables reported in the ISD database.

4.3 Phase-specific information used to compute the growth phases for winter wheat. T_start denotes the variable and corresponding value for each phase. T_start is accompanied by an arrow to indicate whether the threshold is computed for falling (↓) or rising (↑) temperatures. The phase-specific length and base temperature are highlighted in the Duration and T_base columns.

4.4 Comparison of explained model variance using all variables for each data set in a crop-specific regression. MetData model contains all variables listed in Table 3.1, including year. GHCND model contains all variables listed in Table 4.1, and ISD model contains all variables listed in Table 4.2.
4.5 A representative example of change in R² imparted by removing the dust predictor from a model. In this example, we used the core model variables, the MODIS dust metric and included AOD.

S4.1 R² values for each crop when reports of WT07 between 2008 and 2010 are. Top row includes the model with AOD, the bottom does not. In every instance, dust is still the least important predictor.

S4.2 Variable importance for sorghum basis models. Variable importance was calculated using the raw, unscaled permutation importance. Unscaled data should only be compared within single columns (models). Numbers beneath crop names denote R² values for each model.

S4.3 Variable importance for corn basis models. Variable importance was calculated using the raw, unscaled permutation importance. Unscaled data should only be compared within single columns (models). Numbers beneath crop names denote R² values for each model.

S4.4 Variable importance for soybean basis models. Variable importance was calculated using the raw, unscaled permutation importance. Unscaled data should only be compared within single columns (models). Numbers beneath crop names denote R² values for each model.

S4.5 Variable importance for wheat basis models. Variable importance was calculated using the raw, unscaled permutation importance. Unscaled data should only be compared within single columns (models). Numbers beneath crop names denote R² values for each model.

S4.6 R² values for the two-stage regression with the MODIS dust metric. The first regression is performed without dust (with and without AOD), while the second regression is performed using the residuals of the first and the dust metric as the only predictor. We test both random forest (RF) and a linear model (LM).

S4.7 We ranked the variable importance of each dust metric during each phase for several different experiments. In each cell, we include the rank for the dust metric during establishment, critical window, grain filling, and growing season, respectively. For each crop and dust metric, there is a small table of four cells. We perform the experiment without AOD included (left column), with AOD included (right column), and for all counties with data (top row) and with no-dust counties removed (bottom row). When AOD is omitted there are 29 variables in the model, and when AOD is included there are 33 variables in the model.

S4.8 Salient results from the state-by-state analysis.
Acknowledgments

This material is based upon work supported by the National Science Foundation through the Network for Sustainable Climate Risk Management (SCRiM) under NSF cooperative agreement GEO-1240507. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

I want to thank my advisor, Chris E Forest. I came to this program with a peculiar love of dust, and though Chris wasnt overly familiar with the field, he learned quickly and supported my research questions. Once I developed the proposal for this PhD, it was even further outside of his realm of expertise, but he supported me every step of the way and I cannot thank him enough for funding and encouraging me to pursue this work.

I also need to thank Armen Kemanian. When I proposed this project, I severely underestimated the amount of plant science I would need to know. Armen, and his entire research group, took me under their wing and supported this work with unfailing enthusiasm and encouragement.

I would also like to thank the rest of my dissertation committee - Marcelo Chamecki, Greg Jenkins, and Natalie Mahowald - for their advice and feedback. A special shout-out to Natalie Mahowald for letting me know up front how difficult this problem would be to answer. She was not wrong.

A hearty thank you to the Forest Research Team: Alex Libardoni, Judy Tsai, Rob Ceres and Kristina Rolph for their invaluable feedback and snacks in times of need.

I also want to express immense gratitude to my family and friends. I could not have done this without the love and support of my parents James and Cheryl Hoffman, my sister Lauren Hoffman, and my boyfriend, Andy Bogus. His support, unwavering positivity, and breakfast food carried me through the tough times. Thank you to all my friends for balancing out the stress and doubt that come with finishing a PhD.
Chapter 1  
Introduction

1.1 Purpose

As the global population continues to increase and global climate continues to change, we are being forced to understand the factors that drive food production and ultimately food security. The effects of temperature, precipitation, and carbon dioxide are relatively well-understood, but the effect of dust on crop yields has yet to be thoroughly investigated. Though often overlooked in food security and crop yield studies, dust has the ability to physically affect crop yields. Many marginal food-producing countries and food insecure countries lie in regions of the world that experience frequent dust erosion events and large atmospheric dust loadings (Figure 1.1). While there have been studies on how certain aspects of the dust cycle, like sandblasting and suspension, may affect plant productivity and growing conditions (Section 1.3.2), no one has identified a signal linking dust to yield on a large scale. In general dust emissions have a negative effect on plants (Section 1.3.2.2), and dust suspension has a positive effect (Section 1.3.2.1). Despite the fact that these two often occur together, their combined effect has yet to be determined; the dominant effect is unknown. In this research project, we focus on determining whether dust has an effect on crop yields.
1.2 Motivation

We motivate this work based on the potential investments and food security risks that require an informed basis of understanding. Determining whether dust affects crop yields can inform climate change adaptation investments and food security risk mitigation strategies. The ability of an agricultural region to meet the food
demand, which depends on the environmental conditions as well as economic and sociopolitical factors, constitutes a critical component of food security (FAO, IFAD, and WFP, 2014). Changes in climate directly impact food security by altering the growing conditions, and indirectly by affecting income distributions, growth, and conflict (Schmidhuber and Tubiello, 2007; Raleigh et al., 2015).

As a secondary motivation, this work aims to fill a knowledge gap. It is well-established that climate affects both dust emissions and food production (Sections 1.3.1 and 1.2.1). Dust emissions also have an affect on climate, and agriculture has been linked to an increase in anthropogenic dust emissions. However, whether dust has an effect on agriculture has yet to be thoroughly investigated (Figure 1.2).

![Figure 1.2: Flow chart describing knowledge gap.](image)

### 1.2.1 Existing research on climate and crop yields

Research estimating the impact of climate on crop yields evolved from a field scale to a global scale in the last 25 years or so. As the science behind climate change began to converge, the effects on global crop production were unknown and led to a large-scale effort to understand how climate may affect yields on larger scales (Rosenzweig and Parry, 1994). The effects of the changing climate on crop production and yield are estimated using crop models that fall broadly into one of two categories: process-based and statistical models (Rosenzweig et al., 2014; White et al., 2011; Lobell and Burke, 2010, 2009; Müller and Robertson, 2014; Müller et al., 2011; Roberts et al., 2012). This work uses the latter.

#### 1.2.1.1 Crop Model: Process-based crop models

Process-based crop models are the dominant tool used to estimate the impact of long-term (e.g. 2070-2090) climate change on agriculture by utilizing descriptive equations of physical, chemical, and physiological processes in crop growth over time. These models solve differential equations using mathematical descriptions
of physical, chemical, and physiological processes to simulate crop development. These models require high-resolution planting date, hydrology, soil nutrients and composition status, and planting density as initial conditions. Process-based models provide information for field-scale and daily time steps, so weather conditions (temperature, precipitation, solar radiation, etc.) typically need to be specified or stochastically generated on a sub-daily temporal scale. Many of these models were developed for specific sites, crops, and conditions, and many have been calibrated multiple times for different locations and inputs. Unfortunately, the organization and documentation of the existing site-based models and their permutations is chaotic (Lobell and Burke, 2009). Global gridded crop models (GGCMs) are derived from these site-based models, which are then run at thousands of sites with downscaled climate data input to simulate the biophysical responses of crops to growth and management conditions on scales similar to those of a GCM (Elliott et al., 2014).

Process-based models are subject to the following areas of uncertainty and issues: assessing model performance depends on the accuracy and size of the evaluation datasets, models often require extensive high-resolution input data, limited representation of physical processes, and model variability. These models are also subject to uncertainties associated with limited representation of physical processes; arguably one of the largest problems in these models is the poor representation of plant responses to pests and elevated carbon dioxide (Tubiello et al., 2007; Soussana et al., 2010; Gregory et al., 2009). Model variability and parametric uncertainty in physical crop models also cause significant uncertainty in yields (Lobell and Burke, 2010; Iizumi et al., 2009; Asseng et al., 2013; Bassu et al., 2014). While acknowledging the aforementioned sources of uncertainty, physical crop models are critical to understanding the process-level impacts of individual climate variables and the long-term impacts of climate change on crop yields.

1.2.1.2 Crop Model: Statistical crop models

Statistical crop models rely on the mathematical relationships between historical crop production and weather data. These models are typically applied on larger spatial scales than process-based models (i.e. larger than a field), but can be applied at all spatial scales (Lobell and Burke, 2010). Advantages of statistical models are their limited dependence on calibration data, relatively simple uncertainty
assessments, and ability to potentially capture effects, processes, or interactions that are not well understood. While all of these factors played into the decision to use statistical models in this work, the ability to identify effects or interactions that are not well understood (e.g. dust effect on crop yields) was the primary reason.

Before building statistical models, four notable decisions or issues that must be addressed (Lobell and Burke, 2009). First, yield and climate data should be chosen based on spatial and temporal extent. Second, the positive trend in crop yields is primarily due to technological improvements and adaptations. This is often accounted for by approximating the technology trend and detrending the data. Third, selecting the predictors to put in statistical models is quite difficult. When dealing with weather data, covarying and interacting variables become an issue. Fourth, there is a choice in how to prescribe the functional dependence on predictor variables, or how yield relates to climate (or other) predictors.

Statistical models also have notable shortcomings like low signal-to-noise ratios, limited extrapolation, and assumptions of stationarity and no explicit adaptation. Data quality is often an issue in statistical models as well as in process-based crop models.

1.3 Scientific justification of research question

1.3.1 Dust

Dust plays a complex role in the climate system, biogeochemical cycles, and human environments. Once suspended in the atmosphere dust impacts the radiative budget directly via scattering and absorption of radiation, and indirectly by modifying the cloud optical properties and acting as cloud condensation nuclei, which can affect precipitation patterns (Mahowald and Kiehl, 2003). Dust has a local and global cooling effect and the ability to impact the global carbon cycle by providing micronutrients to nutrient-deficient ecosystems via deposition (Evan and Mukhopadhyay, 2010; Jickells et al., 2005; Mahowald et al., 2009; Okin et al., 2008). In addition to these impacts on climate and biogeochemical cycles, dust also affects human environments and health. For instance, major dust events can halt transportation and shipping of goods, while consistently high dust-loads have been linked to premature mortality and meningitis outbreaks (Fang et al., 2004;
1.3.1.1 General Characteristics

The majority of mineral aerosols are emitted from arid and semi-arid regions that lie near the descending branches of the Hadley cells, in what is frequently referred to as the “Dust Belt” (Pye, 1987). However, while not directly in these decent regions, a significant amount of dust is emitted from Australia, China, and parts of the United States.

Dust emission caused by wind erosion is complex and depends on numerous surface characteristics and atmospheric conditions like soil properties, vegetation cover, atmospheric stability, and wind speeds. Dust is typically emitted via sandblasting as a result of saltation of larger particles (Shao et al., 1993; Marticorena and Bergametti, 1995; Gillette and Passi, 1988; Zender et al., 2003; Kok et al., 2012, 2014). Dust emission occurs when friction velocity exceeds the threshold friction velocity. Various factors affecting threshold friction velocity are atmospheric stability, soil composition, and vegetation (Shao et al., 1993; Marticorena and Bergametti, 1995; Zender et al., 2003; Kok et al., 2012, 2014).

Mineral aerosols have undergone historical shifts in time and space due to both natural and anthropogenic forcings. Records indicate that mineral aerosols are sensitive to natural climate variability, with recent glacial periods experiencing dust loadings between 2-20 times larger than interglacial periods at high latitudes (Mahowald et al., 2006; Mahowald, 2007). Another distinct shift in the record of Sahelian dust aligns with the onset of agriculture in the Sahel, indicating that humans began modifying the dust cycle approximately 170 years ago (Neff et al., 2009; Mulitza et al., 2010; Singh et al., 2012). Approximately 20-25% of mineral aerosols are from anthropogenic sources (i.e. irrigation diversion creating ephemeral lakes, disturbing soils, removing plant cover, or indirectly by altering the climate and hydrologic cycle) (Ginoux et al., 2012), although estimates range from -20% to 60% (Tegen and Fung, 1995; Prospero et al., 2002; Tegen et al., 2004; Woodward et al., 2005; Moulin and Chiapello, 2004; Mahowald et al., 2010).

In addition to historical changes in dust loadings, sources of interannual variability in dust emissions are important for interpreting the crop yield responses. Mineral aerosols from northern Africa are correlated with the North Atlantic Oscillation (NAO), El Niño/Southern Oscillation (ENSO), and the Pacific Decadal Oscillation.
(PDO) (Brooks and Legrand, 2000; Mahowald et al., 2003; Evan et al., 2012; Wang et al., 2012; Hoffman et al., 2014; Hand et al., 2016), meaning dust emissions are linked to well-known sources of interannual climate variability.

In the next two sections, we review the physical science behind why mineral aerosols should be considered as a possible driver of crop yield variability in food-insecure countries of the world.

1.3.2 Aerosol and erosion effect on plants

This research investigates the dependency of crop yields on dust and dust events. A considerable body of work addresses the effect of wind erosion on crop growth and productivity, while another has focused primarily on the effect of suspended aerosols on plant productivity. We investigate the total impact of mineral aerosols on crop yields as the aggregate impact of both aerosols and soil erosion.

Mineral aerosols can affect vegetation by altering the amount of photosynthetically active radiation (PAR, 400-700 nm), and by modifying the surface radiative balance. PAR is absorbed by green vegetation and converted to biomass via photosynthesis, while changes in the amount of shortwave and longwave radiation affect surface growing conditions. Past studies have considered the impact of aerosol types like biomass burning smoke, volcanic aerosol, and anthropogenic pollution on PAR, but the impact of suspended mineral aerosols on PAR has not been thoroughly studied (Section 1.3.2.1).

Within the context of agriculture, various stages of the dust cycle are studied under the umbrella of wind erosion. Soil erosion is costly to the agricultural industry so the impact of eroding soil particles on plants is well-studied, however the research is often crop-specific and conducted on fine temporal and spatial scales. Nevertheless, we outline the general crop responses to erosion that can be applied on scales we aim to consider throughout this research in Section 1.3.2.2.

1.3.2.1 Aerosols & plants

Changes in the effects of aerosols on PAR are governed by aerosol type, loading, and optical properties which is important when considering the impact of mineral aerosols because they are highly heterogeneous in composition, size, and shape. Increased dust loading in the atmosphere linearly increases aerosol optical depth
(AOD), but the slope depends on the mineralogy and size distribution of the dust (Ohde and Siegel, 2012; Xi and Sokolik, 2012). Similarly, while the magnitude varies with dust optical properties, enhancing the dust load linearly decreases total PAR, but increases the diffuse component of PAR until it reaches a maximum (when dust aerosol depth at 0.5 μm is approximately equal to 1) and begins to decrease (Ohde and Siegel, 2012; Xi and Sokolik, 2012).

The direct effect of increased AOD on plants is the decrease in total PAR reaching the plant as a result of light attenuation, reducing the photosynthetic rate and primary production in plants (Chameides et al., 1999). However, the increased diffuse component of PAR, or indirect effect, can lead to a higher photosynthetic rate and increase the productivity of a plant. This response is referred to as the diffuse radiation “fertilization” effect and occurs because sunlit leaves receive both direct and diffuse components of radiation are often light saturated, but shaded leaves only receiving diffuse light, are often light limited. Plant productivity increases with the increasing diffuse fraction because light starved leaves experience higher light intensities, while the impact of aerosols on the light saturated leaves remains relatively unchanged (Cohan et al., 2002; Gu et al., 2002, 2003; Xia et al., 2007; Wohlfahrt et al., 2008; Xi and Sokolik, 2012). However, once a critical AOD threshold is reached, the decrease in total PAR dominates the changes in diffuse radiation and all the leaves become light starved.

The total impact of aerosols on net primary productivity (NPP) also depends on cloud cover, canopy architecture, and photosynthetic pathways. For overcast and clear skies, as AOD is increased from a background state of 0.05 to 1, total PAR decreases by 25% and 30%, while the diffuse fraction increases by 5% and 200%, respectively (Greenwald et al., 2006). On overcast days with high AOD, aerosols will impede NPP because the positive effect of increased diffuse radiation is rendered negligible (as the diffuse fraction of radiation is already large), and the decrease in total PAR dominates the response (Cohan et al., 2002; Greenwald et al., 2006). Similarly, the increase in diffuse radiation (due to increased cloud cover and aerosol load) associated with global dimming from 1960-1990 likely enhanced the terrestrial carbon sink by nearly a quarter (Streets et al., 2006; Mercado et al., 2009; Wild et al., 2012). NPP also increases with AOD in forests and crop lands, but decreases in grasslands. Grasslands may not experience a significant benefit from increasing diffuse radiation due to the seasonal stresses experienced by grasslands.
and/or due to the canopy architecture being shallow and open (Niyogi et al., 2004; Wohlfahrt et al., 2008; Wild et al., 2012). C4 plants have higher light saturation points and are less sensitive to changes in diffuse PAR than C3 plants are, so C4 plants like corn are less sensitive to the aerosol type and optical properties that govern changes in the diffuse component of radiation (Xi and Sokolik, 2012).

Mineral aerosols decrease the shortwave radiation (SW; reducing surface temperature and sensible heat) and increase the longwave radiation (LW; increasing temperature) at the surface. Overall, mineral aerosols depress temperatures at the surface, even though enhanced longwave radiation compensates for 20-33% of the cooling caused by the decreased downward SW flux (Xi and Sokolik, 2012). In bright conditions where leaf temperature can exceed the optimal temperature for photosynthesis, this temperature depression is more important to plant productivity than the diffuse radiation fertilization effect (Steiner and Chameides, 2005). Concurrent decreases in temperature, vapor pressure deficit (VPD), evapotranspiration, and plant water stress were found to amplify the diffuse fertilization effect (Stanhil and Cohen, 2001; Wohlfahrt et al., 2008; Estes et al., 2014). Whether increased productivity associated with aerosols is due to the redistribution of light, interactions between air temperature, humidity and VPD and the redistribution of light, or purely thermal responses, is an area of ongoing research (Steiner and Chameides, 2005; Wohlfahrt et al., 2008).

The deposition of aerosols on plant leaves can also affect vegetation productivity by deceasing the amount of PAR reaching leaves (Armbrust, 1986; Dahlman et al., 1972; Bergin et al., 2001). Bergin et al. (2001) estimated that over a two month period, deposition of water insoluble aerosols might account for a 35% reduction in total PAR. Dust can also affect plant function by blocking stomata, although the relevance of this issue depends on the size of the particles. The seminal paper on this topic suggested that this effect is relatively unimportant (Farmer, 1993), but recent work indicates that yields can decrease by 28% due to a 30% decrease in stomatal conductance (Hirano et al., 1995; Zia-Khan et al., 2015). Additionally, deposition of dust on leaves increases leaf temperatures (Hirano et al., 1995; Zia-Khan et al., 2015).

Finally, deposition of dust can also occur in irrigation canals because nearly 30% of deposition from an eroding source field settles out within 50m (Hagen et al., 2007). Infilling of irrigation ditches can lead to reduced yields.
1.3.2.2 Soil erosion & plants

Soil erosion can occur as a result of precipitation, combined with geometry of the land, and wind (Renard et al., 1997; Fryrear et al., 2001). Though soil erosion by wind is thought to occur in arid and semiarid areas, it can be an issue whenever conditions are right (e.g. loose and dry soil, smooth soil, large fields, and strong wind). Wind erosion affects crop growth in three ways: erosion injures plants via abrasion by soil particles, decreases productivity as a result of soil degradation (nutrient loss), and causes undesired sedimentation of soil particles on plants or in irrigation channels. Below, we outline the impacts of soil erosion on plants, although the magnitude of each impact varies by crop.

Yield depression as a result of wind erosion is relatively well-documented. Windblown soils can damage plants when erosion is local (i.e. from the field in which crops are planted) (Armbrust, 1968; Armbrust et al., 1973; Fryrear et al., 1975; Armbrust, 1982, 1984; Michels et al., 1993, 1995; Armbrust and Retta, 2000; Baker, 2007; Baker et al., 2009). Soil particle strikes are directly related to a decreased growth rate and yield, an increase in the number of plants killed, and delayed first bloom (Armbrust, 1968; Armbrust et al., 1973; Fryrear et al., 1975; Baker, 2007). Decreased yields are likely due to the loss of photosynthetically active leaf tissue (defoliation) which causes short-term high-intensity moisture stress resulting from loss of stomata control and epidermis damage (Fryrear et al., 1975; Armbrust, 1982). High levels of erosion have the strongest negative effects on dry weight production when erosion occurs between one and two weeks after emergence, so immature plants are at a higher risk than matured plants (Armbrust, 1984; Fryrear et al., 1975). Abrasion and crop burial result in decreased yield and can necessitate replanting or crop replacement depending on severity (Michels et al., 1993, 1995, 1997; Armbrust and Retta, 2000; Baker, 2007; Baker et al., 2009). In addition to the direct physical damages caused by eroding soil, the resultant damage makes plants more susceptible to pests and pathogens (Fryrear et al., 1975; Claffin et al., 1973; Harvell et al., 2002).

A major consequence of soil erosion is soil degradation via the loss of nutrients and soil organic carbon which reside primarily in the topsoil (Stoltenberg and White (1953); Zobeck and Bilbro (2001); Van Pelt and Zobeck (2007), among others). Agricultural areas susceptible to wind erosion and soil degradation include northern and southern Africa, western and eastern Asia, Australia, parts of South
America, the Great Plains and the Mojave Desert (Zobeck and Van Pelt, 2014; Armbrust and Retta, 2000; Okin et al., 2001; Food and Agriculture Organization of the United Nations, 2014). As a soil erodes it becomes rougher and more coarse, which impacts the stability of the soil and its capacity to hold water, resulting in a positive feedback on erosion rates (Sterk et al., 1996; Sterk, 2003; Zobeck and Bilbro, 2001). As a result of this feedback, prolonged soil erosion can lead to desertification (Ravi et al., 2010; Field et al., 2010).

1.3.3 Summary

Wind erosion leads to dust emission, and though soil erosion has a decidedly negative effect on crop yield, dust can affect yields in several different ways. For instance, suspended dust aerosols change the amount of total PAR and fraction of diffuse radiation reaching a plant, which affects productivity and yield. Dust can also affect the surface radiative budget and impact the growing conditions at the surface like temperature, humidity, and evapotranspiration. Wind erosion can result in sandblasting which can injure the plant through partial defoliation and result in high-intensity moisture stress, decreasing crop yields. Dust may also settle onto plant leaves which decreases the total PAR reaching the plant, and altering the surface radiative balance of the leaf. Indirect effects of mineral aerosols on crop yield occur via accumulation of sand/dust in irrigation canals, the ability of dust to act as a vector for diseases, and changes in local climate caused by dust acting as cloud condensation nuclei which impacts precipitation or as elevated heat sources which impacts atmospheric stability (Miller and Tegen, 1998; Zhao et al., 2011).

We motivate this study by the fact that much of the research estimating the impact of climate change on food security and crop yields only includes temperature, precipitation, and carbon dioxide, but dust is a major climatological component in many food insecure countries (Figure 1.1). Prioritizing climate change adaptation for food insecure countries may depend on dust, a variable thus far emitted from all studies. Based on the literature we can confidently say that dust has the ability to impact crop yields and this research attempts to identify this effect in historical yield observations. Before beginning this work, our hypothesis was that dust has a small, but distinct effect on yields.
1.4 Evolution of research agenda

Our proposed objective is to understand and identify the effect of dust on yields. This work was catalyzed by Lobell et al. (2008) which used a simple linear regression model using two predictors, temperature and precipitation, to address food security in the near future in food insecure regions. This research began with a similar experimental set up. First, we sought to reproduce the results of Lobell et al. (2008). Then, we intended to use a linear model to detect whether dust had an effect on yields and use the regression to project this effect forward to understand the impact of dust on yields in the future. Using gridded weather data and satellite-based optical depth measurements as a (simple) proxy for dust, we quickly realized that a dust signal was going to be more elusive than a temperature or precipitation signal. The yield response to dust was either smaller than other climate predictors and/or was obscured by the interactions with other predictors and noise in the data. We also realized that the response to dust was likely nonlinear, a feature that would not be picked up by a linear regression. Therefore our first objective shifted from simply detecting a dust signal on crop yields, to developing or finding a tool that is capable of detecting a dust signal (i.e. the effect of dust on yields).

If there is an effect of dust on yields, or a ‘dust signal’, it should be present in the underlying climate-yield data. Therefore we sought a statistical tool that could effectively act as a diagnostic crop model and learn from the data. Machine learning is a growing area of research for exploratory data analysis, but while machine learning algorithms are often accurate, many are difficult to interpret. For instance, deep neural networks are highly accurate, but frequently get referred to as a ‘black box’, while more transparent models like linear regression are less accurate and more restrictive. Random Forest (RF) is often used as a diagnostic model because it is one of the most transparent and flexible supervised learning methods available (Breiman, 2001). While machine learning is not a substitute for experimental design, it can prove to be helpful in theory development. RF can approximate non-linear functional forms between predictors and the outcome variable in the data, as well as identify potential interactions among variables. Often used on extremely large data sets with high number of predictors, RF is a useful tool in identifying which variables are important to predicting the outcome variable. The details of RF are outlined in depth in Chapter 2. While no perfect
tool exists, we use RF throughout this work to identify how yields might respond to dust.

Before this tool could be used to detect the dust signal, its ability to detect more significant yield responses (i.e. temperature and precipitation), needed to be assessed. This work was motivated by the potential effect of dust on food security, so Chapter 2 employs RF as a diagnostic crop model to study the drivers of crop yields in sub-Saharan Africa (SSA).

During the development of work for Chapter 2, another issue became clear - data quality and resolution was going to affect the ability of RF to detect how dust may be affecting yields.

It is very difficult to obtain yield data on a sub-national scale for SSA. As a result, it was necessary to aggregate the climate data over each of the countries because we only had access to national yield data. While we accounted for areas in which the crops were grown, extensive spatial aggregation likely obscured fine-scale signals like dust. As in many statistical crop models, this project used monthly climate data aggregated over the growing season of each crop. Temporal aggregation may also obscure fine-scale signals and possible variable interactions, which are critical to understanding the impact of dust on yields. Though never published, a small side project was important to informing the direction of this work. In this mini-project, hourly station data in sub-Saharan Africa were extracted and used to build a model analogous to the one described in Chapter 2. The station data included weather codes that identified days with dust events as well as wind and visibility data which were combined to create dust storm proxies. Using these data, we aggregated over the country and the growing season to try and detect a yield response to dust. We found no evidence of a dust signal, and concluded that spatial and temporal aggregation are likely obscuring the signal. To detect a signal as subtle as the dust effect on yields, we needed higher quality and higher resolution data.

In Chapter 3 we apply RF as a diagnostic crop model in the United States where data quality and resolution are significantly higher. In particular, yields are reported on a county-level as opposed to a country-level, so the extent of spatial aggregation is significantly reduced. Gridded weather products are also of higher quality than in the developing world because station data is used to create these data sets and developed countries have higher station density. As addressed in the
previous sections, we consider the dust signal of interest as the aggregate effect of both dust emission (wind erosion of soil) and dust suspension. In general, plants are negatively affected by wind erosion, which occurs when vegetation is low and plants are young, but positively affected by dust suspension when vegetation is high and diffuse radiation can infiltrate the plant canopy more than direct radiation. Reliable daily data can allow for the growing season of each crop to be partitioned into individual growth phases to be accurately estimated. By partitioning the growing season into phases, we increase the probability of detecting an effect of dust on yields. But first, validation of this new data and method of partitioning the growing season into physiological stages must be validated. Chapter 3 validates the RF model using high quality and high resolution data in the Midwestern United States and determines whether improved data allows the model to identify more nuanced crop responses to climate variables.

While there are limitations to using a data-driven model, Chapter 3 validates that RF can detect more refined signals in the crop record of the United States. Though the dust emission and loadings in the United States are lower than those in food insecure regions of the world, data quality is higher and the RF model has been validated. Chapter 4 finally addresses whether there is a detectable dust signal in the recorded crop yields in the Midwestern United States.

The dissertation is outlined as follows: Chapter 2 is a proof of concept and validation of RF as a diagnostic statistical crop model in sub-Saharan Africa. Chapter 3 is validation of RF in the United States using high quality and high resolution data to detect phase-specific yield responses. Chapter 4 includes dust as a predictor and determines whether the effect of dust can be detected in historic crop yields.
Chapter 2  
Analysis of climate signals in the crop yield record of Sub-Saharan Africa

Published in Global Change Biology in August 2017 (Hoffman et al., 2017)

2.1 Abstract

Food security and agriculture productivity assessments in sub-Saharan Africa (SSA) require a better understanding of how climate and other drivers influence regional crop yields. In this paper, our objective was to identify the climate signal in the realized yields of maize, sorghum, and groundnut in SSA. We explored the relation between crop yields and scale-compatible climate data for the 1962-2014 period using Random Forest, a diagnostic machine learning technique. We found that improved agricultural technology and country fixed effects are three times more important than climate variables for explaining changes in crop yields in SSA. We also found that increasing temperatures reduced yields for all three crops in the temperature range observed in SSA, while precipitation increased yields up to a level roughly matching crop evapotranspiration. Crop yields exhibited both linear and nonlinear responses to temperature and precipitation, respectively. For maize, technology steadily increased yields by about 1% (13 kg/ha) per year while increasing temperatures decreased yields by 0.8% (10 kg/ha) per °C. This study
demonstrates that although we should expect increases in future crop yields due to improving technology, the potential yields could be progressively reduced due to warmer and drier climates.

2.2 Introduction

Sub-Saharan Africa (SSA) is particularly vulnerable to climate-driven shortages in food production. Impact assessments used to prioritize adaptation and mitigation strategies should benefit from understanding how different climate variables, as well as improving technologies, affect country yields in the region. As of 2014, agriculture accounted for 17% of this region’s GDP (World Bank, 2016), providing food but also surplus production which facilitates trade, boosts local economies, and diversifies nutrient intake (Morton, 2007; FAO, 2015). Understanding the impacts of climate factors on crop production in this region is critical because climate change may impose external penalties on food security and the economy of these nations (FAO, IFAD, and WFP, 2015).

Numerous studies project detrimental impacts of climate on crop yields in SSA (Jones and Thornton, 2003; Lobell et al., 2011a; Thornton et al., 2011; Waha et al., 2013), particularly when adaptation is omitted. While crop yields are responding to changes in temperature and precipitation patterns, these responses occur simultaneously with technological advancements which increase crop yields. Thus, it is challenging to account for the interacting factors that will determine the future productivity of each country in the region. Both process-based and statistical models of crop yield are often used for such purpose.

Process-based crop models represent the mechanisms that control plant growth on a field-scale. Currently, these models are applied to assess large-scale impacts of climate change on crops (Challinor et al., 2014; Sultan et al., 2014; Lobell and Asseng, 2017). This approach requires either identifying climate and soil clusters representative of large areas, so that results may be applied to larger domains, or performing data-dense simulations that require aggregation according to the spatial domain of interest (Auffhammer et al., 2013; Angulo et al., 2014; Ewert et al., 2015; Hoffmann et al., 2016). These models are also subject to initial condition uncertainties (i.e. soil properties and management constraints), which is problematic in data-deficient regions like SSA.
As an alternative to process based models, statistical models tend to operate at larger scales which are compatible with the scales of reported yields. Statistical crop models used to estimate the impact of climate on crop yields and food security have historically been relatively simplistic linear models aimed at projecting yields on global and/or regional scales (Lobell and Field, 2007; Lobell et al., 2008, 2011b). To use these models with relatively short temporal ranges, variable selection must occur a priori to isolate important variables and reduce collinearity, as linear models can be easily overfit (Lobell and Burke, 2009; Lobell et al., 2008). Temperature and precipitation are the most commonly used yield predictor variables, with crop yields relating negatively to increasing temperatures both linearly (Lobell et al., 2008) and nonlinearly (Schlenker and Roberts, 2009) and positively to increasing precipitation both linearly (Lobell et al., 2008) and nonlinearly (Sadras et al., 2011).

While other climate predictors like extreme degree days and vapor pressure deficit are used in some process-based crop models (Stöckle et al., 2014), including these variables in statistical crop models did not occur until recently (Schlenker and Roberts, 2009; Schlenker and Lobell, 2010; Lobell et al., 2011a, 2013, 2014; Roberts et al., 2012). At high levels of warming, each additional degree (computed as extreme degree days) causes rapid, nonlinear yield loss (Lobell et al., 2015). Plant growth also responds negatively to increasing vapor pressure deficit, usually through a decrease in the biomass produced per unit of water transpired (Tanner, 1981; Kemanian et al., 2005). It is expected that warming may increase vapor pressure deficit particularly during heat stress events above 30°C, and thus temperature and vapor pressure deficit effects can be confounded and difficult to disentangle with statistical models. When soil water does not limit transpiration, a high vapor pressure deficit increases transpiration and reduces the canopy temperature well below that of the air (Idso et al., 1982), but in water-limited conditions the increased transpiration demand is not met and crop canopies experience both a high water deficit and a temperature several degrees above that of the air, reducing growth and yield. While most process-based crop models may not properly capture the interaction between crop water use and temperature (Webber et al., 2017), statistical crop models can be used to extract the realized interactions as well as the nonlinear yield responses to temperature, vapor pressure deficit, and precipitation.

The extent to which the representation of variable interactions and nonlinear yield responses are relevant may depend on the scale of the analysis. Over large
spatial and temporal scales, linear regression is often justified because the nonlinear dependence of crop yields on climate variables are diluted by averaging over space and time (Schlenker and Roberts, 2009). Many studies have investigated the impact of spatial and temporal aggregation of climate data on process-based crop models as a source of prediction bias, but the impact on statistical crop models remains unclear (Pierce and Running, 1995; Baron et al., 2005; Tack et al., 2015; Hoffmann et al., 2016).

We employed a machine learning method, Random Forest (Breiman, 2001), to understand the crop yields dependence on technology and climate in SSA from 1962 to 2014. Random Forest is a nonparametric regression technique that can be used for both detecting the sensitivity to predictive variables and identifying the functional form of these relations. Specifically, we address whether nonlinear crop yield responses to environmental factors are manifested in aggregated data for 35 countries in SSA based on 53 years of yield and scale-compatible climate records. This report focuses on maize, groundnut, and sorghum, which are the most important sources of calories and protein in SSA (Waha et al., 2013). Because food security and climate impact assessments have focused on regional scales (Lobell et al., 2008; Giorgi and Francisco, 2000), we consider four regions in SSA. Out of the 35 countries considered in this report, twenty-four have undernourishment rates greater than 10% (Figure 2.1).

We demonstrate that despite the paucity and country-level aggregation of data from SSA, Random Forest allows an in-depth analysis that provides a deeper understanding than previously used parametric methods. It reveals the underlying dependencies contained in the yield record and discriminates between technology and climate signals to enable more accurate projections of their impact on crop yield.
Figure 2.1: The 35 countries from the four regions of interest used in this study. Countries with greater than 10% undernourishment rates are denoted by solid colors, all other countries are hatched (FAO, IFAD, and WFP, 2015).
2.3 Materials & Methods

2.3.1 Data

We developed a climate-crop database from 1962 to 2014 for all countries following standard practices that account for variations in crop area and growing season dates (e.g., Lobell and Field (2007); Lobell et al. (2008, 2011b); Schlenker and Lobell (2010); Iizumi et al. (2013); Lobell and Tebaldi (2014); for additional details see Section S1 in the supplement). For each crop, we generated a dataset that includes climate variables indexed by year, latitude, and longitude on a 0.5° x 0.5° grid. For each crop, these data were averaged by country and growing season to create a panel dataset that includes both spatial and temporal variation (Section S1). While this averaging removes intra-country climate and crop variability, it produces records that are spatially compatible with crop yield records.

2.3.1.1 Crop yield

The Food and Agricultural Organization of the United Nations (FAO) (Food and Agriculture Organization of the United Nations, 2016) provides crop yield data in units of kilogram of grain yield per hectare (kg/ha). Annual crop yields are supplied to the FAO by individual governments in the form of national questionnaires or publications, and as a result, are not always reliable and may contain biases. These biases, however, are likely to be uncorrelated from country to country. We flagged and removed data for countries which repeat the same crop yield for three (or more) consecutive years.

2.3.1.2 Crop area

Crop area fractions represent the fraction of a grid cell (5’ x 5’) used to grow a given crop in the year 2000 (Monfreda et al., 2008). We used crop area data to weight the grid cells, so that climate data were only considered in places where a given crop has been grown. We initially discriminated between irrigated and rainfed regions; however, we removed this masking because the FAO database does not distinguish irrigated and rainfed yields.
2.3.1.3 Crop calendar

Crop calendar data represent the average day of the year in which planting and/or harvest of a given crop occurs in a grid cell (5’ x 5’) representative of the year 2000 (Sacks et al., 2010). We used crop calendar data to compute the growing-season cumulative or average value (as appropriate) for each climate variable, and thereby, only considered the periods in which a crop was actually growing.

2.3.1.4 Climate data

We used gridded monthly data from the Climate Research Unit of the University of East Anglia (CRU TS 3.24.01) at 0.5° x 0.5° spatial resolution produced for years 1901-2014 (Harris et al., 2014). These data are an archive of monthly mean variables produced using observational data from weather stations around the world. Crop yields primarily respond to temperature and precipitation provided that other soil properties are not limiting for agriculture (e.g. soil depth, salinity, rocky outcrops), but other variables (e.g. vapor pressure deficit and radiation) can improve the explanatory power of simple models (Roberts et al., 2012). Because the explanatory power (or variable importance) of individual climate variables has not been studied systematically before in statistical crop models, in the first stage of our analysis, we assessed ten unique climate variables (Table 2.1). In addition to the CRU variables, we also computed vapor pressure deficit, aridity index, and the standardized precipitation index, or SPI (McKee et al., 1993). SPI was used as a measure of soil moisture status over the growing season.

2.3.2 Climate-crop dataset

**Crop Area:** First we aggregated the crop area fractions to the 0.5°x 0.5° climate grid by summing the fractions within the larger grid cell.

**Growing Season:** We defined the growing season as the 120 days prior to the harvest day using the crop calendar data from (Sacks et al., 2010). This deviates from a more common methodology, which calculates the growing season as the period between average planting and average harvest date (Lobell et al., 2008, 2011b). We used the cumulative precipitation (\(pre\)) and estimated reference evapotranspiration (\(ETo\)) over the growing season. We computed the aridity index (\(aridity\)) as the ratio \(pre/ETo\). For each crop, we created a 24 month time-series
Table 2.1: Abbreviations and descriptions of variables included in the dataset. Variable abbreviations are used when referring to modeled responses, while variable names are generally reserved for discussing general plant responses.

<table>
<thead>
<tr>
<th>Variable abbreviation</th>
<th>Variable description</th>
</tr>
</thead>
<tbody>
<tr>
<td>tmp</td>
<td>Average growing season temperature (°C)</td>
</tr>
<tr>
<td>tmn</td>
<td>Average minimum daily temperature over the growing season (°C)</td>
</tr>
<tr>
<td>tmx</td>
<td>Average maximum daily temperature over the growing season (°C)</td>
</tr>
<tr>
<td>pre</td>
<td>Average monthly precipitation accumulated over growing season (mm)</td>
</tr>
<tr>
<td>vap</td>
<td>Average growing season vapor pressure (hPa)</td>
</tr>
<tr>
<td>vpd</td>
<td>Average growing season vapor pressure deficit (hPa)</td>
</tr>
<tr>
<td>cld</td>
<td>Average growing season cloud cover (percentage)</td>
</tr>
<tr>
<td>ETo</td>
<td>Reference evapotranspiration accumulated over the growing season (mm)</td>
</tr>
<tr>
<td>aridity</td>
<td>Aridity index calculated by dividing precipitation by reference evapotranspiration (unitless)</td>
</tr>
<tr>
<td>SPI</td>
<td>Three-month average Standardized Precipitation Index derived with a Pearson Type III distribution over the growing season (unitless)</td>
</tr>
<tr>
<td>time</td>
<td>Simulation year</td>
</tr>
<tr>
<td>country</td>
<td>Country index (categorical variable for each of the 35 countries)</td>
</tr>
</tbody>
</table>

for each 5’x 5’ grid cell, to account for growing seasons spanning January of the reporting year (e.g. crop yield reported in 1965 may be grown between November of 1964 and March of 1965). Then, we aggregated this 5’x 5’ grid to the 0.5°x 0.5° climate grid using local area averaging. Next, the 120-day period in each climate grid cell was computed from the monthly climate data from 1962 to 2014 such that we only considered the climate signal during the growing seasons of each crop (Section S1). Finally, we multiplied the climate data by the 24 month time-series for each 5’ x 5’ grid cell (aggregated to the 0.5°x 0.5° grid), and summed
over the 24-month vector. For all variables except \( \text{pre} \), \( \text{ETo} \), and \( \text{aridity} \), this weighted sum was averaged.

**Country-Averaged Climate Data:** The growing season-averaged climate data were next multiplied by crop area fraction (Section S1). The resulting array was a time series of climate data that only considered climate during the growing season of each crop and only in areas where crops are grown. Finally, this gridded climate data was averaged for each crop over each country and year to create a national average for each climate variable.

Unlike many regional statistical crop models, we did not transform the data to remove any common, long-term trends. Statistical crop models which assess the impact of climate on crop yields typically detrend the data to account for the consistent positive impact of improving technology and possibly CO\(_2\) fertilization on crop yields because those trends are not often the signals of interest (Lobell and Field, 2007; Lobell et al., 2008, 2011b; Schlenker and Lobell, 2010; Iizumi et al., 2013; Lobell and Tebaldi, 2014). We were able to use the raw values of both yield and climate without detrending (as discussed next). To account for differences in technology among countries, we included country and time in addition to the climate variables (Table 2.1). The data input into the regression models for each crop is a panel dataset with observations on the country-year level.

### 2.3.3 Statistical methods

We primarily used the Random Forest regression method (Breiman, 2001) although we also used multiple linear regression to compare our results to previous regression techniques in the following section. Significant effort focused on isolating a single Random Forest regression model which works for all crops so that we could analyze and compare the drivers of crop yields in these countries. We briefly outline and describe the Random Forest method below. (For more in-depth descriptions, we refer the reader to Section S3, as well as Breiman (2001), Liaw and Wiener (2002), and Grömping (2009).)

#### 2.3.3.1 Random Forest

Random Forest is a machine learning algorithm for classification and regression that can handle large numbers of variables without assuming functional relations
between predictors and the dependent variable. The Random Forest algorithm has built-in variable importance assessment and error monitoring while allowing the user to ignore distributional assumptions in many cases. The Random Forest method is a special ensemble of classification and regression trees (Breiman et al., 1983). Decision trees are created through recursive partitioning of the data space, wherein each partition is fit with a simple prediction. The Random Forest algorithm is built on the concept of bootstrap aggregation, or bagging, wherein each tree is constructed using a bootstrap sample of the data set and the split at each node is selected from a random subset of $p$ total predictors, resulting in independent trees (Breiman, 2001). Random Forest is called such because it contains many regression trees.

After each regression tree is grown, the individual trees are averaged to form the trained predictor, $F$. In each tree, about one third of instances are left out and these are called the out-of-bag (OOB) data. OOB data are used to estimate the generalization error (i.e. out-of-sample error). Using Random Forest analysis, it is possible to visualize the functional forms relating yield and predictors in the data using partial dependence and feature contribution plots (Welling et al., 2016) allowing users to discover linear and nonlinear responses that would otherwise be missed or difficult to find using other methods. One limitation of a Random Forest model is that predictions are limited to the observed data range and extrapolation requires further interpretation.

### 2.3.3.2 Building the model

Prior to training a Random Forest model, the following two parameters must be set: the number of trees in the forest ($n_{\text{tree}}$) and the number of variables tried at each node ($m_{\text{try}}$). Both of these parameters are discussed in Breiman (2001) and Liaw and Wiener (2002). First, we determined the appropriate number of trees ($n_{\text{tree}}$) by observing the reduction in error rates as $n_{\text{tree}}$ increases. Second, we set $m_{\text{try}} = 3$ and this provided robust results. Using a similar method, combined with cross-validation, we determined that including more than five predictors did not necessarily provide more useful information (Figure S2.7). Interactions and correlations between predictors are common when using climate variables and can be identified through systematic use of partial dependence and feature contribution plots in both two and three dimensions (Section S3.1). We then selected which
Table 2.2: Variable importance for each crop and each of the variables in the final regression model. The number of observations used in each regression is provided beneath each crop. Values in the table represent percent increase in MSE as a result of slightly permuting a particular variable.

<table>
<thead>
<tr>
<th></th>
<th>Maize (1015)</th>
<th>Groundnut (870)</th>
<th>Sorghum (920)</th>
</tr>
</thead>
<tbody>
<tr>
<td>tmx</td>
<td>17.7</td>
<td>28.6</td>
<td>19.9</td>
</tr>
<tr>
<td>pre</td>
<td>23.7</td>
<td>27.4</td>
<td>15.1</td>
</tr>
<tr>
<td>vpd</td>
<td>17.8</td>
<td>21.7</td>
<td>20.1</td>
</tr>
<tr>
<td>time</td>
<td>74.0</td>
<td>62.0</td>
<td>41.5</td>
</tr>
<tr>
<td>country</td>
<td>63.6</td>
<td>89.2</td>
<td>49.0</td>
</tr>
</tbody>
</table>

The variable importance metric (built into the randomForest R package) measures the mean decrease in accuracy when a given variable is permuted (Section S3.1.2). As mentioned before, the OOB data in each of the trees is used to compute the mean squared error (MSE) (which is then averaged across all trees for the model) by comparing predicted vs observed yield. To compute variable importance, the candidate predictors in the OOB data are permuted, and a second MSE is computed in the same manner. The decrease in accuracy that results from permuting a predictor is directly proportional to its importance. This metric indicated that time was the most important predictor in our analysis, followed by country (Table 2.2).

Variables should be included in the model.

A distinct feature of crop yields over the last half century is the consistent positive trend (Figure 2.2a). This trend is generally attributed to technological advances within each country, like breeding and improved management. Carbon dioxide fertilization also contributes to the trend, though the relative magnitude is small when compared to the increased yields due to technology (McGrath and Lobell, 2011). Also, crop responses to elevated CO₂ vary between rainfed and irrigated systems as well as between C3 and C4 crops (Lobell and Burke, 2009; Stöckle and Kemanian, 2009; O’Leary et al., 2015; Gray et al., 2016). Statistical crop modelers typically remove this trend before building models because the long-term technology trend is not the signal of interest (Lobell et al., 2008; Iizumi et al., 2013). However, this may remove important feedbacks within the statistical analysis that alters the estimated interdependencies (Rougier, 2008). The country index accounted for time-invariant effects like the rate of innovation within a country, soil fertility, and
Figure 2.2: (a) Average annual yield for maize, groundnut, and sorghum from 1962 to 2014 over all countries. (b) Partial dependence plot for time. Maize, groundnut, and sorghum are denoted by green, orange, and black lines, respectively. Partial dependence plots graph the independent variable against the model outcome, after accounting for the average effect of other independent variables in the model.

diversity in agricultural practices. Together, the two surrogate variables time and country can accurately represent the increase in technology that is well-documented as the primary reason for increasing crop yields over the last 50 years (Lobell and Burke, 2009; Carter et al., 2016) and we included both of them in the final model.

We selected three climate variables to include in the final model. Starting
with ten climate variables (Table 2.1), we systematically dropped one variable at a time and assessed the impact on the model fit. For significantly correlated climate variables, we used partial dependence and feature contribution (FC) plots, to estimate the average effect of each variable on crop yields and how each interacts or correlates with other variables. FC plots are in the same family as partial dependence plots, but allow visual identification of latent interactions or correlations between predictor variables with a color gradient and provide a measure on how well the plotted line matches raw data (Welling et al., 2016). If one variable in a correlated cluster or pair interacted less with other climate variables, we selected that variable (e.g. $ETo$ was more often correlated with other climate variables than $vpd$). We selected maximum temperature over mean or minimum temperature to capture extreme heat conditions during daytime. The final climate variables in our model were maximum temperature, accumulated precipitation, and vapor pressure deficit (Table 2.2). Therefore, the final regression model for each crop was:

$$\hat{y}_{it} = F(t, i, tmx_{it}, pre_{it}, vpd_{it}) + \epsilon_{it} \quad (2.1)$$

where $\hat{y}_{it}$ represent crop yields for year $t (t = 1, ..., 53)$ and country index $i (i = 1, ..., 35)$. $F$ represents the Random Forest regression function, $i$ represents country fixed effects, and $\epsilon_{it}$ represents the error.

Additionally, because the bootstrap technique samples from the distribution of observed values which concentrate in the midrange, Random Forest often overestimates small yields and underestimates large yields. To minimize this effect, we sorted the data by yields and partitioned the data into five bins and randomly sampled $n$ times from each bin, where $n = (2/3) \times \text{size(smallest bin)}$. The remaining data was held out for testing, despite the fact that Random Forest is designed to be run without the need for a test set.

### 2.4 Results

#### 2.4.1 Overall performance

We built the final Random Forest model using the three climate variables ($tmx, pre, vpd$) and two variables representing technology ($time$) and regional/country effects
Table 2.3: Model comparison using R² (RMSE in kg/ha is reported in parentheses) of the baseline, linear, and Random Forest models in this study. The variables for the Random Forest model are the same as Equation 2.1. We also include the average weather model from Schlenker and Lobell (2010) for comparison with a similar study (where $h_{it}$, $p_{it}$, and $c_i$ represent the mean temperature, total precipitation, and country fixed effects, respectively). The data used in this study are not the same as in Schlenker and Lobell (2010), so these models cannot be compared directly.

<table>
<thead>
<tr>
<th>Model</th>
<th>Maize</th>
<th>Groundnut</th>
<th>Sorghum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Model - Average Weather (Schlenker and Lobell, 2010)</td>
<td>$\hat{y}<em>{it} = \gamma_1 t + \gamma_2 t^2 + c_i + \alpha_i h</em>{it} + \beta_1 p_{it} + \epsilon_{it}$</td>
<td>$0.59$</td>
<td>$0.39$</td>
</tr>
<tr>
<td>Linear Model - Baseline</td>
<td>$\hat{y}<em>{it} = \gamma_1 t + c_i + \epsilon</em>{it}$</td>
<td>$0.53 (411)$</td>
<td>$0.43 (236)$</td>
</tr>
<tr>
<td>Linear Model</td>
<td>$\hat{y}<em>{it} = \gamma_1 t + c_i + \alpha_i t m x</em>{it} + \beta_1 p_{it} + \varphi_1 v p d_{it} + \epsilon_{it}$</td>
<td>$0.55 (403)$</td>
<td>$0.45 (231)$</td>
</tr>
<tr>
<td>Random Forest Model</td>
<td>$\hat{y}<em>{it} = F(t, i, t m x</em>{it}, p r e_{it}, v p d_{it}) + \epsilon_{it}$</td>
<td>$0.73 (310)$</td>
<td>$0.61 (193)$</td>
</tr>
</tbody>
</table>

(country) as described above and summarized in Equation 2.1. We found that the Random Forest model predicted out-of-sample yields better than a linear model when we compared the non-parametric approach with a linear model containing the same predictors, and a baseline model that only included a linear time trend and country fixed effects (Table 2.3). The Random Forest captured the majority of yield variability through the time and country variables, as in Schlenker and Lobell (2010) with a linear model. In general, when compared with linear regression, using the Random Forest analysis improved the modeling of historical yields for all three crops based on increases in the R². While Random Forest does not require withholding data (i.e. internal error monitoring, OOB RMSE), we chose to withhold a random selection of one third of the data to estimate the root mean squared error (RMSE) for both models on an equal basis (Figure 2.3). Comparing the OOB RMSE for the training set and RMSE for the test set (black and red, respectively), we found that the training and test set RMSE were similar, indicating that we did not overfit the model (Figure 2.3). We also noted that the Random Forest model over (under) predicted low (high) values (see the clockwise rotation with respect to the 1-to-1 line in Figure 2.3). We attempted to minimize the tilt by partitioning the data into bins based on yield and using the same number of observations within
each bin, but this did not completely eliminate the biases.

### 2.4.2 Dependence on Individual Variables

#### 2.4.2.1 Time/Technology

From 1962 to 2014, maize, groundnut, and sorghum yields displayed positive, first-order trends due to improving technology (Figure 2.2a). For all crops, we found that yields depend linearly on *time*, with a weaker linear trend in groundnut (Figure 2.2b). Detrending crop yields prior to the analysis would have removed trends of 15, 3.7, and 3.8 kg/ha of grain per year for maize, groundnut and sorghum, respectively. However, as estimated by the partial dependence plots, trends for *time* were 13, 2.8, and 3.2 kg/ha per year for maize, groundnut and sorghum, respectively. This indicates that the model identified the long-term trend without imposing it and eliminated the need for a priori detrending to estimate the climate effect on yields.

#### 2.4.2.2 Maximum temperature

Maximum temperature (*tmx*) was the most important climate predictor and in general, we found that yields decreased with increasing temperatures, albeit with some variations when maximum temperatures exceeded 30°C (Figure 2.4a). Yields consistently decreased for all three crops between 23°C and 30°C.

In general, maize and sorghum responded similarly to temperature because both are annual C4 crops and are grown in similar regions (Figure S2.1). Maize and sorghum were occasionally grown in regions with cooler temperatures than groundnut, but these instances (cooler than 22°C) were limited (Figure 2.4a).

Each of the three crops behaved uniquely for temperatures above 30°C. Only Lesotho and South Africa experienced growing seasons with average maximum temperatures in this range (Lesotho was consistently cooler), and the results were impacted accordingly (Figure 2.5a).

Maize yields showed small increases when temperatures were above 30°C, but the partial dependence deviations indicated that maize yields remained below average and decreased again above 33°C (Figure 2.4a). Groundnut yield response to temperatures above 30°C had a similar but sharper pattern than maize, increasing between 30°-33°C and decreasing at higher temperatures. Sorghum yields remained steady from 30° to 33°C, and decreased again at higher temperatures. In general, the
model captured the negative response of yields to increasing maximum temperature.

2.4.2.3 Precipitation

Modeled crop yields increased rapidly with precipitation ($pre$) until a threshold after which yields became insensitive to further precipitation increases (Figure 2.4b). Partial dependence plots indicated precipitation thresholds for maize, groundnut, and sorghum occurred between 300-500 mm. The bump in sorghum yields at 250 mm was caused by data from South Africa only, as this feature disappeared from the plot when we omitted its data. As expected, the minimum in maize and sorghum yields occurred in conditions with high temperatures and low precipitation (Figure 2.6a,b).

2.4.2.4 Vapor pressure deficit

For all crops, we estimated that yield decreased with increasing vapor pressure deficit ($vpd$) (Figure 2.4c). Although we designed the methodology to avoid having pairs of highly correlated variables, a color gradient applied along the temperature axis highlights the correlation (by definition) of $tmx$ and $vpd$ (Figure 2.5a,c). However, we estimated that the yield response to $vpd$ was independent of $tmx$, indicating that a drier atmosphere can additively penalize crop yield and amplify grain losses at higher temperatures. While the impact of vapor pressure deficit on crop yields was less important than maximum temperature and accumulated precipitation, it was distinctly negative, as expected.

2.4.2.5 Country index

Country fixed effects (country) were included to account for technological heterogeneity by functioning as a scale-factor for the regression. Each country implements adaptations and new management techniques at its own speed, so allowing for varying degrees of technological innovation was particularly important given the low quantity and quality of data in SSA (i.e. soil, fertilizer usage, tillage practices). The model did not permit in-filling when it encountered missing yields, therefore countries with more missing data contain fewer instances and less certain country indices. There appeared to be a correlation between country index and maximum
temperature because we grouped the countries by regions which experience similar climates (Figure 2.5a,e). For example, countries in the Sahel are warmer and listed first (Table S2.1), therefore the first six countries appear correlated with high maximum temperatures.
Figure 2.3: Modeled yield is plotted against observed yield for (a,c,e) the simple linear model (identical to the model included in Row 3 of Table 2.3) and (b,d,f) the Random Forest model (identical to the model included in Row 4 of Table 2.3) for (a,b) maize, (c,d) groundnut, and (e,f) sorghum. The data used to train the model are designated by black diamonds, while the test data is designated by red circles. The $R^2$ and mean squared errors for the training and test sets are included in the top left corner of panel in blue, black, and red, respectively.
Figure 2.4: Partial dependence plots for (a) maximum temperature (°C), (b) precipitation (mm), and (c) vapor pressure deficit (hPa). The y-axis represents the average deviation in yields caused by a given variable (kg/ha). We plot data between the 2.5-97.5<sup>th</sup> percentiles of the climate data to focus on the more robust signals, though several small, exaggerated features remain.
Figure 2.5: Feature contribution (FC) plot for groundnut. Panel titles designate which variable is being plot along the x-axis: (a) maximum temperature, (b) accumulated precipitation, (c) vapor pressure deficit, (d) time, and (e) country index. Panel titles also include the $R^2$ (leave-one-out goodness of fit) of the average FC line (denoted in orange). The color gradient is applied in all panels along the maximum temperature axis changing from green-blue-red with increasing temperature. The country indices are: Mali (1), Niger (2), Chad (3), Sudan (4), Burkina Faso (5), Senegal (6), The Gambia (7), Guinea Bissau (8), Guinea (9), Sierra Leone (10), Cote D’Ivore (11), Ghana (12), Togo (13), Benin (14), Nigeria (15), Somalia (16), Kenya (17), Tanzania (18), Uganda (19), Ethiopia (20), Rwanda (21), Burundi (22), Zambia (23), Malawi (24), Mozambique (25), Zimbabwe (26), Botswana (27), Namibia (28), South Africa (29), Lesotho (30), Swaziland (31). Groundnut yield data for Lesotho is missing, therefore the country index for Swaziland is 30 for this crop only (Table S2.1).
Figure 2.6: Dual (or 3D) partial dependence plot for (a) maize and (b) sorghum yield, with maximum temperature ($tmx$) and precipitation ($pre$) as the predictor variables. Maximum temperature ($tmx$) and precipitation ($pre$) are plotted in the horizontal, xy-plane, and the average modeled yield on the vertical, z-axis. Color gradient aligns with average modeled yield.
2.5 Discussion

We used Random Forest as a diagnostic tool to study historical crop responses to past climate in SSA and to understand the drivers of crop productivity in this region. We found that crop yields were considerably affected by improving technology but also exhibited distinct responses to maximum temperature, accumulated precipitation, and vapor pressure deficit.

The most important driver of crop yields from 1962-2014 was time, which we assumed was representative of improving technology. These technologies include better agronomic management, increased inputs to realize the yield potential in areas of high precipitation, and the use of adapted genotypes (Lassaletta et al., 2014). We acknowledge that the model variable included other factors that can improve yield, for example merging technology and CO$_2$ positive effects on crop yield. While it can be argued that CO$_2$ rather than technology can account for the gradual increase in yield with time, the differences among countries suggest that technology was more influential. Countries with low average yield suggest limiting factors that may decrease the ability of crops to respond to CO$_2$. For instance, the CO$_2$ effect is expected to be stronger in C3 crops and in drier conditions (Sultan et al., 2014), although Gray et al. (2016) reported lower benefits of CO$_2$ fertilization in dry conditions in locations of high productivity. McGrath and Lobell (2011) reported that 10% of the trend in yields between 1960 and 2009 could be attributed to CO$_2$ while technology explained the remainder. Thus, while CO$_2$ fertilization is an additional component, we are confident the dominant signal identified by time is representative of technological advances.

The impact of technology was particularly evident in maize yields, which increased faster in the last decade (Figure 2.2). This trend is robust across countries and not caused by the highest-yielding countries. We estimated that technology increased average maize yields by roughly 13 kg/ha per year (just over 1% per year), albeit from a low baseline. Though the magnitude of yield is not comparable between SSA and the United States, the relative rate of 1% per year is similar to that observed in the US after 1960 (1.2% growth based on NASS data). Statistical crop modeling projections often assume the technology trend that persisted over the last half century will continue throughout the modeled time period, which is most likely true (or an underestimation) in these regions because the base yield is low.
It will be important to continue implementing technologies that increase yield and provide tolerance to increasing abiotic stress as a result of increasing temperatures. We note that biotic stress is a source of unexplained yield variability, which we assume is unrelated to climate. However, there is uncertainty in this assumption because crop pests and disease risks may have increased due to a changing climate (Morton, 2007).

Country fixed effects were critical to modeling historical crop yields in SSA because they account for heterogeneities which result in large yield variations among SSA countries. We assumed that the country index is an appropriate scale for the data, but uniform averaging over small and large countries may preferentially introduce more uncertainty in large countries. The country index encompassed a wide set of indicators and did not express intra-country climate or soil type variability, but it did account for variation in soil fertility and other time-invariant differences between countries. Although we averaged the climate for each country, we only considered the areas where crops were grown, which minimized the intra-country climate variance. These results indicated that time and country were both important to predicting yields accurately.

Because technological advances were largely responsible for driving crop yields over the 53-year period, climate variables played a comparatively small role (Table 2.2). Climate may exhibit a relatively low impact on yields because low-yielding or continuously stressed crops tend to show minimized sensitivity to variations in weather. For instance, when sorghum is exposed to high temperature stress during flowering, a decrease in harvest index occurs, but if the same plant is continuously exposed to high temperatures (i.e. grown in above-optimal conditions), responses to additional heat stress are minimal because the productivity is already low (Prasad et al., 2006, 2008).

The most important climate predictor was maximum temperature. Maize, groundnut, and sorghum yields responded negatively to higher maximum temperatures (Figure 2.4a). Between 23°C and 30°C, all yields responded linearly, and as temperatures increase to 33°C, sorghum yields stagnated, maize yields increased slightly, and groundnut yields increased in a pattern that is difficult to explain. For maximum temperatures above 33°C, all yields dropped off again as temperature increased.

Existing research has indicated that yield responses to increasing temperatures
may be nonlinear or exhibit threshold behaviors that reflect the underlying response of physiological processes (Kemanian et al., 2004; Porter and Semenov, 2005; Parent and Tardieu, 2012), but we did not observe this response in our modeled yields. Nonlinear modeling of temperature effects has provided superior out-of-sample forecasts from statistical models, but these studies have only been performed with predefined response functions and fine, high quality yield and weather data (Schlenker and Roberts, 2009; Lobell et al., 2011a, 2013; Rosenzweig et al., 2014). One exception is Schlenker and Lobell (2010) which utilized monthly CRU TS 2.1 data and approximated the distribution of daily temperatures via downscaling. While their results indicated that the nonlinear models that included these data performed well, we chose not to introduce additional uncertainty into the model with data interpolation. There are several possible explanations for the lack of non-linear or threshold responses in the modeled crop yields. The most probable is that averaging over a fixed time scale and over different geographic regions obscured the nonlinear nature of the signal by removing the extreme values. Similar considerations apply to country averages, particularly for countries with large variations in growing conditions. It is also possible that we limited the potential for identifying nonlinear yield responses to maximum temperatures because we averaged over diverse areas and soil conditions.

The results for maize agree with ensemble model results in Bassu et al. (2014) for locations where the baseline average temperatures are above 25°C (average daily maximum temperatures were typically 5°C warmer than the average daily temperature). The increase in yields as temperatures decreased from 30°C indicates that maize is typically grown above optimal temperatures in these warm regions. While a growing body of evidence suggests that maize yields exhibit strong negative responses to temperatures exceeding 30°C in rainfed or water-limited conditions (Schlenker and Roberts, 2009; Lobell et al., 2011a, 2013), we did not detect this signal. Maize yields tended to increase slightly between 30°C to 33°C but decreased again for temperatures above 33°C, with an overall negative response of yield to temperature (Figure 2.4a). Reductions in modeled maize yields also appear to be more acute when precipitation was limited (Figure 2.6a) (Anderson et al., 2015).

The increase in groundnut yields from 30°C to 33°C requires further comment. While the density of data that occurred in this temperature band indicates that it
is a robust response (Figure 2.5a), this result is contrary to basic crop physiology. The $R^2$ value atop each panel describes how well the feature contribution can be understood as a main effect. A main effect is one that can be understood isolated from other effects, while an interaction effect cannot be understood on its own (S. Welling, personal communication). The low $R^2$ for the maximum temperature feature contribution plot ($R^2 = 0.24$) indicates that the interactions with other predictors are needed to explain the model response to maximum temperature. Indeed, when joint, or three-dimensional, feature contributions for maximum temperature and precipitation ($tmx$-$pre$) and maximum temperature and vapor pressure deficit ($tmx$-$vpd$) were computed, the explanatory power (i.e., $R^2$) of the surface increased to 0.49 and 0.44, respectively. In addition, more uncertainty exists for groundnut’s response curves than those of maize and sorghum because the model captured less variability (Table 2.3), and part of the uncertainty might be intrinsic to the quality of the FAO database for this crop.

As maximum temperatures increased, modeled sorghum yields consistently decreased across all regions (Figure 2.4a). The rapid increase in yields at $23^\circ$C should not be considered robust, since it is caused by only one country, Lesotho.

A distinct yield response to precipitation emerged that agrees relatively well with agronomic expectations: yields in SSA rapidly increased in a linear fashion while water-limited, but then plateaued after the 350-500 mm threshold (Figure 2.4b). The shapes of the yield responses agree well with the idealized maize response to precipitation in Figure 2.2 of Anderson et al. (2015). The thresholds observed in this experiment were consistent, albeit slightly below, estimates of 500-800 mm for optimal growing conditions (Kremer et al., 2008; Sadras et al., 2011). We did not explicitly consider stored soil water, which contributes to available crop water, and could explain why yield response to precipitation leveled off at a low threshold. Alternatively, water use efficiency decreases at low nutrient levels (Bennett et al., 1989), so low precipitation threshold estimates could be a result of limited soil nutrients like phosphorous or nitrogen. This work identified precipitation as the second most important climate variable, but the magnitude of yield responses was likely moderated as a result of aggregating across many soil types with varying responses to rainfall. Despite the caveats of using real data, the Random Forest model clearly indicated a nonlinear yield response to precipitation consistent with agronomic literature and without explicit pre-specification.
As expected, our results also show that vapor pressure deficit compound the limitations imposed by heat stress and had a negative effect on yields (Figure 2.4c). For all crops, the modeled contribution of vapor pressure deficit was better explained when it included interactions with the other climate variables, indicating latent interactions between variables.

Physiologically, vapor pressure deficit can be used to standardize the response to water availability. All other conditions equal, the production of biomass decreases as vapor pressure deficit increases when water stress limits growth, particularly above 10 hPa (Tanner, 1981; Tanner and Sinclair, 1983; Kemanian et al., 2005). When linear trend lines were plotted through the partial dependence values for \( vpd \), the average trend line intersected zero near 11.2 hPa, indicating that the model likely captured the detrimental effects of high vapor pressure deficit. Large maize yield losses may be associated with high evaporative losses in high vapor pressure deficit and high temperature conditions (Roberts et al., 2012; Lobell et al., 2013, 2015). However, in a study that controlled for genetics and management in irrigated maize, Carter et al. (2016) found no negative impact of increasing vapor pressure deficit on seasonal time scales. Our results have indicated a weak, but distinctly negative, impact of increasing vapor pressure deficit on maize yields, likely because the majority of crop areas in SSA are rainfed.

Our results indicate distinct yield responses to climate variables, but we have not discussed how the growing season climate has changed in SSA. Over the time series, every country included in the analysis experienced a positive trend in maximum temperatures, warming by an average of 1.2°C (Figures S20-22). In 95% of country-crop combinations, the warming trend exceeded one standard deviation of historic year-to-year variability. Meaning, the average conditions and locations in which maize, groundnut, and sorghum grew have become progressively warmer for each country between 1962 and 2014. Precipitation exhibits more spatial variability than temperature, so the trends in growing conditions for each country are more varied, and often significantly smaller, than the historic year-to-year variability. So, while the trend in precipitation was negative across 72% of the country-crop combinations, the respective magnitudes for each country were insignificant. If either trend continues for temperature or precipitation, the impact could diminish the positive trend in crop technology over time. Under a conservative emission scenario (RCP4.5), the majority of average temperatures in SSA in 2030 are
expected to increase between 0.75°C and 1.0°C (Kirtman et al., 2013). While these results present a modest but significant shift in the mean, evidence suggests that maximum daily temperatures will increase more rapidly than the mean (Kirtman et al., 2013). Again, because precipitation has higher spatial heterogeneity than temperature, projected precipitation trends under RCP 4.5 are more uncertain than temperature (Kirtman et al., 2013). In 2030, areas which are wet are expected to become wetter, and drier areas are expected to become drier (Kirtman et al., 2013). If we assume the trend that existed between 1962 and 2014 persist, it is likely that much of the agricultural areas in SSA will become drier. Increasing production in areas with excess water supply can be a promising avenue to increase the average production in SSA significantly.

In summary, while technology will likely keep improving crop yields in SSA, the potential gains will be lowered by increases in maximum temperatures and vapor pressure deficits because the regional thermal regime implies that crops are grown within the range of temperatures that decrease yield. In addition, increasing frequency of heat stress events may shift part of SSA into increasingly marginal conditions for crop production. Conversely, increases in precipitation within the range of response to precipitation (Figure 2.4b) or irrigation (when feasible) can boost yields considerably.

We recognize that statistical crop models can be constructed in numerous ways, and while variable selection in this work was constrained by available data and selected in a systematic fashion, the choice of variables is still subjective. Recent work has shown that other statistical crop modeling teams have used average monthly temperature and soil moisture data over the growing season (Iizumi et al., 2013), while others have used average monthly temperature and total precipitation over the growing season (Lobell et al., 2008; Schlenker and Lobell, 2010). In lieu of using average monthly temperature, when daily data was available or interpolated from monthly data, growing degree days and extreme degree days were computed and used (Schlenker and Roberts, 2009; Schlenker and Lobell, 2010; Lobell et al., 2011a; Roberts et al., 2012). It is less common to use vapor pressure deficit as a predictor, but it has been done with daily data (Roberts et al., 2012; Urban et al., 2015). However, our selection of maximum temperature, total precipitation, and vapor pressure deficit as predictors fits within the community attempting to address the complex dynamics of crop yields at regional scales.
Model performance and accuracy was potentially decreased because we omit radiation data. In the US Corn Belt, radiation has been shown recently to be the primary limiting factor on irrigated maize yields (Carter et al., 2016), while the contribution of solar brightening could be responsible for up to 27% of the increase in maize yields from 1984-2013 (Tollenaar et al., 2017). We omitted radiation from the analysis because reliable observations did not extend over the time series.

We also recognize that the use of spatially and temporally averaged (or aggregated) climate data can result in loss of potentially important heterogeneities and introduce biases. It is well-documented that the use of nonlinear functions with linearly averaged input data at larger scales may lead to biased predictions (Pierce and Running, 1995; Ewert et al., 2015; Hoffmann et al., 2016). However, the impact of spatial aggregation on statistical crop models, particularly a nonparametric method like Random Forest, has not yet been investigated.

The area considered in this analysis spans a variety of soil, climate, and socioeconomic conditions, and we assume that this complexity is captured with the time and country variables. However, sub-Saharan Africa is dominated by smallholder agricultural systems which are inherently complex and the impacts of human health as well as biotic and abiotic crop stressors have not been well studied in these systems (Morton, 2007; Roxburgh and Rodriguez, 2016). While our results indicate distinct yield responses to climate, we acknowledge that the realized outcomes are likely more complex than described in this manuscript.

In this research, we highlight the utility of Random Forest analysis as a critical tool for analyzing factors impacting agricultural yields in SSA and identified dependencies on both climate and other driving factors. Although this analysis covers a large area with a great variety of soil types, climates, and socioeconomic conditions with coarse country-level data, the impact of climate and technology on crop yields was detected. The Random Forest model predicted more accurate yields when compared to a linear model containing the same variables. The Random Forest model not only outperformed previous approaches in which functional crop responses are predefined, but it eliminated the need for strong prior assumptions. A major potential use for Random Forest in this field is to identify the importance of less-studied climate variables and how they affect crop yields or interact with the variables that have been studied. This machine learning tool should help accelerate progress in our understanding of crop models and possibly reduce uncertainty in
predictions of the climate impacts on crop productivity at different spatial scales.

2.6 Acknowledgments

We gratefully acknowledge funding from the National Science Foundation through the Network for Sustainable Climate Risk Management (SCRiM) under NSF cooperative agreement GEO-1240507 and the USDA NIFA Award 2014-68002-21768. We would like to thank S. Welling and D. Lobell for help and discussions as well as the anonymous reviewers for helpful insights and suggestions. We declare no conflicts of interest for this research.
2.7 Supplement

Section 2.7.1 describes the methodology used to create the climate data used in our models. Section 2.7.3 briefly comments on the quality of data and how certain issues may manifest in the model results described in the manuscript. While Random Forest minimizes prior assumptions compared to parametric models, \textit{a priori} decisions regarding model structure are still necessary. Section 2.7.4 provides more information on Random Forest as well as the choices made to create the final model. Section 2.7.5 presents some results which were not presented in the manuscript. Finally, Section 2.7.6, provides additional discussion not presented in the manuscript.

2.7.1 Calculation of Climate Data

In the following section and equations, allow latitude and longitude indices for the 5 minute and half degree grid to be \([i, j]\) and \([k, l]\), respectively. We created the climate data using the NCAR Command Language (NCL), and all referenced functions are available online.

2.7.2 Crop area fraction and irrigation

The crop area fraction datasets are 5’ x 5’ (Figure S2.1). To scale the masked crop area fraction dataset (5’ x 5’) to the climate grid (0.5° x 0.5°), we systematically looped through each of the climate grid cells and averaged the 36 crop area fraction cells inside to compute the crop area fraction on the climate grid \((CA[c, k, l])\).

Reported yields likely include yields from irrigated regions, but we did not mask out irrigated land in the analysis. We assumed this does not significantly contaminate the large-scale signals because only a small fraction of area considered in this study is irrigated. Using the Global Map of Irrigated Areas (GMIA) from AQUASTAT we computed the actual irrigated area as a fraction of the total area of each 5’ x 5’ (Siebert et al., 2013) grid cell. We then estimated the upper limit of the percent of irrigated land on maize, groundnut, and sorghum. For the entire area included in this study, we estimated that only about 3.2% is actually irrigated.

Crop area data indicated that Djibouti did not grow any groundnut or sorghum, so there are no climate data for that country.
Figure S2.1: Crop area map for (a) maize, (b) groundnut, and (c) sorghum based on data from (Monfreda et al., 2008). The countries included in the analysis are outlined in black. The scale is logarithmic and ranges from white to red as crop area fraction increases.

(a) **Maize − Crop Area**  
(b) **Groundnut − Crop Area**  
(c) **Sorghum − Crop Area**

2.7.2.1 Crop calendar

Maize and sorghum have secondary seasons (Sacks et al., 2010). Since the two growing seasons cannot be discerned from the final yield reports, we used the following method. We chose the primary season for maize and sorghum because the spatial distribution of grid cells from (Sacks et al., 2010) indicates that the secondary maize and sorghum seasons are less prevalent in SSA. We made the assumption that crop yield is directly proportional to crop area, and relatively small areas of second-season crops did not significantly contribute to the total yield. As stated in the text, we assume growing seasons of 120 days based on the average harvest day from the Sacks et al. (2010) data (Figure S2.2). The crop calendar data was used in the following manner.

We let the crop calendar vector for each crop \( c \) and grid cell \( (i, j) \) be \( CC[c, i, j, m] \) where \( m \) spanned from 0 to 23 (24-month dimension). This array was initialized to 0. Allow \( DOY \) be the cumulative days in a normal year on the first day of each month \( (DOY = 0, 31, 59, ..., 304, 334) \). If the average planting date lied in a given month, the number of days remaining in that month (including the planting day) was used to compute the fraction of that month in which a crop is actively growing. For the days in the harvest month preceding the harvest day, the number of days (including the harvest day) were used to compute the fraction...
of that month in which a crop is actively growing. If an entire month was used to
grow a crop, the value for that month equals 1. Let’s consider an example: for a
given crop and grid cell, allow the harvest date \( h_{date} \) to equal 215 (215\textsuperscript{th} day of
the year; early August). The planting date \( p_{date} \) is 95 because \( p_{date} = h_{date} - 120 \).
Values \( p_{frac} \) and \( h_{frac} \) represent the bookends of the growing season in fractions.
These values are computed from \( p_{date} \) and \( h_{date} \) so that we can compute the values
for each month in \( CC[c, i, j, m] \). \( p_{frac} \) was computed in the following manner,
\[
p_{frac} = \frac{p_{date} - \text{DOY}[n]}{\text{DOY}[n + 1] - \text{DOY}[n]} + n
\]  
(2.2)
where \( n \) was calculated by a search function. In this example \( n = 3 \) and the value
of \( p_{frac} = 3.167 \). An equivalent calculation (except that \( p_{date} \) was replaced by
\( h_{date} \)) was performed to calculate \( h_{frac} \). In this example \( n = 7 \) and \( h_{frac} = 7.096 \).
Therefore the crop is grown in a fraction of April, all May, June, July, and a fraction
of August. Note that functions \( ceil \) and \( floor \) round up and down to the nearest
integer, respectively. In this example, \( CC[c, i, j, m] \) would be filled in as follows

Now,
\[
CC[c, i, j, ceil(p_{frac}) : floor(h_{frac}) - 1] = 1.00
\]  
(2.3)
and
\[
CC[c, i, j, floor(p_{frac})]_{i,j} = ceil(p_{frac}) - p_{frac} = 0.833
\]  
(2.4)
and
\[
CC[c, i, j, floor(h_{frac})]_{i,j} = h_{frac} - floor(h_{frac}) = 0.096
\]  
(2.5)

Next, \( CC[c, i, j, m] \) had to be interpolated to the climate grid \( CC[c, k, l, m] \).
As stated in the text, we used the NCL function \( hi2lores \), which uses local area
averaging. We note that the code is open source and available online.

2.7.2.2 Climate Data

Let the original array of climate data equal \( CL[y, n, v, k, l] \), where the indices \( [y, n, v] \)
represent year (53), month (12), and variable (9), respectively. Recall that \( [k, l] \)
represent latitude and longitude on the climate grid. Next, we altered the climate
array so that we could consider 24 months for each year. The new array is now
\( CL[y, m, v, k, l] \) and the year 1962 now contains climate data for all months in 1961
Figure S2.2: Average harvest day for (a) maize, (b) groundnut, and (c) sorghum based on data from (Sacks et al., 2010). The countries included in the analysis are outlined in black. The scale is logarithmic and ranges from white to red as crop area fraction increases.

and 1962. Next, we computed a dummy array, $D$, by multiplying the the climate data by the crop calendar, $D[c, y, m, v, k, l] = C\cdot L[y, m, v, k, l] \cdot C\cdot C[c, k, l, m]$. For each crop ($c$) and year ($y$), we summed the array along the month dimension and then, for all variables ($v$) except $pre$, $ETo$, and $aridity$, this value was averaged in the following manner to compute $CGS[c, y, v, k, l]$,

$$CGS[c, y, v, k, l] = \frac{\sum_{m=0}^{23} D[c, y, m, v, k, l]}{\sum_{m=0}^{23} C\cdot C[c, k, l, m]} \quad (2.6)$$

For each country, $ct$, we used Global Administrative Areas shapefiles to create a data mask on the climate grid, ($S[ct, k, l]$), where grid cells within a country are equal to one and all others are missing. The growing season averaged climate, $CGS[c, y, v, k, l]$, was finally weighted and normalized (for each country) by the crop area fraction in the following manner to produce the final gridded climate data, $C[ct, y, c, v]$.

$$C[ct, y, c, v] = \frac{\sum_k \sum_t C\cdot G S[c, y, v, k, l] \cdot C\cdot A[c, k, l] \cdot S[ct, k, l]}{\sum_k \sum_t C\cdot A[c, k, l] \cdot S[ct, k, l]} \quad (2.7)$$
2.7.3 Data Overview

Many African countries are data-deficient, but there has yet to be an in-depth analysis of the available data and how the results compare to existing research with high resolution data. We provide the histograms of data used to build the regression models, and many climate variables do not have normal distributions (Figures S2.3, S2.4, and S2.5). Based on these distributions, and to allow climate predictors to interact in producing the yield response, we chose to use a non-parametric regression model. We also note that the FAO yield data can also be inaccurate.

When dealing with climate data in a regression model, it is very common to run into issues of correlated predictors. We provide a correlation matrix below to illustrate how all candidate predictor variables covary with one another (Figure S2.6).

2.7.4 Statistical Methods

In this section, we provide more information on variable importance, as well as details on the specific methods and algorithms used in the paper. This research uses the R language for all statistical analyses (R Core Team, 2016).

2.7.4.1 Random Forest

The CART algorithm (Breiman et al., 1983) chooses the split at each node so that reduction in overall node impurity is maximized (impurity is calculated as the sum of squared deviations from node centers). Unless otherwise specified, CART trees grow very large and are subsequently pruned. Pruning effectively "cuts off branches" that do not add predictive performance (according to a pruning algorithm that is not necessarily the same as the splitting criterion).

Random Forest (RF) uses a combination of trees in which each tree depends on the values of random, independently-sampled vectors with the same distribution (Breiman, 2001). The governing logic is laid out as follows: most trees will be good for most of the data, and most trees will make errors in different locations, therefore the ensemble average should be a good predictor of the data. RF models are built on the concept of bootstrap aggregation, or bagging, wherein each tree is constructed using a bootstrap sample of the data set resulting in independent
trees (Breiman, 2001). Random Forest combines the method of bagging with that of random split selection, where the split at each node of each tree is selected at random from all the predictors. At each split $m_{try}$ predictors are chosen. In this work, we set $m_{try} = 3$. If $m_{try} = 1$ the splitting variable is determined completely at random (if $m_{try} = p$ the random split selection component is eliminated and the method is effectively bagging). A Random Forest is therefore random in two ways: each tree is based on a random subset of the observations, and each split within each tree is treated based on a random subset of predictors.

We produce a Random Forest model ($F$) with a set number of trees, $n_{tree}$, for the dataset ($S$) with $N$ observations of crop yields ($y_k$) and $p$ covariates denoted $x_{i,k}$ for $i = 1, 2, ..., p$ and $k = 1, 2, ..., N$. Given the training set $S$, Step 1 is to form bootstrap training sets, $\theta_j$ for each tree where $j = 1, 2, ..., n_{tree}$. Step 2 is to construct regressions $h(x, \theta_j)$ with those data for each tree. Step 3 is to compute the average of all the trees to form the trained predictor, $F$. The regression model, $F$, generates predictions, $\hat{y}_k = F(x_{1,k}, x_{2,k}, ..., x_{p,k})$.

### 2.7.4.1.1 Error Monitoring

In each training set, $\theta_j$, about one third of instances are left out and called the out-of-bag or OOB data. These OOB data are used to estimate the generalization error, which is calculated as the MSE of the tree grow with $\theta_j$ on the OOB data on the training set (Grömping, 2009). This technique eliminates the need to set aside a test set, and provided $n_{tree}$ is large enough, the OOB estimate of MSE is accurate (Breiman, 1996). For all observations, $l$, that are OOB ($l = 1, 2, ..., n_{OOB}$). The accuracy of a Random Forest prediction is estimated as follows:

$$\text{OOBMSE} = \frac{1}{n} \sum_{l=1}^{n_{OOB}} (y_l - \hat{y}_l^{OOB})^2$$

(2.8)

where $\hat{y}_l^{OOB}$ represents the average prediction for the $l^{th}$ observation from all trees for which the observation was OOB. Similar to linear regression, the variance is equal to $\sigma_y^2 = \sum_{l=1}^{n} (y_l - \bar{y})^2$ and as a result, $R^2$ for the OOB data can be obtained as $1 - \text{OOBMSE}/\sigma_y^2$.

We used the OOB error rate to determine the number of trees ($n_{tree}$) in the regression. In Figure S2.7, it is easy to see that the OOB MSE begins to flatten around 300 trees.
2.7.4.1.2 Variable Importance  To compute the values in Table S2.2 and Table 2.2, the randomForest function calculates the average increase in MSE that comes from permuting a variable. For tree $j$, the OOB MSE is calculated as the average of the squared deviations of OOB responses from their respective predictions,

$$
OOBMSE_j = \frac{1}{n_{OOB,j}} \sum_{l=1}^{n} (y_l - \hat{y}_{l,j})^2
$$

(2.9)

where the hat indicates predictions, $OOB_j = \{ l : \text{observation } l \text{ is OOB for tree } j \}$, and $n_{OOB}$ is the number of OOB observations in tree $j$. The idea is that if regressor $X_i$ does not have a predictive value for the response, it should not make a difference if the values for $X_i$ are randomly permuted in the OOB data before the predictions are generated. Therefore

$$
OOBMSE_j(X_i \text{ permuted}) = \frac{1}{n_{OOB,j}} \sum_{l=1}^{n} (y_l - \hat{y}_{l,j}(X_i \text{ permuted}))^2
$$

(2.10)

For each tree, the difference between equations 2.10 and 2.9 and the MSE reduction for regressor $X_i$ for the forest is the averaged over all $n_{tree}$ trees of these differences to compute the variable importance metric. Variable importance for all variables is provided in Table S2.2. Larger numbers indicate that a variable is more important to the regression. It is possible that $time$ and $country$ arise as most important because they are largely uncorrelated with the other variables and the importance metric is biased towards predictors with unique values (Strobl et al., 2007). However, we do not think this is the case because it is well-documented that technology increase is important. For more information on RF please see Breiman (2001) for theory or Liaw and Wiener (2002) for the specifics of the randomForest R package.

2.7.4.1.3 Variable Selection  We used RF as a diagnostic model to create the most parsimonious model that can be used with all three crops. We chose to use five predictors after analysis of cross validation (Figure S2.8), OOB MSE error plots
We initially considered the inclusion of cloud fraction (cld) in the final model because it has a relatively high variable importance and the ability to act as a radiation proxy. However, the cloud fraction response was consistently difficult to interpret because it was related to other predictors. Cloud fraction may have captured (1) incoming solar radiation reaching the crops (clouds decrease direct and increase diffuse radiation), (2) a secondary impact of vapor pressure deficit (as clouds reduce moisture stress and signal higher humidity and lower vapor pressure deficits), (3) a secondary impact of maximum temperatures (clouds suppress high temperatures), or (4) a secondary precipitation signal (clouds produce precipitation). Because we did not find evidence of orthogonal information from the cloud fraction variable (e.g. radiation), its inclusion in the model increased the model MSE for all three crops, and its exclusion did not affect the shape of the yield responses to the other three climate variables, we omitted cloud fraction from the analysis.

Additionally, we did not include any explicit variable interaction terms in our initial list of predictors. This was because Random Forest enables identification of variable interactions within the model (when a variable appears repeatedly in different tree branches it is clearly an indication of interactions). Variable interactions were then identified through the systematic use of partial dependence and feature contribution plots in both two and three dimensions. Joint feature contributions plots were particularly useful in this process, and we refer the reader to Section 4.1 of Welling et al. (2016) for a more in-depth explanation.

Our final model included maximum temperature, accumulated precipitation, vapor pressure deficit, time, and a country index (given in Table S2.1).

2.7.4.1.4 Partial Dependence & Feature Contribution Plots To compute the partial dependence of one variable, $x_i$, the following function (Pearson, 2016) is computed and plotted over a range of observed $x_i$ values,

$$
\phi(x_i) = \frac{1}{n} \sum_{k=1}^{N} F(x_{1,k}, \ldots, x_{i-1,k}, x_{i,k}, x_{i+1,k}, \ldots, x_{p,k})
$$

(2.11)

Recall that $i = 1, 2, \ldots, p$ and $k = 1, 2, \ldots, N$, so the function $\phi(x_i)$ plots how the value of $x_i$ influences $\hat{y}_k$ after the influence of the other variables is effectively
averaged out. Feature contribution plots are more complex, so we direct the reader to Welling et al. (2016) for further information.

### 2.7.5 Model Results and Comparison

With Figure S2.13, we were able to visualize the average impact for all three crops. Interestingly if we estimated a linear trend line through these yields responses, the average yield impact of maximum temperature flipped from positive to negative when temperatures increase above 30°C, which is consistent with thresholds in agronomic literature. The average yield impact of accumulated precipitation flipped from negative to positive around 400mm which is also a reasonable estimate in rainfed conditions. Lastly, the average yield impact of vapor pressure deficit flipped from positive to negative at 10hPa, which is also consistent with thresholds in agronomic literature.

Further investigation into why modeled groundnut yields increase above 30°C found that the increase in groundnut yields is caused by countries in the Sahel and West Africa, which are warmer than the other countries in this study (Figure S2.9). Countries in the Sahel and in West Africa produced larger yields than countries in eastern and southern Africa (839 kg/ha compared to 747 kg/ha). This regional impact was minimized for all crops by the inclusion of the country index, but remained most pronounced for groundnut.

To supplement model comparison in the manuscript (Figure 2.3), we also provide residual plots from the linear and Random Forest regression models. The color gradient on the residual plots indicates both models poorly-predicted extreme yields. We sampled the data from bins to balance the data density along the yield range, but it did not eliminate the biased behavior of the models. Weak heteroscedasticity exists in the residuals of the linear model; by transforming the dependent variable to the natural log of yields this effect was corrected. Model $R^2$ for both Random Forest and linear models did not change considerably when using the natural log of yields as the dependent variable.

Once the final model was constructed, we allowed $m_{try}$ to vary and found minimal sensitivity this parameter.
2.7.6 Discussion

2.7.6.1 Fertilizer:

Fertilizer is important to bridging yield gaps in regions with poor growing conditions, though the response of agricultural systems to increased nitrogen fertilization over the past 50 years is different for each country (Lassaletta et al., 2014). Consumption of both nitrogen and phosphorous fertilizers have been steadily increasing in SSA since the FAO began archiving records in 1961 (Figure S2.16). In our manuscript, we discuss the impact of technology on yields, and fertilizer is a component of technology. In this analysis we considered nitrogen fertilizer, since the impact on model behavior was slightly better with nitrogen.

Because fertilizer has generally increased with time and is a component of technology, it could be assumed that time (which primarily captures the response to technology) could be omitted in lieu of fertilizer. Swapping out time for nitrogen fertilizer not only decreased the R\(^2\) of the maize and groundnut considerably (R\(^2\) drop to 0.56 and 0.43 for maize and groundnut, the sorghum model is largely unaffected), but it altered the maximum temperature response of yields. So while fertilizer was correlated with time, fertilizer use does not explain all of the technology signal observed in our results.

However, fertilizer could supply some orthogonal information not provided by time. So, we ran the model from our manuscript with an added predictor of nitrogen fertilizer consumption by country (variables: time, fertilizer, country, tmx, pre, vpd). An unsurprising result of this experiment was that the impact of technology (time) became less important and explained less variability than before. We compared the original yield response to technology to that of the new model in Figure S2.17. The R\(^2\) values of the new model were 0.73, 0.54, and 0.8 for maize, groundnut, and sorghum, respectively. Adding fertilizer to the model decreased the amount of explainable variability in the groundnut model, but increased it for sorghum. Adding fertilizer to the model smoothed out the model response to precipitation for sorghum by eliminating the jump in yields at 250mm caused by South Africa, and did not change the model responses to temperature or vapor pressure deficit (Figure S2.18). South Africa is the largest consumer of fertilizer in the analysis, and the jump in precipitation from Figure 2.3 likely disappeared because fertilizer (by country) provided another scale for the regression. It is also possible that fertilizer
scales similarly to irrigated area, and we were allowing the model to pick up on an irrigation signal via the fertilizer variable.

One of the major reasons we did not use fertilizer in our final model was because it is not crop-specific. It is unlikely that maize, groundnut, and sorghum receive equal proportions of fertilizer, to assume otherwise would introduce more uncertainty into our model. Another reason is that we were already picking up the fertilizer signal with the time variable. Additionally, the inclusion of fertilizer consumption degraded the explanatory power of the groundnut model, which was already low.

South Africa was not omitted from this analysis, despite its relatively high levels of fertilization and irrigation because we wanted an accurate representation of yields from SSA. Because South Africa is one of the largest producers of all yields, to omit this country would be misleading. Additionally, the impact of each country is easily identified with model response plots. Random Forest is built to be transparent, and because we identify and discuss the impact of South Africa we did not see a valid reason to omit it.

2.7.6.2 Standardized Precipitation Index:

One of the candidate predictors in our model that we did not end up using was the Standardized Precipitation Index developed by McKee et al. (1993). We computed this index from the gridded CRU TS 3.24.01 data using the NCL function\texttt{dim\_spi\_n}. We computed the three-month average for each month (after the first two months), with a Pearson-III distribution (Guttman, 1999) as opposed to the traditional Gamma distribution (McKee et al., 1993). In our data, SPI ranged from -2.8 to 2.5, the metric is designed to be symmetrical and indicate severe drought when the index drops below -2. We included SPI as a more thorough method of estimating drought conditions. We compared the impact of the standardized precipitation index (SPI) in lieu of accumulated precipitation (Figure S2.19). We found that the model response to SPI is linear and positive, as expected because low values of SPI indicate drought and high values indicate more water than normal. Swapping SPI for precipitation did not significantly improve the model, indicating that these two variables were identifying similar signals related to water availability.
2.7.6.3 Climate trends:

Support for material in the main text is presented in Figures S2.20, S2.21, and S2.22.
Figure S2.3: Sample distributions for data used in regression analyses: (a) annual yields (b)-(k) climate variables during the growing season. The bin sizes for each variable are 25 kg/ha, 0.1°C, 0.1°C, 0.1°C, 10 mm, 0.1 hPa, 2.5 mm/month, 0.5%, 0.025 (unitless), 0.075 hPa, 0.025 (unitless), for yield, tmp, tmn, tmx, pre, vap, ETo, cld, SPI, vpd, and aridity, respectively.
Figure S2.4: Same as Figure S2.3 but for groundnut.
Figure S2.5: Same as Figure S2.3 but for sorghum.
Figure S2.6: Heat map visualization of the correlation matrix computed for all candidate predictors listed in Table 2.1.
Table S2.1: Country indices included in manuscript. Asterisks denote a special case: groundnut yields for Lesotho do not exist, so the index for Swaziland becomes 30 in that instance.

<table>
<thead>
<tr>
<th>Country</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mali</td>
<td>1</td>
</tr>
<tr>
<td>Niger</td>
<td>2</td>
</tr>
<tr>
<td>Chad</td>
<td>3</td>
</tr>
<tr>
<td>Sudan (former)</td>
<td>4</td>
</tr>
<tr>
<td>Burkina Faso</td>
<td>5</td>
</tr>
<tr>
<td>Senegal</td>
<td>6</td>
</tr>
<tr>
<td>The Gambia</td>
<td>7</td>
</tr>
<tr>
<td>Guinea Bissau</td>
<td>8</td>
</tr>
<tr>
<td>Guinea</td>
<td>9</td>
</tr>
<tr>
<td>Sierra Leone</td>
<td>10</td>
</tr>
<tr>
<td>Liberia</td>
<td>NA</td>
</tr>
<tr>
<td>Ivory Coast</td>
<td>11</td>
</tr>
<tr>
<td>Ghana</td>
<td>12</td>
</tr>
<tr>
<td>Togo</td>
<td>13</td>
</tr>
<tr>
<td>Benin</td>
<td>14</td>
</tr>
<tr>
<td>Nigeria</td>
<td>15</td>
</tr>
<tr>
<td>Eritrea</td>
<td>NA</td>
</tr>
<tr>
<td>Djibouti</td>
<td>NA</td>
</tr>
<tr>
<td>Somalia</td>
<td>16</td>
</tr>
<tr>
<td>Kenya</td>
<td>17</td>
</tr>
<tr>
<td>Tanzania</td>
<td>18</td>
</tr>
<tr>
<td>Uganda</td>
<td>19</td>
</tr>
<tr>
<td>Ethiopia</td>
<td>20</td>
</tr>
<tr>
<td>Rwanda</td>
<td>21</td>
</tr>
<tr>
<td>Burundi</td>
<td>22</td>
</tr>
<tr>
<td>Angola</td>
<td>NA</td>
</tr>
<tr>
<td>Zambia</td>
<td>23</td>
</tr>
<tr>
<td>Malawi</td>
<td>24</td>
</tr>
<tr>
<td>Mozambique</td>
<td>25</td>
</tr>
<tr>
<td>Zimbabwe</td>
<td>26</td>
</tr>
<tr>
<td>Botswana</td>
<td>27</td>
</tr>
<tr>
<td>Namibia</td>
<td>28</td>
</tr>
<tr>
<td>South Africa</td>
<td>29</td>
</tr>
<tr>
<td>Lesotho*</td>
<td>30</td>
</tr>
<tr>
<td>Swaziland*</td>
<td>31</td>
</tr>
</tbody>
</table>
Figure S2.7: Solid black line plots OOB error rate as a function of the number of trees for a model with all predictors. Dashed red line represents the same function for a model with 6 variables: time, country, tmx, pre, vpd, and cld. Dashed blue line represents the same function using the 5 variables in the final model: time, country, tmx, pre, and vpd. Dashed green line represents the same function using only four variables: time, country, tmx, and pre. Vertical grey line delineates $n_{tree} = 300$. 
Table S2.2: Variable importance metric for all variables initially considered in the analysis.

<table>
<thead>
<tr>
<th></th>
<th>Maize</th>
<th>Groundnut</th>
<th>Sorghum</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>tmp</em></td>
<td>15.4</td>
<td>15.5</td>
<td>15.7</td>
</tr>
<tr>
<td><em>tmn</em></td>
<td>14.7</td>
<td>16.8</td>
<td>12.9</td>
</tr>
<tr>
<td><em>tmx</em></td>
<td>14.5</td>
<td>16.0</td>
<td>12.9</td>
</tr>
<tr>
<td><em>pre</em></td>
<td>16.7</td>
<td>16.5</td>
<td>13.0</td>
</tr>
<tr>
<td><em>vap</em></td>
<td>17.9</td>
<td>17.2</td>
<td>14.5</td>
</tr>
<tr>
<td><em>ETo</em></td>
<td>13.2</td>
<td>19.3</td>
<td>12.4</td>
</tr>
<tr>
<td><em>cld</em></td>
<td>18.0</td>
<td>19.4</td>
<td>13.3</td>
</tr>
<tr>
<td><em>spi</em></td>
<td>7.3</td>
<td>7.4</td>
<td>7.5</td>
</tr>
<tr>
<td><em>vpd</em></td>
<td>13.6</td>
<td>17.7</td>
<td>13.5</td>
</tr>
<tr>
<td><em>aridity</em></td>
<td>12.8</td>
<td>14.8</td>
<td>13.0</td>
</tr>
<tr>
<td><em>time</em></td>
<td>42.7</td>
<td>34.5</td>
<td>24.7</td>
</tr>
<tr>
<td><em>country</em></td>
<td>33.2</td>
<td>38.1</td>
<td>23.5</td>
</tr>
</tbody>
</table>
Figure S2.8: Prediction performance of Random Forest model with systematic reduction of predictors for maize, groundnut, and sorghum, respectively, using k-fold cross validation (k=10). This Random Forest model uses $n_{tree}=300$ and $m_{try}=3$. 

(a) maize  (b) groundnut  (c) sorghum
Figure S2.9: Dependence of groundnut on country index for data used in the regression. Yields which occurred when maximum temperatures were greater than 30°C are denoted by red dots. The horizontal blue segments represent the mean groundnut yields for SAF/WAF and EAF/SAF.
Figure S2.10: Feature contribution plot for maize with the color gradient applied along the maximum temperature axis. A consistent y-axis is maintained across each panel, which limits the ability to distinguish features in the climate variable plots.
Figure S2.11: Feature contribution plot for sorghum with the color gradient applied along the country index axis. The color scale identifies each country.

(a) **tmx (R^2 = 0.83)**

(b) **pre (R^2 = 0.52)**

(c) **vpd (R^2 = 0.19)**

(d) **time (R^2 = 0.25)**

(e) **country (R^2 = 0.98)**
Figure S2.12: Raw partial dependence plot for maximum temperature (not deviations) for maize, groundnut, and sorghum, respectively. Similar to Figure 2.3, data within the innermost 95th percentile are marked by a green 'x'. The red line represents the linear trend line through all plotted values; the fit represents the $R^2$ for each trend line. The red number in the bottom left of each panel represents the percent change relative to the average yield caused by one degree of maximum temperature increase.
Figure S2.13: Composite standardized partial dependence plots for maize, groundnut, and sorghum. A composite model is based on combining the data for all three crops, so that panel data are specified on the crop-country-year level. The 5th and 95th percentile bounds are designated by vertical grey lines. If data is missing for a given interval (i.e. if data exists for maize, but not groundnut and sorghum at 23°C), the average is computed based on quantities which are present. The thick, red line represents a trend line based on the data between the 5th and 95th percentiles to highlight the most robust trends in the yield responses to climate.
Figure S2.14: Residuals from linear models (a,c,e) and residuals from Random Forest models (b,d,f). Residuals are plotted as a function of the modeled yields and a color gradient is applied along the observed yield axis (not plotted). This gradient indicates increasing yields as colors shift from blue-green-red.
Figure S2.15: Total fertilizer consumption from all countries in the analysis from 1961-2014. Units of consumption are 10,000 metric tonnes of nutrient. Black and red lines represent nitrogen and phosphorous fertilizer consumption. The dashed lines represent the total when South Africa is omitted.
Figure S2.16: Average nitrogen fertilizer consumption of each country between 1961-2014.
Figure S2.17: Average yield for maize, groundnut, and sorghum between 1962 and 2014 for all countries in the final regression model. (b) Partial dependence plot for time for the model used in the manuscript. (c) Partial dependence plot for time with nitrogen fertilizer included as a predictor. Yield response to time should be interpreted as response to technology. As in the manuscript, maize, groundnut, and sorghum are denoted by green, orange, and black lines, respectively.
Figure S2.18: Partial dependence plots for (a) maximum temperature (°C), (b) precipitation (mm), and (c) vapor pressure deficit (hPa) for the model which includes consumption of nitrogen fertilizer. The average deviation in yields caused by a given variable (kg/ha) are plotted against each variable. Data between the innermost 95\textsuperscript{th} percentiles are plotted to focus on the robust signals.

![Partial Dependence: Maximum Temperature](image1)

![Partial Dependence: Accumulated Precipitation](image2)

![Partial Dependence: Vapor Pressure Deficit](image3)

Figure S2.19: Partial dependence plots for (a) maximum temperature (°C), (b) standardized precipitation index (unitless), and (c) vapor pressure deficit (hPa) in a model where SPI is used in lieu of accumulated precipitation. Data between the innermost 95\textsuperscript{th} percentiles are plotted to focus on the robust signals.

![Partial Dependence: Maximum Temperature](image4)

![Partial Dependence: SPI](image5)

![Partial Dependence: Vapor Pressure Deficit](image6)
Figure S2.20: Time dependence of maximum temperature for maize. Red line denotes the best fit line through the data using a simple linear regression. In the bottom right of each panel, the first blue number represents the average change in maximum temperature from 1962-2014 computed with the slope of the trend line. The second blue number in parentheses represents one standard deviation of the year-to-year variability in maximum temperature.
Figure S2.21: Same figure as S2.20, but for groundnut.
Figure S2.22: Same figure as S2.20, but for sorghum.
Chapter 3  
Effects of historical weather on US yields since 1980

3.1 Introduction

The geographical distribution and yield of grain crops depends strongly on the regional climate, and therefore, a determinant factor in food security. Our understanding of how specific climate variables affect the yield of different crop species has been advanced recently by analysis using historical climate data and yield records (Schlenker and Roberts, 2009; Lobell et al., 2014; Hoffman et al., 2017). This type of analysis takes advantage of the non-controlled experiments conducted by producers as they grow crops in different locations and years for which records are aggregated at scales from fields to regions (e.g. counties), and from regions to countries. When paired with the climate during the growing season and by using appropriate statistical techniques, these datasets may provide robust responses between crop yields and climate.

Traditionally, relatively simple statistical crop models like linear regression have been used to study the impact of climate variable on crop yields. However, as data quality and resolution improve, the use of simple models has dropped off in lieu of methods that allow for and/or prescribe nonlinear crop responses to climate. Schlenker and Roberts (2009) provided a new way of analyzing the crops response to temperature by calculating the grain yield response of corn, cotton and soybean east of the 100th meridian, and used the number of days exposed to a given temperature range and other variables as predictors. Their analysis revealed that
corn and soybean yield have a broad optimum temperature range, that falls off abruptly when the average temperature increases above 29° and 30°C, respectively. Secondly, they found no evidence that technological adaptation to temperature exists because the threshold is not sensitive to time (an indicator of the progress of technology). Collectively, this analysis highlighted the power of combining climate and yield records, despite the potential contamination of the climate signal by other factors such as diseases, pests, and even different soils.

As an alternative to models that specify predictor responses and their interactions, machine learning models are ideally suited to handle large data sets, require minimal assumptions, and above all, let data guide the analysis. As a potential drawback, most machine learning techniques are not transparent, and when the signal of interest is implicitly represented in the model, a "black box" is not informative. But this is not an insurmountable challenge. Random forests (RF) is a unique, non-parametric machine learning tool that is more transparent than other machine learning algorithms (e.g. support vector machines, neural networks, etc.). In fact, some suggest it should be called a 'grey box' instead of a black box (Prasad et al., 2004).

Random forests originated in bioinformatics and is increasingly found in the fields of ecology (Cutler et al., 2007), agronomy (Saha et al., 2017), and has recently been applied as a diagnostic statistical crop model (Hoffman et al., 2017) (presented in Chapter 2). In that research, we were able to separate technology from climate signals, despite the major socioeconomic discrepancies in the data originated from countries with different technologies and spatial domains. The immediate implication is that if one were to use a data-rich region that covers a large climate domain as used in Schlenker and Roberts (2009), it would be possible to extract a more detailed response of crops to climate conditions.

We summarize prior work on the impact of fundamental climate variables on yield as follows. Precipitation effects on crop yield are related to the water balance: when soil water limits the ability of crop canopies to match the atmospheric demand for evaporation, stomata close reducing photosynthesis. Beyond minimum temperatures that damage the vegetative or reproductive structures of crops, higher temperatures tend to decrease yield as a result of several factors (Lobell et al., 2014; Bassu et al., 2014; Liu et al., 2016). First, high temperatures enable a faster developmental rate (e.g. Sofield et al. (1977)) and therefore a shorter growing
season usually associated with lower interception of solar radiation. Second, high temperatures cause direct heat stress effects on grain sink size which is associated with reduced endosperm cell counts and floral structure size (Brocklehurst, 1977; Quattar et al., 1987; Miralles et al., 1998; Commuri and Jones, 2001). High temperature in C3 plants can also favor photorespiration and reduce net CO\textsubscript{2} assimilation during portions of the day, reducing productivity (Zelitch, 1982). Though not addressed in this study, in wheat data the net effect has been reported to be a yield reduction of 4 and 6% per °C in yield (Liu et al., 2016), although these should be considered as upper bounds because adaptation through planting date or phenology has not been systematically assessed. With no room for adaptation, warmer regions will likely suffer yield losses in the order of 5% per °C. Other factors like dryness of the atmosphere quantified through the vapor pressure deficit (VPD, hPa) have also been reported as directly linked to crop yield (Lobell et al., 2014). Composite variables can be created from multiple climate variables to represent a more complete snapshot of the environmental conditions driving yields. For instance, the photothermal coefficient (ratio of intercepted radiation to thermal time for a given phase; (Fischer, 1985)), or the photothermal coefficient further normalized by the VPD, are used (Sadras et al., 2011; Ernst et al., 2016). These composite variables are informative, but they may not provide new information to a regression when compared to the individual fundamental variables that comprise the composite variable. A systematic approach to address the actual importance of each variable while accounting for variable interactions has not been addressed yet.

In this paper, we apply RF as a diagnostic crop model using high resolution data in the Midwest region of the United States. Our specific objectives are to: (1) quantify the importance of precipitation, maximum and minimum temperature and solar radiation in determining the yield of the summer crops corn, sorghum and soybean; (2) obtain response curves to these variable for different phases of each crop growing season; and (3) test if composite variables provide additional information that improves the modeled response of crops to climate. Specifically, we analyze how climate has affected sorghum, corn, and soybean yields from 1980 to 2016 in 18 states in the central region of the United States. Ultimately, this knowledge could help us design effective pathways for adaptation, and understand the crop response to non-climate factors.
3.2 Data

3.2.1 Yield data

This work focuses on yield responses in the central portion of the United States, as well as the a portion of the High Plains and Southwest (Figure 3.1). We selected this region because it contains states that produce the majority of sorghum, corn, and soybean in the US. We used county-level yield data for each state from the United States Department of Agriculture National Agriculture Statistics Service (USDA NASS). These are higher quality data than those used in Hoffman et al. (2017) and are available on a much higher spatial resolution. Yield data on the county-level reduces the need to aggregate climate data, which can potentially obscure subtle yield responses in statistical crop models and bias predictions in process-based crop models (Pierce and Running, 1995; Baron et al., 2005; Tack et al., 2015; Hoffmann et al., 2016; Hoffman et al., 2017).

Yield data from the USDA is available from census and survey data, but because the census does not report county-level yields, we used survey data in this work. Survey data is partitioned into various sectors and further into groups. The yield data used in this study come from the field crop group in the crop sector. We extracted the irrigated, rainfed, and unspecified yield data for sorghum, corn, and soybean. We attempted to preserve as much yield data as possible while limiting contamination from irrigated yields. If rainfed yields exist for a given county and year, that yield is included. However, rainfed yields are inconsistently reported and often not reported at all, so we filled in missing data in this record with unspecified yields. Since most irrigated area is west of the 100th meridian, instead of eliminating these records altogether (as in Schlenker and Roberts (2009)), we filtered out potentially irrigated yields using census data from the 2012 Census of Agriculture (USDA NASS, 2012). In general, irrigation has increased with time, so using irrigation data from this decade will likely make sure our estimates are conservative. For each reporting county, we compute the area irrigated as a function of area planted. If that fraction is greater than 25%, we eliminate all yields from those counties. In the case that irrigation is reported, but unpublished (denoted by ‘D’ in the record) we assume that the county is irrigated. If no irrigation is reported, we assume the county is primarily rainfed. Therefore, if the fraction irrigated is
Figure 3.1: Average estimated planting dates for corn. Dark colors represent earlier in the year, while light colors represent later. We have plotted a planting date for every county in which at least one year reported corn yields that meets the irrigation criteria.

less than 25%, we allow the unspecified yields to fill in when available. While low yields are permitted, we only consider the records from a county that contains at least one instance of non-zero and non-missing data. All yields were converted from bushels per acre to kg/ha assuming 13%, 15.5%, and 13% moisture for sorghum, corn (grain), and soybean, respectively.
Table 3.1: Potential climate predictors used in this study. Fundamental variables used in the final models are denoted with an asterisk. Throughout the text, each variable may be referenced for an individual phase, in which case the variable will appear as below, but appended with a dash and phase abbreviation: ES, CW, GF, and GS for establishment, critical window, grain filling, and growing season, respectively.

<table>
<thead>
<tr>
<th>Fundamental variables</th>
<th>Composite variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMAX* (°C) daily maximum temperature</td>
<td>RMIN (%) daily minimum relative humidity</td>
</tr>
<tr>
<td>TMIN* (°C) daily minimum temperature</td>
<td>RMAX (%) daily maximum relative humidity</td>
</tr>
<tr>
<td>EDD* (°C d) extreme degree days (calculated via cap method)</td>
<td>SPH (kg of water vapor/kg of air) daily mean specific humidity</td>
</tr>
<tr>
<td>PRCP* (mm) accumulated precipitation</td>
<td>PDSI (unitless) Palmer drought severity index</td>
</tr>
<tr>
<td>VPD* (hPa) vapor pressure deficit</td>
<td>FDIFF (unitless) fraction of diffuse radiation Bristow et al. (1985)</td>
</tr>
<tr>
<td>SRAD* (W/m²) surface downwelling shortwave flux</td>
<td>SRADTMP (x) adjusted solar radiation for optimal temperature conditions Olson et al. (2012)</td>
</tr>
<tr>
<td>WINDSPEED (m/s) daily mean wind speed</td>
<td>SRADVPD (x) adjusted solar radiation for optimal temperature and vapor conditions Olson et al. (2012)</td>
</tr>
<tr>
<td>TAVG (°C) daily average temperature</td>
<td></td>
</tr>
</tbody>
</table>

### 3.2.2 Climate data

We used daily weather data at 4km resolution from 1980 to 2016. These data were taken from the MetData high-resolution gridded surface meteorological product from the University of Idaho (Abatzoglou, 2011). We initially considered 15 possible variables as predictors (Table 3.1). Raw MetData variables are denoted in plain text, while those computed from the raw data are denoted in bold letters. Year is also included as a potential predictor in our analysis to capture the long-term trend of improving technology with time. For each climate predictor in Table 3.1, we compute four values based on the phenological growth stages for each crop (as described next).
3.2.3 Methods to compute growing season phases

3.2.3.1 Planting Date

For each year, we estimate the planting date for each county based on county-level temperatures. Other researchers used latitude and elevation to estimate planting dates (as in Yang et al. (2017)), but using temperatures allowed us to naturally adapt to cold or warm years (i.e. if the year is colder than average, a farmer can delay planting later to decrease the risk of early season frost damage). We estimate that planting occurs once the 21-day moving average rises to a crop-specific threshold temperature (Table 3.2). Planting temperatures for sorghum, corn, and soybean are 10°C, 12°C, and 15°C, respectively. While there is some latitudinal variation in these planting thresholds, a single threshold for each crop is chosen for simplicity and we note that our results still lie within the range of observed planting dates and these thresholds agree with other work (Bondeau et al., 2007; Waha et al., 2012; Yang et al., 2017; National Agricultural Statistics Service, 2012). Additionally, we assume planting occurs on a single day, though it realistically takes approximately 3 weeks to plant 80-90% of the area once planting conditions are achieved.

We implemented a frost restriction while computing sorghum and soybean planting dates because of their relatively low tolerance for cold temperatures. When a frost restriction is imposed, the planting date occurs when the planting threshold is reached only after the last early frost has occurred. If threshold temperatures for planting are never met, the crop is not planted (typically only true for high elevations).

3.2.3.2 Growth stages

We considered the entire growing season but also partitioned the growing season for each crop into three distinct growth phases: establishment, critical window, and grain filling. Overall, this approach should provide a more consistent analytical basis than calendar-based approaches, except that it does not account for planting and harvest time windows, or differences in phenology among planted genotypes, both of which are arguments in favor of a wider calendar window.

The establishment phase is meant to capture conditions during germination,
emergence, and early vegetative stages. The critical window aims to capture the reproductive phases that determine yield potential, and the grain filling phase captures the conditions important during the last phase of grain growth. There is some unavoidable overlap between consecutive phases. These phases represent the phenology well for corn and sorghum, but for soybean, the staggered flowering and pod filling phases prevents a distinct outline of phases for a single plant. For soybean, we used the phases and thermal times reported in Setiyono et al. (2007) and we consider the establishment phase to contain planting, all of the vegetative phase and the first reproductive phase (R1), the critical window to contain the second through fourth reproductive stages (R2-R4), while grain filling contains the final stages of the reproductive phase (R5-R8) (Berglund and McWilliams, 2015; Fehr and Caviness, 1977).

The planting date algorithm was designed to allow for year to year adaptation, and the growing season algorithm was designed to allow for early termination of the grain filling phase under extreme conditions. For all crops, the length of these phases is determined by thermal time and crop-specific base temperature. Using thermal times to compute the growth phases results in a shorter growing season under warm conditions. In the event that the plant experiences extremely cold late-season conditions we force end grain filling prematurely. We consider harvest for all crops to occur once grain filling is completed, and losses of yield post-maturity are not accounted for.

Because planting data is difficult to obtain, there have not been many statistical crop modeling studies which compute phase-specific crop responses. Lobell et al. (2014) partitioned the growing season into five successive 30-day periods to estimate the crop responses to different periods of growth, but did not relate these periods back to the respective growth phases of maize and soybean. Phase-specific crop responses have been traditionally studied using processed based crop models (Grassini et al., 2011).

### 3.2.4 Modeling

RF is a non-parametric regression technique based on a special ensemble of classification and regression trees (CART) (Breiman et al., 1983; Breiman, 2001). RF models are constructed by using an ensemble of CARTs wherein each tree is grown
Table 3.2: Phase-specific thermal times, base temperature, and planting thresholds for sorghum, corn, and soybean. Asterisk denotes that a frost restriction is in place. The unit °C d represents a degree day.

<table>
<thead>
<tr>
<th>Phase Abbrev.</th>
<th>Sorghum</th>
<th>Corn</th>
<th>Soybean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Establishment</td>
<td>ES 600°C d</td>
<td>600°C d</td>
<td>500°C d</td>
</tr>
<tr>
<td>Critical window</td>
<td>CW 400°C d</td>
<td>600°C d</td>
<td>500°C d</td>
</tr>
<tr>
<td>Grain filling</td>
<td>GF 600°C d</td>
<td>850°C d</td>
<td>1000°C d</td>
</tr>
<tr>
<td>Growing season</td>
<td>GS 1600°C d</td>
<td>2050°C d</td>
<td>2000°C d</td>
</tr>
<tr>
<td>$T_{\text{base}}$</td>
<td>- 8°C</td>
<td>8°C</td>
<td>5°C</td>
</tr>
<tr>
<td>$T_{\text{plant}}$</td>
<td>- 15°C *</td>
<td>10°C</td>
<td>12°C *</td>
</tr>
</tbody>
</table>

using a random sample of the training data and random predictor selection at each node of each tree. The use of an ensemble reduces variance in the final model, while the random predictor selection reduces the correlation between trees and keeps the bias low. Finally, predictive performance of RF only increases when unique information is added by predictors, so correlated variables do not affect the performance of the model. However, the importance of correlated predictors is split among correlated variables.

RF can be used to estimate the sensitivity of yields to individual climate predictors and estimate their functional responses (Breiman, 2001). Though many of the predictors in this kind of analysis are closely related, the relation can be explored and made explicit. In Hoffman et al. (2017), we showed that RF is a useful tool in understanding the drivers of crop yields in sub-Saharan Africa.

For this work, we explicitly defined a set of fundamental climate variables that represent the most basic climate variables relevant for crop yields (Table 3.1). These fundamental variables can be used to derive more complex climate variables (e.g. maximum temperature and vapor pressure are used to compute minimum relative humidity). A powerful feature of RF is its ability to compute and rank variables by their importance to predicting the response variable. These complex, or composite variables are a better representation of the environmental conditions, so naturally they arise as more important to a regression. However, whether these variables actually provide new information through variable interactions is unknown. An objective of this work was to test if composite variables are necessary to understand the response of crops to climate, or if they simply repeat information.
Each crop was modeled independently. Each model consists of all six fundamental variables and the year of harvest as predictors. We also include all four phases: establishment, critical window, grain filling, and the growing season average. Though it is not required to withhold a test set using Random Forest, we set aside 25% of the data in each model for testing.

We considered the following states: New Mexico, Colorado, Texas, Oklahoma, Kansas, Nebraska, South Dakota, North Dakota, Wisconsin, Minnesota, Iowa, Michigan, Ohio, Illinois, Indiana, Missouri, Arkansas, and Louisiana for a total of 1631 counties. We allow for zero and spuriously low yields (see Section ). Models for sorghum, corn, and soybean have 17,973, 29,678, and 27,251 records, respectively.

### 3.3 Results

In this section we examine the results from this work. We generalize the model performance in the first section, followed by a crop-specific breakdown of yield responses to each predictor.

#### 3.3.1 Model performance

Overall, the RF crop models captured between 71%, 86%, and 81% of the variance in sorghum, corn, and soybean yields, indicating that the models are high-performing (Figure 3.2). For reference, the analogous figures for corn and soybean in Schlenker and Roberts (2009) were 77 and 63%, respectively. In Table 3.3, the variable importance metric highlighted that yield dependence on year (i.e. time) is the most important variable in each crop model. The yield dependence on time was approximately linear, indicating that the average corn yield increased by 90 kg/ha (1.3%) per year. The corresponding figures for both sorghum and soybean are 29.3 kg/ha (0.79%) and 29.4 kg/ha (1.1%). To determine which variable and growth phase had the largest effect on yields, we utilized the permutation accuracy importance measure. This variable importance measure is computed by measuring the change in model error as a result of permuting a single predictor; if the variable is important, the permutation error is large. In general, we found that yields were largely driven by environmental conditions during the grain filling period. However, accumulated precipitation and extreme degree days over the entire growing season...
had a significant effect on yields, particularly for corn (Table 3.3). Though VPD during the grain filling period (VPD-GF) drove sorghum and soybean yields, corn yields were driven by VPD during the critical window (VPD-CW). Yield sensitivity to solar radiation was low when compared with the other fundamental variables; ranking in the bottom 40% of variables (Table 3.3). Next, we discuss the individual results for each crop.

### 3.3.2 Sorghum

The model for sorghum yields had the lowest explanatory power of all three crops with an R$^2$ of 0.713 (Figure 3.2a). Following time, the most important yield driver was vapor pressure deficit during grain filling (Figure 3.3d). While sorghum was most sensitive to environmental conditions during grain filling, precipitation and extreme degree days accumulated over the entire growing season were more informative than their phase-specific counterparts (Figure 3.4a,d, 3.3a,d, and 3.5a). The effect of solar radiation variability on yields was minor (Table 3.3). On average, grain yields decreased by 44 kg/ha (1.2%) for every 1hPa increase in
Table 3.3: Variable importance for the fundamental variable models for sorghum, corn, and soybean. Numbers beneath crop names denote $R^2$ values for each model. Variable importance for sorghum basis models. Variable importance was calculated using the raw, unscaled permutation accuracy importance measure (Strobl et al., 2007). Unscaled data should only be compared within single columns (models). Higher values for raw permutation importance measure indicate more important variables. Average rank of each variable-phase combination is included in the rightmost column; higher ranks denote more important variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>sorghum</th>
<th>corn</th>
<th>soybean</th>
<th>Average Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMAX-ES</td>
<td>111094</td>
<td>279181</td>
<td>19184</td>
<td>3.67</td>
</tr>
<tr>
<td>TMAX-CW</td>
<td>87485</td>
<td>668761</td>
<td>31561</td>
<td>9.00</td>
</tr>
<tr>
<td>TMAX-GF</td>
<td>386453</td>
<td>899946</td>
<td>182738</td>
<td>21.67</td>
</tr>
<tr>
<td>TMAX-GS</td>
<td>152822</td>
<td>253232</td>
<td>44666</td>
<td>9.00</td>
</tr>
<tr>
<td>TMIN-ES</td>
<td>180936</td>
<td>372121</td>
<td>19963</td>
<td>9.33</td>
</tr>
<tr>
<td>TMIN-CW</td>
<td>163461</td>
<td>428806</td>
<td>36333</td>
<td>12.33</td>
</tr>
<tr>
<td>TMIN-GF</td>
<td>362413</td>
<td>1033885</td>
<td>161897</td>
<td>21.33</td>
</tr>
<tr>
<td>TMIN-GS</td>
<td>316584</td>
<td>594050</td>
<td>82417</td>
<td>18.00</td>
</tr>
<tr>
<td>PRCP-ES</td>
<td>110435</td>
<td>310734</td>
<td>20598</td>
<td>5.33</td>
</tr>
<tr>
<td>PRCP-CW</td>
<td>90328</td>
<td>386066</td>
<td>16389</td>
<td>5.33</td>
</tr>
<tr>
<td>PRCP-GF</td>
<td>267309</td>
<td>311301</td>
<td>58884</td>
<td>14.33</td>
</tr>
<tr>
<td>PRCP-GS</td>
<td>405500</td>
<td>1327530</td>
<td>39213</td>
<td>20.00</td>
</tr>
<tr>
<td>SRAD-ES</td>
<td>124514</td>
<td>329092</td>
<td>19435</td>
<td>6.00</td>
</tr>
<tr>
<td>SRAD-CW</td>
<td>85191</td>
<td>308509</td>
<td>22174</td>
<td>4.00</td>
</tr>
<tr>
<td>SRAD-GF</td>
<td>153644</td>
<td>306490</td>
<td>40820</td>
<td>10.00</td>
</tr>
<tr>
<td>SRAD-GS</td>
<td>189655</td>
<td>247272</td>
<td>33471</td>
<td>9.33</td>
</tr>
<tr>
<td>VPD-ES</td>
<td>153021</td>
<td>344308</td>
<td>27813</td>
<td>8.67</td>
</tr>
<tr>
<td>VPD-CW</td>
<td>163721</td>
<td>1332811</td>
<td>27621</td>
<td>14.33</td>
</tr>
<tr>
<td>VPD-GF</td>
<td>589232</td>
<td>739364</td>
<td>87386</td>
<td>21.67</td>
</tr>
<tr>
<td>VPD-GS</td>
<td>543623</td>
<td>712807</td>
<td>31594</td>
<td>16.67</td>
</tr>
<tr>
<td>EDD-ES</td>
<td>174386</td>
<td>382914</td>
<td>36547</td>
<td>12.67</td>
</tr>
<tr>
<td>EDD-CW</td>
<td>106905</td>
<td>722930</td>
<td>33128</td>
<td>11.00</td>
</tr>
<tr>
<td>EDD-GF</td>
<td>194104</td>
<td>498190</td>
<td>86837</td>
<td>17.00</td>
</tr>
<tr>
<td>EDD-GS</td>
<td>233306</td>
<td>124538</td>
<td>62727</td>
<td>19.33</td>
</tr>
<tr>
<td>year</td>
<td>603215</td>
<td>3965215</td>
<td>341616</td>
<td>25.00</td>
</tr>
</tbody>
</table>
Figure 3.3: Partial dependence plot for phase-specific accumulated precipitation (PRCP) and vapor pressure deficit (VPD) for each growth phase, including the growing season. Colors represent the growth phase: establishment (blue), critical window (green), grain filling (red), growing season (black). Responses from the 2.5-97.5th quartiles are plotted.

VPD-GF. Interestingly, at very high VPD-GF (greater than 15hPa) sensitivity to increasing VPD-GF is reduced and yields only drops by 3.8 kg/ha. This reduced sensitivity is also evident in the response to accumulated extreme degree days over the growing season (EDD-GS). In general, yields will decrease by about 19 kg/ha for every extra 10 °C d EDD-GS, but once 80 EDD are accumulated, that response drops to less than 1 kg/ha. However, flattening at the edges of these plots, called partial dependence plots, is quite common with RF and should not be considered a robust part of the signal.

Optimal temperature ranges emerged for both minimum and maximum temperatures during grain filling, TMIN-GF and TMAX-GF, respectively (Figure 3.4a,d). Optimal TMIN-GF ranges from 14.5°C to 19°C, while the optimal TMAX-GF
ranges from 25°C to 30°C (Table 3.4). Though sorghum yields exhibit optimal temperature ranges, yields were likely to be above average so long as TMIN-GF ranged from 12.5°C and 20°C and TMAX-GF remained under 32.5°C.

Sorghum yields were highest when TMAX-GF was near 29°C, but the most rapid decline in yields did not occur until 32.5°C indicating that sorghum can withstand warmer temperatures than corn or soybean (Figure 3.4a). Yet, sorghum is also sensitive to cooler minimum temperatures, and yields benefit from increased TMIN throughout the growing season, so long as TMIN-GF does not exceed 19°C (Figure 3.4d).

Sorghum yields optimized when approximately 420 mm of precipitation accumulated over the growing season. When PRCP-GS was less than optimal, sorghum yields steadily increased by about 90 kg/ha per 50mm until 330 mm accumulated, after which the benefit reduced to 23 kg/ha until 420 mm was achieved. If more
than 420mm accumulated during the growing season, yields decreased by about 13 kg/ha with each additional 50 mm. While PRCP-GF is not as important as PRCP-GS, we found that precipitation accumulated during this phase consistently increases yields (Figure 3.3d). Finally, in cooler regions, there is a distinct yield decrease when PRCP-ES exceeded 140 mm (not shown).

### 3.3.3 Corn

The fundamental variable model for corn yields explained the most variability of the three crops with an $R^2$ of 0.857. Corn yields were primarily driven by VPD during the critical window (Figure 3.3e). In general, corn yields decreased by 127kg/ha (1.9%) per 1 hPa increase in VPD during the critical window.

More than any other crop, extreme degree days accumulated throughout the growing season (EDD-GS) were critical to explaining historical yields (Figure 3.5e). In general, if more than EDD-GS exceeded 55 °C d, yields dropped below average.

Similar to sorghum, this analysis revealed optimal TMIN-GF and TMAX-GF ranges for corn growth. Yields optimized when TMIN-GF ranged from 8-19°C and TMAX-GF ranged from 21-29°C (Table 3.4). Once TMAX-GF > 29°C yields rapidly decreased by 82 kg/ha (1.2%) for every additional °C. Increased in TMIN consistently increased corn yields, so long as TMIN-GF did not exceed 16°C (Figure 3.4e).
Corn yields optimized when approximately 628 mm of precipitation accumulated over the growing season (Figure 3.3e). At low levels of precipitation, increasing precipitation by 50 mm steadily and linearly increased yields by 316 kg/ha (4.8%). This higher response of corn likely reflects that it grows in a moister environment (lower VPD) where evaporation is more subdued. Unlike sorghum, yield response to accumulated precipitation is not phase-specific; that is no particular growth phase is significantly more important than another for precipitation. There is no significant yield response to variability in solar radiation (Table 3.3).

3.3.4 Soybean

The fundamental variable model for soybean had an $R^2$ of 0.811, and grain filling was clearly the most important phase of growth (Table 3.3). Yields respond most strongly to TMAX-GF with yields slowly increasing until 30°C and then falling off sharply at 98 kg/ha (4.2%) with each additional degree (Figure 2c). These results indicated an optimal range of both TMIN-GF and TMAX-GF between 9-19°C and 23-29°C, respectively.

Yields are also affected by vapor pressure deficit during grain filling, VPD-GF. In general, yields decrease by 31 kg/ha (1.3%) per 1hPa increase in VPD-GF.

Unlike what was observed with sorghum and corn, soybean yields responded to extreme degree days and precipitation accumulated during the grain filling phase, EDD-GF and PRCP-GF. Yields increase by approximately 71 kg/ha (3%) per 50 mm until 200 mm have accumulated during this phase. Once 200mm accumulates, yield improvement stagnates. Yield response curves to precipitation accumulated over the growing season indicates that yield optimize at 494 mm. Similar to observations for sorghum, if too much precipitation was accumulated during the establishment phase (PRCP-ES > 120 mm) yields decreased.
Table 3.4: Optimal grain filling temperature ranges, and optimum temperature for both TMAX and TMIN during grain filling (derived from partial dependence information plotted in Figure 3.4).

<table>
<thead>
<tr>
<th></th>
<th>TMAX</th>
<th>TMIN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>min</td>
<td>max</td>
</tr>
<tr>
<td>sorghum</td>
<td>25.3 °C</td>
<td>31.1 °C</td>
</tr>
<tr>
<td>corn</td>
<td>20.7 °C</td>
<td>29.8 °C</td>
</tr>
<tr>
<td>soybean</td>
<td>21.3 °C</td>
<td>30.4 °C</td>
</tr>
</tbody>
</table>

3.4 Discussion

3.4.1 General

An important objective of this work was to determine whether composite variables are necessary to explain yield responses. Composite variables are not necessary. In an experiment with all four phases, all 15 climate variables from Table 3.1, as well as year and two variables that measured cold temperature strain during grain filling, the explanatory power of the model marginally increased (Table 3.5). Fundamental variables are sufficient to explain a large fraction of the yield variance from 1980-2016 in the central US.

3.4.2 Variable effects

On average, sorghum, corn, and soybean yields are most sensitive to vapor pressure deficit, maximum temperature, minimum temperature, precipitation, extreme degree days, and solar radiation, respectively.

Sorghum, corn, and soybean yields are strongly affected by VPD, and increasing VPD typically leads to a reduction in yield (Figure 3.3d-f). Over the region of interest, VPD over the growing season has increased by 1.4 hPa, 1.2 hPa, and 1.3 hPa for sorghum, corn, and soybean, respectively over the time series. These values are significantly smaller than those found by Zhang et al. (2017) over various parts of China, but we observe similar yield responses for soybean and corn. These results also agree with Hoffman et al. (2017) which found a linear response of yields in sub-Saharan Africa to increasing VPD.

Recent work has shown that the sensitivity of corn yields to VPD has increased
Table 3.5: Comparison of explanatory power ($R^2$) between crop models that only include fundamental variables and those that include all variables from Table 3.1. Models with all predictors also include two variables measuring cold temperature strain during grain filling, while both models include year as a predictor.

<table>
<thead>
<tr>
<th></th>
<th>sorghum</th>
<th>corn</th>
<th>soybean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fundamental variable model</td>
<td>0.713</td>
<td>0.857</td>
<td>0.811</td>
</tr>
<tr>
<td>All variable model</td>
<td>0.738</td>
<td>0.869</td>
<td>0.826</td>
</tr>
</tbody>
</table>

over time, likely as a result of increasing planting density or warming temperatures (Lyon et al., 2003; Lobell et al., 2014). However, we find no interaction or correlation of VPD with time.

While corn yields are largely responsive to climate during grain filling, corn yields are extremely sensitive to VPD during the critical window (Figure 3.3e). A similar response was found using multivariate adaptive spline regression on data from Iowa, Illinois, and Indiana (Lobell et al., 2014). Lobell et al. (2014) found that VPD between 61-90 days after planting (roughly the critical window - see Figure S3.4) was the primary driver of corn yields. While the general response is similar, the VPD used by Lobell et al. (2014) differs from the one we used; we used the average VPD for the day, while Lobell et al. (2014) used the maximum VPD of the day. Another reason the threshold varies is because our area of study was much broader and spans a variety of different climates. When Lobell et al. (2014) cooler states from the Corn Belt were included in their analysis, the VPD threshold dropped down to 20 hPa. Because our work spanned a much broader geographic region, we show that the corn yield response is broader and more prevalent than previously concluded.

It is well-documented that corn is open pollinated and drought stress causes pollination failures as well as abortion of fertilized ovules. While it is logical that dry periods are characterized by large VPD, but it is surprising that this drought signal (during the critical window) in corn is picked by VPD rather than by PRCP. This may indicate that atmospheric dryness, not low soil moisture, is what leads to difficult pollination, pollen germination, or pollen receptivity.

Temperatures during grain filling emerge as an important driver of yields in the central US. The conclusion that temperature is most important during grain filling has been found before (Lobell et al., 2014; Prasad et al., 2015). Prasad et al. (2015) found that average temperatures ranging from $25^\circ$ to $37^\circ$C quadratically
decreased individual grain weight when imposed at the start of grain filling as a result of decreased floret fertility. For sorghum, TAVG-GF of 25°C corresponds to TMAX-GF of 30°C, indicating very similar temperature responses. Similarly, Lobell et al. (2014) found that maximum temperatures during the last month of growth was a primary driver of corn and soybean yields in their work, although their work focused only on Illinois, Indiana, and Iowa (Lobell et al., 2014). One reason that temperature during grain filling may be emerging as an important yield driver is because we assume planting to occur on a single day of the year. In reality, it takes approximately 3 weeks to plant 80-90% of the area once planting conditions are achieved. Therefore, it is possible that the grain filling period may reflect some critical phase yield effects.

Modeled yield responses of corn and soybean yields to TMAX-GF compare favorably to those found in other work wherein yields increase slowly with increasing temperature until a threshold, and then fall off rapidly (Schlenker and Roberts, 2009; Lobell et al., 2014). We estimate the threshold temperatures at approximately 29°C and 30°C for corn and soybean, respectively. Given the areas in which sorghum is grown, we expected to see warmer optimum temperatures for TMAX-GF (Table 3.4). However, the threshold response is not as well-defined for sorghum as it is for corn and soybean, indicating that sorghum is able to withstand warmer temperatures (up to about 32°C) before dropping rapidly.

Unlike previous work, we found that yields exhibit distinct responses to TMIN throughout the growing season (Figure 3.3d-f). Previously, the distinction in yield responses to TMAX and TMIN had not well-defined in statistical crop models due to the correlation between the two variables. In general, we find that yields will improve as TMIN-GS increases, indicating that within the observed range, warm nights benefit yields. All three crops indicate that increasing the minimum temperature in the establishment phase will increase yields, though this response is most pronounced for sorghum. This is not to imply that a warmer climate will benefit yields, because warm nights are linked to warmer days, and a strong threshold response to TMAX is evident with all three crops. We found that sorghum yields increase 43 kg/ha (1.2%) for every degree TMIN-ES increases. While it is well known that sorghum has lower tolerance for cold temperatures than corn, our analysis quantifies the response and allows to consider it in predictive models in a systematic fashion.
These results also reveal distinct optimal ranges for TMIN and TMAX during the grain filling periods of sorghum, corn, and soybean (Table 3.4).

The impact of TMAX on yields in this study is distinctly nonlinear while Hoffman et al. (2017) found that the effect of increasing TMAX was predominantly linear. An important difference between these two studies is the quality and resolution of data. When we build the fundamental variable model using only growing season values, as was done in Hoffman et al. (2017), the shape of the functional forms remains largely the same for all variables except for TMAX. For corn, the response becomes predominantly linear and negative, however there is still a distinct nonlinear signal for sorghum (not shown).

Extreme degree days accumulated over the growing season (EDD-GS) are important to sorghum, corn, and soybean yields, but have had a disproportionately large effect on corn (Figure 3.5). While it appears that there is decreased yield sensitivity to high EDD (EDD-GS>150), we cannot confidently distinguish this response from a model artifact. The 'flattening' at extreme values of EDD is likely a feature of RF. As noise increases, the more conservative predictions (i.e. regressions toward the mean) will become. At the bounds of our data, there are fewer data points, and thus more noise. However, there is evidence to suggest that this response may also be due to decreased yield sensitivity to high temperatures if the crops are continuously stressed (Schlenker and Roberts, 2009; Prasad et al., 2006, 2008).

As expected, sorghum, corn, and soybean yields are strongly affected by precipitation accumulated over the growing season, and exhibit nonlinear responses to accumulated precipitation (Figure 3.3a-c). The modeled yield response curves are similar to those in other work wherein yields show significant linear increases until a threshold is reached, after which yields are relatively insensitive to more precipitation, but decrease slightly if too much precipitation is accumulated (Sadras et al., 2011; Hsiang et al., 2013; Anderson et al., 2015; Hoffman et al., 2017).

Our results outline environmental conditions for optimized yield. If TMAX-GF and TMIN-GF lie in the optimal ranges, and if an adequate amount of precipitation is accumulated for each crop, we consider the crop to be in ideal conditions; these conditions are plotted for corn in Figure 3.6 for 2016 and 2012. Crop yields in 2012 were historically low as a result of drought and a very warm growing season, while the year 2016 was quite average. Unsurprisingly, optimal conditions for corn
growth in 2016 occur in the Corn Belt, but it is clear why 2012 had historically low yields. Very few counties met the conditions for optimal growth. The band of optimal TMAX-GF temperature shifted north, while precipitation was inadequate in the majority of the central US, severely limiting yield. Though accumulated precipitation over the course of the growing season was important, precipitation accumulated during grain filling was more important to driving yields than any other phase. In particular, we found that soybean yields were highly sensitive to precipitation during grain filling, which is likely because soybean is most sensitive to water deficits during reproduction, and grain filling represents the longest duration of that phase (R5-R8). Additionally, if our estimation of the critical window is too early, the grain filling phase may be capturing the entire reproductive phase (R1-R8).

Though yields are less sensitive to PRCP accumulated in the establishment phase, excessive precipitation during this phase decreased in yields. Corn yields exhibit maximum yield reduction (in a soil water study) when excessive precipitation occurred within 36 days of planting as a result of delayed emergence, poor soil aeration and increased susceptibility to pest and disease (Mukhtar et al., 1990).

Until now, we neglected to assess the impact of SRAD on yields simply because the other variables are more important (Table 3.3). Based on theory, we know that solar radiation is an important fundamental variable, but the results of the model are not showing this. For this reason, we tested two separate permutations of the SRAD variable: (1) solar radiation adjusted for optimal temperature conditions (SRADTMP) and (2) solar radiation adjusted for optimal temperature and atmospheric moisture conditions (SRADVDPD). However, this was not a directly comparable test because we effectively created a composite variable in which more than one environmental variable was represented. For each of these composite solar radiation variables, the importance increased.

These results do not validate those of (Tollenaar et al., 2017) which suggested that solar brightening has caused corn yields to increase from 1983-2013. However, we suspect that this effect has been absorbed into the time (year) variable because the increase in solar brightening has been consistently increasing since 1983 (Figure 1 in Tollenaar et al. (2017)). Despite possible issues with the variability in the SRAD variable, we observe an increase in SRAD-GF from 1980-2016 in line with Tollenaar et al. (2017), but we only observe half (650 MJ/m2) of what they report.
While environmental conditions undoubtedly drive crop yields, this work identifies time (year) as the most important variable in the analysis. This result is unsurprising given the results in Grassini et al. (2013) and Hoffman et al. (2017). However, we caution the reader because of the temporal inconsistency in reporting yields (discussed in Section 3.4.3.4). Not every county reports yield data each year and each year does not represent the same spatial coverage. Interestingly, year does not arise as one of the three most important predictors in Lobell et al. (2014).

To determine whether our results depended on location, we computed average latitude within a county and included it as a predictor. We found no changes in any of the temperature thresholds. However when latitude was added as a predictor, the steepness of the slope of corn yields once TMAX-GF > 29°C decreased; a response very similar to what Schlenker and Roberts (2009) found. We did not include latitude as a predictor in the final model because it did not improve the explanatory power of the fundamental variable model, it did not significantly affect the shape of the variable response functions, and it was tightly correlated with temperature.

### 3.4.3 Potential sources of error

#### 3.4.3.1 Growth Phases:

A novel aspect of this work was the estimation of annual planting dates so that we can approximate the yield response to climate during individual growth phases, but it is possible that we introduced additional uncertainty to our model as a result. While planting dates are largely dependent on environmental conditions, farmers also use historic knowledge to inform their decision making which is not included in this work (Waha et al., 2012). Similar to the methodology of Waha et al. (2012), it could be worthwhile to estimate the planting date in the warmest states (i.e. Texas and Louisiana) using seasonal precipitation as the limiting factor rather than temperature. Another potential source of this uncertainty is using a single planting threshold temperature for each county.

Based on the results and the physiology of soybean, we suspect that there may be several small limitations regarding soybean growth phases. Our results suggest that we are still capturing the establishment, or cool-season, signal with...
the modeled critical window. In other words, the vegetative (establishment) phase for soybean is likely not long enough. Therefore, the phase which we intended to be the critical window is actually contained within the grain filling phase, and the grain filling phase effectively represents all of soybean’s reproductive phases (R1-R8). This is likely why grain filling arises as the most informative phase for every variable.

3.4.3.2 Partial dependence sensitivity to noise:

As mentioned before, a prominent feature in many partial dependence plots is the decreased sensitivity of yields to extreme values on either end of the x-axis. This ‘flattening’ is a feature of random forest, and as noise increases, the more conservative predictions will become. At the bounds of our data for each predictor variable, there are fewer data points, and thus more noise. Though we often plot the innermost 95th percentile of data, the effect is still evident in many cases. These features should not be considered robust.

3.4.3.3 Scale:

We do not parse out field-level crop information or mask out areas of a county that do not grow a given crop. We elected to estimate the county averages in this manner because the available data for crop areas are calibrated to represent the year 2000 (Sacks et al., 2010). Though temporally varying crop areas can potentially be estimated using satellite data (Schlenker and Roberts, 2009), the use of satellite products introduces additional uncertainty because these algorithms are based on single-cropped systems.

We assume there are minimal differences from state to state. However, soil composition, the use of applications (fertilizer, fungicide, insecticide, etc.), and farm-subsidies can change with state. When including state as a predictor, we did not find any significant differences in the shape or thresholds of the yield responses to individual climate variables (only vertical compression).

3.4.3.4 Data quality:

Yield data from the USDA NASS is commonly used for crop modeling studies, but few studies acknowledge the limitations of these data, and we address several
of these here. We note concerns with reporting inconsistencies in irrigated data and county-level yields. First, we note spatial and temporal inconsistencies in the reporting of irrigated, rainfed, and unspecified yields. For instance, since the year 2008 only five states have reported corn yields that discriminate irrigated and rainfed yields (personal comm. Chris Hawthorne). Additionally, if an insufficient fraction of survey responses is collected in a given county for rainfed or irrigated yields of any crop, then the responses are aggregated for a larger area called 'other (combined) counties' (Johnson, 2014; Bock and Kirkendall, 2017). These counties were omitted from our analysis because the NASS does not record which counties were aggregated in the data.

We used unspecified yields in this study, but masked for counties that irrigate over 25% of the harvested area for a given crop. This was because there were no consistent patterns in the data that would’ve allowed us to extrapolate what fraction of the unspecified yields were rainfed and which were irrigated. For this reason, studies that use rainfed yields from the USDA NASS have not used it to study large regions (Lobell et al., 2013).

Another major issue in the USDA NASS data is that the number of counties for which survey-based yield estimates are available (irrigated, rainfed, or unspecified) has been steadily decreasing with time (Bock and Kirkendall, 2017). By 1998, the halfway point in this analysis, 70%, 55%, and 54% of yields had already been reported for sorghum, corn, and soybean, respectively. A major reason for this discrepancy is because yield data from aggregated areas cannot be utilized, and aggregating yields has become more common.

Lastly, we considered all yields from a given county provided that there was one instance of non-zero, non-missing data.
Figure 3.6: Optimal temperature and precipitation ranges are plotted for corn in 2016 (top) and 2012 (bottom). Counties which lie within the optimum TMAX and TMIN ranges are highlighted in red and blue, respectively. Counties which experienced more than 400 mm of precipitation are colored grey. Counties that experienced optimal growing conditions are therefore darkest (dark purple).
3.5 Conclusions

In this work, we used random forest to estimate the drivers of sorghum, corn, and soybean yield in the central United States from 1980-2016. To do this, we needed to develop a climate-crop data set that discriminated between various growth phases of each crop. We found that yields are sensitive to environmental conditions during the grain filling stage, but corn yields are influenced by conditions during the critical window. These results identify optimal TMAX and TMIN ranges for yield during the grain filling phase. Corn yields are driven by VPD during the critical window and EDD accumulated over the growing season. We found that in general, increasing TMIN-GS will increase yields, indicating that a warming climate may improve yields, so long as TMIN doesn’t get too warm during grain filling. Using corn as an example, warm and dry years like 2012, shrink the area of optimum growing conditions and shift it northward toward higher latitudes and westward toward higher elevations (Figure 3.6). This can be considered an indication of what’s to come if the frequency of heat stress increases. While climate is undoubtedly important, technological advancements appear to be the main driver of crop yields in this region.
3.6 Supplement

The supplement elaborates on the methodology of the climate data as well as an important feature about RF.

3.6.1 Climate Data:

This work utilized several variables not included in MetData (see bold variables in Table 3.1). For each of these variables, the appropriate MetData variables were used to compute the new variable at each individual grid point for each year and each day. These values were subsequently aggregated to the county-level and used to compute the phase-specific variables used in the analysis. Though variables like temperature show small daily variations across a county, we elected to compute these variables at each individual grid points (instead of using the county average) for consistency with the MetData.

3.6.1.1 TAVG

Daily average temperature was computed in a simple manner:

\[ TAVG = \frac{(TMAX + TMIN)}{2} \]  (3.1)

Many crop models use a skewed value so that it is more representative of daytime conditions, which are arguably more important to plants.

3.6.1.2 EDD

The "cap method" is used to compute extreme degree days.

\[ EDD = \sum_{t=1}^{N} DD_{30,t} * normFactor_t \]  (3.2)

\[ DD_{30,t} = \begin{cases} 
0 & TAVG_t < 30^\circ C \\
TAVG_t - 30 & TAVG_t \geq 30^\circ C
\end{cases} \]  (3.3)

where

\[ normFactor_t = \frac{DD_{30,t}}{TMAX_t - TMIN_t} \]  (3.4)
Where \( t \) represents each day, while \( N \) represents the number of days in each growth phase. We normalize EDD by the diurnal temperature because EDD is typically computed with hourly data (as in Lobell et al. (2013)).

### 3.6.1.3 VPD

Vapor pressure deficit was computed using the maximum and minimum values for temperature and relative humidity. The standard means were computed for both temperature and relative humidity. Average temperature was used to estimate the saturated vapor pressure using Teten’s equation while average relative humidity was used to estimate vapor pressure using the approximation that \( RH = \frac{e}{e_s} \). We use the following variation of Teten’s equation:

\[
e_s(hPa) = 6.112 \times \exp\left(\frac{17.269 \times T(\circ C)}{T(\circ C) - 237.3}\right)
\]

There are many ways to compute VPD (Abtew and Melesse, 2013). Lobell et al. (2014) used the difference in saturation vapor pressure computed using TMAX and TMIN. As with average temperature, we could’ve also used average temperature skewed towards the daytime to compute the saturation vapor pressure.

### 3.6.1.4 Derived radiation variables: FDIFF, SRADTMP, SRADVPD

These variables were computed because plants do not utilize solar radiation uniformly under all conditions, under extreme conditions radiation is utilized less efficiently. We use the equations outlined in Bristow et al. (1985), Olson et al. (2012), and Olson et al. (2012) for FDIFF, SRADTMP, and SRADVPD, respectively.

### 3.6.1.5 Flags and frost events

Though not discussed much in the manuscript, the early grain filling termination flag is thrown and the grain filling phase is terminated when the 10-day moving average temperature drops below the base temperature for the grain filling phase. This variable was categorical: early termination of grain filling due to cold temperatures or not. We also computed the number of late-season frosts during grain filling using the minimum temperature to compute how many times the minimum temperature dropped below 0\(^\circ\)C.
3.6.2 Random forest

Though RF is a good model to use when predictors are correlated, estimating the importance of variables is biased (Strobl et al., 2007, 2008; Boulesteix et al., 2011). For instance, if the response variable ($y$) is dependent on $x$, but we include a predictor set of $x$ and a variable correlated to $x$ called $x_{corr}$, the importance of $x$ will reduce and importance of $x_{corr}$ will increase (Kimes, 2006). Including correlated variables does not affect the explanatory power of the model, only the variable importance. For instance, if TMAX-GF is duplicated for sorghum and the fundamental variable model is re-run, the explanatory power of the model does not change significantly (from 0.713 to 0.711), but the variable importance of TMAX-GF decreases by 50% because the information is now shared between two variables. If this experiment is repeated twice more, the explanatory power of the model remains at 0.711, but the importance of TMAX-GF is reduced to about 25% of its original value.
Figure S3.1: Dual partial dependence plot for TMAX and PRCP for sorghum.
Figure S3.2: 3-dimensional partial dependence plot for corn with vapor pressure deficit averaged over the critical window and accumulated precipitation over the growing season. VPD-CW and PRCP-GS are plotted in the horizontal, xy-plane, and the average modeled yield on the vertical, z-axis. Color gradient aligns with average modeled yield.
Figure S3.3: Segment plot depicting average growing season for sorghum. If a state has no data, then sorghum is either not planted there, or is planted in counties with harvest area irrigation over 25%. Blue represents the establishment phase, green represents the critical window, red represents the grain filling phase, while black represents a two-week drying period (not included in this study).
Figure S3.4: Same as Figure S3.3, but for corn.
Figure S3.5: Same as Figure S3.3 and S3.4, but for soybean.
Chapter 4  
Detecting the effect of dust events on crop yields in the central US

4.1 Introduction

Dust plays an important role in the global climate through its effects on radiation, biogeochemical cycles, and atmospheric chemistry (Miller and Tegen, 1998; Prospero, 1999; Mahowald and Kiehl, 2003; Jickells et al., 2005; Mahowald et al., 2009, 2010). Dust is emitted from natural and anthropogenic sources, and the latter is primarily driven by agricultural activities (Tegen and Fung, 1995; Tegen et al., 2004; Ginoux et al., 2012). In North America, dust activity to the west of the Rockies is primarily natural in origin, while dust to the east is anthropogenic (Ginoux et al., 2012). Reaching from Montana and North Dakota to New Mexico and Texas, the High Plains represent the largest dust source (by area) in the United States. This region is semi-arid and contains a large fraction of the farming industry, with cropped lands producing disproportionately more dust than land used for other purposes (Lee et al., 2012). While farming and agriculture has a well-documented effect on dust emissions (Baker et al., 2004; Ginoux et al., 2012; Singh et al., 2012; Lee et al., 2012), this causal chain has yet to be thoroughly examined in reverse.

Wind erosion is comprised of three distinct phases: emission, transport, and deposition (Shao, 2008; Kok et al., 2012). All three phases of wind erosion can affect plants. In source fields where soil is being eroded by wind, yield reductions can be attributed to plant damage like defoliation and seedling stand damage caused abrasive action of sandblasting or by seedling burial (Woodruff, 1956; Armbrust,
This damage results in short-term high-intensity moisture stress caused by loss of stomata control and epidermis damage, which reduces plant survival and growth (Fryrear et al., 1975). Wind erosion can also lead to nutrient removal and modified soil properties, further reducing yields (Zobeck and Bilbro, 2001; Zobeck and Van Pelt, 2014). These experiments have also noted that plants have the ability to recover from abrasion events if water is available (or added) after injury (Woodruff, 1956). The impact of suspended aerosols on plant productivity has been extensively studied (Chameides et al., 1999; Cohan et al., 2002; Gu et al., 2002; Greenwald et al., 2006; Wohlfahrt et al., 2008; Ohde and Siegel, 2012; Xi and Sokolik, 2012; Ozdes, 2012). In general, moderate increases in optical depth will positively affect plants with a canopy, because diffuse photosynthetic active radiation (PAR) can reach light-deprived leaves. This effect is moderated by aerosol type, photosynthetic pathway (C3 vs C4) of the plant, leaf structure, and sky conditions (Cohan et al., 2002; Greenwald et al., 2006; Xi and Sokolik, 2012). Suspended aerosols also tend to decrease water stress in plants by reducing soil evaporation and plant transpiration (Greenwald et al., 2006). Though plant responses vary depending on the composition and size distribution of dust as well as leaf size, deposition of dust on plants generally decreases yield through decreased photosynthesis as a result of less photosynthetic active radiation reaching the plant, by interfering with the gas exchange between the leaf and air, as well as a reduction in leaf stomatal conductance (Smith, 1978; Farmer, 1993; Hirano et al., 1995; Bergin et al., 2001; Zia-Khan et al., 2015). The deposition of inert particles on plant leaves mimics drought stress, leading to stomatal closure and higher leaf temperature (Eller, 1970; Hirano et al., 1995; Zia et al., 2013; Zia-Khan et al., 2015). The net impact of dust deposition on plant productivity depends on leaf type; for example, resinous leaves are more susceptible to dust coating (Farmer, 1993; Sharifi et al., 1999).

Based on these well-documented impacts, dust should have a predominantly negatively affect on yields. Much of the research regarding the effects of dust on plants has been compartmentalized, focusing on the three components of wind erosion independently. A holistic approach to estimating the effect of historical dust events on yields has yet to be undertaken over large regions. A significant hurdle to taking this approach was the lack of adequate dust and yield records. A potentially more problematic issue was the the lack of tools capable of isolating the
actual 'dust effect' from the other climate variables.

Traditional crop modeling techniques could not be applied to answer this question because calibration and validation of process-based models would require extensive data that is not available on the scales of interest. Many statistical crop models could not be applied out-of-the-box because dust not only correlates with other climate variables, but interacts with them (Xi and Sokolik, 2012). However, increasingly transparent tools for exploratory data analysis presented a potential solution to these issues. Random forest (RF) analysis has recently been applied as diagnostic statistical crop models to estimate the drivers of crop yield in sub-Saharan Africa (Hoffman et al., 2017) and in the central US (Chapter 3). In this work, we apply RF as a data mining tool to test whether dust has a detectable effect on sorghum, corn, soybean, and winter wheat yields in the central United States from 1980 to 2016.

To identify the impact of dust on crop yields, we create a comprehensive data set that includes dust, climate, and yield data. Creating this data set has been a significant and meticulous feature of this project and is outlined in Section 4.2 along with the methods used in this work. In Section 4.3, we present an overview of the results, with a discussion and conclusions in Section 4.4 and 4.5, respectively.

4.2 Data & Methods

For this work, we required a comprehensive data set that included dust, climate, and yield data in order to identify the effect of dust on crop yields. Creating this data set was a major component of this research and required incorporating multiple sources and types of data. Using data from the USDA NASS and filtered for irrigation using the Census of Agriculture (USDA NASS, 2012), we estimated rainfed county-level yields. We extracted daily weather data for each county using MetData and station data from Global Historical Climatology Network - Daily (GHCND) (Menne et al., 2012) and Integrated Surface Data Hourly Global (ISD) (Smith et al., 2011) datasets. We computed dust event data using station data from GHCND, ISD, and satellite data from MODIS-Aqua. We define a dust event as any event in which dust is either suspended, emitted, or both.
4.2.1 Yield and irrigation data

This work uses the same yield data as Chapter 3. Though described in Section 3.2.1, we briefly outline the yield data again for clarity. County-level yield data is from the United States Department of Agriculture National Agriculture Statistics Service (USDA NASS). This work, and the work in Chapter 3, specifically focuses on the yield responses of rainfed crops, however because we knew that dust was most prevalent in the High Plains and parts of Texas, we did not omit data west of the 100\textsuperscript{th} meridian due to high prevalence of irrigation as in Schlenker and Roberts (2009) (Figure 4.1). Instead we computed the prevalence of irrigation in each county (by harvested area) and permitted yields from counties if the harvested area in 2012 was less than 25\% irrigated (USDA NASS, 2012). All yields were converted from bushels per acre to kg/ha assuming 13\%, 15.5\%, 13\%, and 13.5\% moisture for sorghum, corn (grain), soybean, and winter wheat, respectively.

4.2.2 Climate data

We extracted and merged multiple sources of climate data for this project. MetData (Abatzoglou, 2011), a gridded weather product from the University of Idaho is combined with two datasets derived from station data. We use data from the Global Historical Climatology Network - Daily (GHCND) (Menne et al., 2012) and the Integrated Surface Data Hourly Global (ISD) (Smith et al., 2011).

We distinguish non-dust climate variables (e.g. temperature and precipitation) from dust for clarity. Dust characterization studies in nearby regions have used diverse sources of data such as satellite (optical depth), station-based visibility, and other records (Okin and Reheis, 2002; Stout and Lee, 2003; Novlan et al., 2007; Ginoux et al., 2012; Neff et al., 2013; Lei and Wang, 2014; Hand et al., 2017). Reconciling results from these data sets is notoriously difficult, so we estimated our own dust metrics from satellite and station data for consistency (Lei and Wang, 2014; Hand et al., 2017). Unfortunately dust data derived from satellite and station data both have unique spatial and temporal coverage, as well as different sources of uncertainty. To test for robust conclusions that are independent of the metric, we used the various dust metrics independently.

Because we suspected that the yield-effect of dust may depend on the growth phase, we developed the algorithm to estimate phases of the growing season in...
Figure 4.1: Region of interest plotting corn yield from 2005 in kg/ha. States without county outlines are not included in the analysis. Counties with missing yield data (white) are due to missing or irrigated yields.

Chapter 3. Dust events that occur when the plant is young are likely smaller in scale and caused by local soil erosion from high winds and dry conditions. This is because limited vegetation results in low threshold friction velocities for dust emission (Shao et al., 1993; Marticorena and Bergametti, 1995; Zender et al., 2003; Hoffman et al., 2014). These dust events may have a significant effect on yield as a result of defoliation and short-term, high-intensity moisture stress. Because leaf and vegetation growth is significant by the end of the establishment phase, high
threshold friction velocities decrease the frequency of local wind erosion events. However, it is still possible for dust events to occur after the establishment phase (Lei and Wang, 2014). Although we suspect that the effect of dust events on yield is largely negative, late-season dust events that increase AOD and the fraction of diffuse radiation could negate the negative effects of wind erosion (Chameides et al., 1999; Xi and Sokolik, 2012; Farmer, 1993). To summarize, dust events can occur during all phases of crop growth, but the characteristics of the event and the effect on the plant may depend on the phase.

4.2.3 Dust event from MODIS data

As one source of dust information, we use the Moderate Resolution Imaging Spectroradiometer (MODIS) instrument aboard the Aqua satellite launched in May of 2002 (Levy and et al., 2015). MODIS provides a massive data set that can be used for measurements of various optical properties like aerosol optical depth (AOD) or used to create a proxy for mineral dust. We provide an overview of the satellite data product and then identify a set of criteria that selects dust events for our analysis. These criteria are based on previous attempts to use MODIS as a dust event proxy.

We used MODIS Level 3 Collection 6, gridded daily atmospheric product MYD08, from July 2002 to December 2016. The Collection 6 atmosphere products over land provide data based on the Deep Blue algorithm to improve satellite retrievals over land, particularly over bright surfaces (Hsu et al., 2004, 2013). Unless otherwise specified, all satellite-derived variables are the Deep Blue products. Level 3 data is mapped on uniform spatial and temporal grids and are typically more complete than Level 2 data. MYD08 D3 (daily, 1°) data are used. The MYD08 D3 product is created using the 10 x 10 km pixel arrays, and each 1° cell of the MYD08 D3 product contains simple statistics (mean, standard deviation, maximum, and minimum), as well as quality control information.

With a pass over time of 1330 LST (ascending node), we used data from MODIS Aqua because it is more likely to capture convective dust events than its sister satellite, Terra which has a pass over time of 1030 LST (descending node). Based on day-time frequency of occurrence climatologies, mesoscale dust events are most likely between 1400-1600, are probable between 1300-1800, and are possible between
1000-2100 (Novlan et al., 2007). Because a pass over is effectively a snapshot of the
day, these data will likely not capture small dust disturbances. It is also possible
that we lose some smaller disturbances as a result of using the 1° data instead of
the 10 km data (Ginoux et al., 2010).

Past efforts have noted that the identification of dust events from satellite
data is exceptionally difficult (Ginoux et al., 2010, 2012; Ciren and Kondragunta,
2014). Using MODIS Deep Blue data, we define a dust event proxy using AOD_{550},
Angstrom exponent (\(\alpha_{412-470}\)), single scattering albedo values (\(\omega_{412}\) and \(\omega_{670}\)), and
standard deviation of AOD_{550}. Mean AOD_{550} is not a proxy for dust events (nor
is maximum AOD_{550}) because the 550 nm band can also capture pollution and
smoke, but the first criteria of dust event detection is that an event must have a
reasonably large AOD (AOD_{550} > 0.25). The Angstrom exponent is computed
using AOD and wavelength and it is meant to provide information on the aerosol
size distribution. Dust particles are generally coarse so we filtered out instances
where the Angstrom exponent for land (\(\alpha_{412-470}\)) is less than 1 (Eck et al., 1999;
Ginoux et al., 2010). Several studies have used smaller threshold values for \(\alpha\),
but these projects sought to filter out ‘old dust’, which we did not (Schepanski
et al., 2007; Ginoux et al., 2010, 2012). Single scattering albedo (\(\omega_{\lambda}\) is the ratio
of scattering efficiency to total extinction efficiency and is related to the composition
of dust (Eck et al., 1999; Moosmüller et al., 2012). While it is unlikely that this area
of the world experiences significant amounts of sea salt and dust simultaneously,
we filter out other coarse-mode particles like sea salt using the single scattering
albedo. Dust absorbs in the visible spectrum, while sea salt does not; therefore we
consider events when \(\omega_{670} > \omega_{412}\) (Ginoux et al., 2010). Finally, dust storms are
often characterized by high spatial variability and can be identified by data with
high standard deviations (Lei and Wang, 2014).

To recap, we consider MODIS data to capture a dust event when it meets the
following criteria:

- \(AOD_{550} > 0.25\)
- \(\omega_{670} > \omega_{412}\)
- \(\alpha_{550} < 1\)
- \(\sigma_{AOD_{550}} > 0.05\)
This simple method of dust event detection filters out many of the pollution events, but we still get false positives from wildfire smoke. We calibrated this method using the events outlined in Lei and Wang (2014) and found that many of the mesoscale and cyclogenesis dust events are captured using these detection criteria. The area that experiences a high number of dust events (as detected by these criteria) align well with those found in other work (Stout and Lee, 2003; Lee et al., 2012; Ginoux et al., 2012). Dust events are aggregated for each growth phase to create a metric of dust event frequency similar to DSF in Engelstaedter et al. (2003) and FoO Ginoux et al. (2012).

4.2.4 Station data

To supplement the satellite-based dust events, we used two different station data sets: the Global Historical Climatology Network - Daily (Menne et al., 2012) and Integrated Surface Data Hourly Global (Smith et al., 2011). Similar to the methodology used in Hoffman et al. (2017) and Chapter 3, we compute the daily weather statistics for each day and county. Next, we introduce the variables extracted from each data set as well as the methodology used to compute the various station-based dust metrics.

4.2.4.1 Global Historical Climatology Network - Daily

We extract GHCND data for the variables in Table 4.1. There are five core variables in these data: precipitation, snow fall, snow depth, and maximum and minimum temperature. We extract the station-level data for each of the five core variables as well as 21 other elements, including data for each 'Weather Type' variable. 'Weather Type' variables were computed as frequencies; if a county reported a certain code on any day, that day was given a value of 1, and 0 otherwise. The GHCND is a database that consists of composite climate records from numerous sources that are subsequently merged and undergo quality control. Daily statistics are computed for each variable, and compiled into a station-specific file. Nearly all of the GHCND observations are in LST (only one source reports observation in UTC and that source is very rarely used as it is 27 out of 27 in terms of priority). In order to extract daily information for each county, we first needed to identify all stations in each county. Then, all stations in a given county were used to
compute county-level variable statistics (mean, minimum, maximum, and standard deviation) as well as quality control flags.

While GHCND has a large number of stations (Figure 4.2), there are significant inconsistencies in reporting that limit the spatial extent of useful data.

Figure 4.2: Station locations for GHCND data (blue plus signs) and ISD data (red diamonds).
Table 4.1: Station-level daily variables extracted from GHCND database. Asterisks denote a variable computed from the GHCND station data. These data were computed for each day and each station, which were then aggregated to form county-level statistics.

<table>
<thead>
<tr>
<th>Variable abbreviation</th>
<th>Variable description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRCP</td>
<td>precipitation (mm)</td>
</tr>
<tr>
<td>SNOW</td>
<td>snowfall (mm)</td>
</tr>
<tr>
<td>TMAX</td>
<td>maximum temperature (°C)</td>
</tr>
<tr>
<td>Tmin</td>
<td>minimum temperature (°C)</td>
</tr>
<tr>
<td>TAVG*</td>
<td>computed average daily temperature (°C)</td>
</tr>
<tr>
<td>ACSC</td>
<td>average cloudiness sunrise to sunset from 30-second ceilometer data (percent)</td>
</tr>
<tr>
<td>ACSH</td>
<td>average cloudiness sunrise to sunset from manual observations (percent)</td>
</tr>
<tr>
<td>AWDR</td>
<td>average daily wind direction (°)</td>
</tr>
<tr>
<td>AWND</td>
<td>average daily wind speed (m/s)</td>
</tr>
<tr>
<td>EVAP</td>
<td>evaporation of water from evaporation pan (mm)</td>
</tr>
<tr>
<td>TSUN</td>
<td>daily percent of possible sunshine (percent)</td>
</tr>
<tr>
<td>WDFG</td>
<td>direction of peak wind gust (°)</td>
</tr>
<tr>
<td>WESD</td>
<td>water equivalent of snow on the ground (mm)</td>
</tr>
<tr>
<td>WSFG</td>
<td>peak gust wind speed (m/s)</td>
</tr>
<tr>
<td>VPD*</td>
<td>vapor pressure deficit computed by taking the difference in saturated vapor pressure at TMIN and TMAX (hPa)</td>
</tr>
<tr>
<td>WT04</td>
<td>weather type = ice pellets, sleet, snow pellets, or small hail</td>
</tr>
<tr>
<td>WT05</td>
<td>weather type = hail (may include small hail)</td>
</tr>
<tr>
<td>WT06</td>
<td>weather type = glaze or rime</td>
</tr>
<tr>
<td>WT07</td>
<td>weather type = dust, volcanic ash, blowing dust, blowing sand, or blowing obstruction</td>
</tr>
<tr>
<td>WT08</td>
<td>weather type = smoke or haze</td>
</tr>
<tr>
<td>WT15</td>
<td>weather type = freezing drizzle</td>
</tr>
<tr>
<td>WT16</td>
<td>weather type = rain (may include freezing rain, drizzle, and freezing drizzle)</td>
</tr>
<tr>
<td>WT17</td>
<td>weather type = freezing rain</td>
</tr>
<tr>
<td>WT18</td>
<td>weather type = snow, snow pellets, snow grain, or ice crystals</td>
</tr>
<tr>
<td>WV07</td>
<td>weather in the vicinity: ash, dust, sand, or other blowing obstruction</td>
</tr>
</tbody>
</table>

4.2.4.2 Integrated Surface Data Hourly Global

We extract ISD data for the variables in Table 4.2. The ISD database requires six mandatory variables to be reported: wind, ceiling height, visibility, air temperature, dew point temperature, and sea level pressure. To be consistent with daily data from GHCND, we computed daily statistics from the hourly (or sub-hourly) data for each of the six mandatory variables. Daily statistics were only computed if the observation contained adequate quality control flags. After identifying all stations within a county, we aggregated these stations to compute the county-level data (Table 4.2). The ISD database consists of global hourly observations from several different sources and is hosted by National Climatic Data Center (NCDC). Observations from ISD stations are occasionally included in the GHCND database.
Table 4.2: Station-level daily variables extracted from the mandatory variables reported in the ISD database.

<table>
<thead>
<tr>
<th>Variable abbreviation</th>
<th>Variable description</th>
</tr>
</thead>
<tbody>
<tr>
<td>wdir</td>
<td>wind direction (°)</td>
</tr>
<tr>
<td>wspd</td>
<td>average windspeed (m/s)</td>
</tr>
<tr>
<td>wspdx</td>
<td>maximum recorded windspeed (m/s)</td>
</tr>
<tr>
<td>ceil</td>
<td>ceiling height (m)</td>
</tr>
<tr>
<td>vis</td>
<td>average visibility distance (m)</td>
</tr>
<tr>
<td>visx</td>
<td>maximum recorded visibility (m)</td>
</tr>
<tr>
<td>visn</td>
<td>minimum recorded visibility (m)</td>
</tr>
<tr>
<td>temp</td>
<td>average daily air temperature (°C)</td>
</tr>
<tr>
<td>tempx</td>
<td>maximum recorded air temperature (°C)</td>
</tr>
<tr>
<td>tempn</td>
<td>minimum recorded air temperature (°C)</td>
</tr>
<tr>
<td>tdew</td>
<td>average daily dew point temperature (°C)</td>
</tr>
<tr>
<td>tdex</td>
<td>maximum recorded dew point temperature (°C)</td>
</tr>
<tr>
<td>tdewn</td>
<td>minimum recorded dew point temperature (°C)</td>
</tr>
<tr>
<td>patm</td>
<td>average sea level pressure (hPa)</td>
</tr>
<tr>
<td>vpd*</td>
<td>vapor pressure deficit computed using an equation (hPa)</td>
</tr>
</tbody>
</table>

4.2.4.3 Dust event identification: station data

In addition to the satellite-based dust metric (MODIS), we also consider three different station-based dust metrics. First, we used the "WT07" code from the GHCND database to identify dust events (Table 4.1). Days in which this code was issued were given a value of 1 and those that did not were given a value of zero. Next, we use wind and visibility data from the ISD database to compute events that met the National Weather Service criteria for a dust event. The NWS issues a Dust Storm Warning when dust causes visibility to reduce to a quarter mile (409 m) with winds greater than 25 miles per hour (11.1 m/s) and a Dust Storm Advisory when dust reduces visibility between a quarter mile and 1 standard mile with winds greater than 25 miles per hour. Days that met these conditions were given a value of 1 and those that did not were given a value of 0. For each metric, events are aggregated for each growth phase to create a metric of dust event frequency.
4.2.5 Statistical crop model

As in the previous chapters, we use random forest (RF) as the diagnostic statistical crop model. Chapter 2 and 3 validate the use of this model to identify the phase-specific functional form between crop yields and climate predictors. This project included dust as an additional set of potential predictors to evaluate the influence of dust on crop yields.

4.2.6 Growing season phases

4.2.6.1 Methods to compute growth phases

Using the same methodology and data as Chapter 3, we estimate the planting date and phase lengths for the summer crops (Section 3.2.3). To recap, planting dates are estimated for each year based on county-level temperatures because its permits adaptation, unlike stationary planting dates (as in Sacks et al. (2010)). We estimate that planting occurs once the 21-day moving average rises to a crop-specific threshold temperature (Table 3.2). We also implemented a frost restriction for sorghum and soybean planting dates because of their relatively low tolerance for cold temperatures. When a frost restriction is imposed, the planting date occurs when the planting threshold is reached only after the last early frost has occurred.

We partitioned the growing season for each crop into three distinct growth phases: establishment, critical window, and grain filling. For the summer crops, the length of these phases is determined by thermal times and crop-specific base temperatures (Table 3.2). The establishment phase is meant to capture conditions during germination, emergence, and early vegetative stages. The critical window aims to capture the reproductive phases that determine yield potential, and the grain filling phase captures the late-season conditions important for grain growth. The planting date algorithm was designed to allow for year to year adaptation, and the growing season algorithm was designed to allow for early termination of the grain filling phase under extreme conditions. We consider harvest for all crops to occur once grain filling is completed, and losses of yield post-maturity are not accounted for.

The growth phases for winter wheat were computed differently. Instead of planting when the 21-day moving average temperature exceeded a crop-specific
Table 4.3: Phase-specific information used to compute the growth phases for winter wheat. $T_{\text{start}}$ denotes the variable and corresponding value for each phase. $T_{\text{start}}$ is accompanied by an arrow to indicate whether the threshold is computed for falling (↓) or rising (↑) temperatures. The phase-specific length and base temperature are highlighted in the Duration and $T_{\text{base}}$ columns.

<table>
<thead>
<tr>
<th>Phase</th>
<th>$T_{\text{start}}$</th>
<th>Duration</th>
<th>$T_{\text{base}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Establishment</td>
<td>$T_{\text{plant}} = 20{}^\circ\text{C} \downarrow$</td>
<td>-</td>
<td>$4{}^\circ\text{C}$</td>
</tr>
<tr>
<td>Critical Window</td>
<td>-</td>
<td>400°C d</td>
<td>$4{}^\circ\text{C}$</td>
</tr>
<tr>
<td>Grain Filling</td>
<td>$T_{\text{heading}} = 19.5{}^\circ\text{C} \uparrow$</td>
<td>500°C d</td>
<td>$7{}^\circ\text{C}$</td>
</tr>
</tbody>
</table>

threshold as we did for summer crops, winter wheat is planted when the 21-day moving averages falls below 20°C (Table 4.3). In order to ensure fall or early winter planting dates, we only considered days after the peak summer temperature. Of course there is variation in this temperature threshold, but we chose a single temperature for simplicity as we did in Chapter 3.

Unlike summer crops, the length of the establishment phase for winter wheat is not uniform and therefore the same methodology cannot be applied. Instead, we estimated the heading date for winter wheat, which marks the beginning of the grain filling phase, and worked forwards and backwards from there. The heading date is estimated when the 10-day moving temperature average reaches 19.5°C (calibrated using Hu et al. (2005)). To estimate the start of the critical window, growing degree days are computed "backwards" from the heading date. Using the phase-specific values for the base temperature and thermal time for this phase (Table 4.3, the critical window starts when 400 °C d can be achieved by the end of the phase. In the event that the 10-day rolling average drops below the base temperature, we began the critical window immediately after that instance (and throw a flag). The remaining period between the beginning of the critical window and the planting date is considered the establishment period. The duration of the grain filling phase in winter wheat is relatively consistent, and once 500 °C d have accumulated after the heading date, we consider harvest to occur.

In many of the colder states, these criteria may result in an overly long establishment phase (Figure 4.3), but start of the critical window and grain filling phases are reasonable (Hu et al., 2005). To account for this long establishment phase when computing variable-specific averages, we omit days in which the average temperature drops below 4°C because the plant is not actively growing. However,
we did not omit any days of data for precipitation and snow, since soil water storage is critical for this crop. Additionally, we did not omit any days for of the dust variables because a plant does not necessarily need to be actively growing to be affected by wind erosion.

Figure 4.3: Segment plot depicting average growing season for winter wheat. Blue represents the establishment phase, green represents the critical window, red represents the grain filling phase, while black represents a two-week drying period (not included in this study).
4.2.6.2 Data used to compute growth phases

As in Chapter 3, we use MetData to compute the dates of the growing season and length of individual growth phases. Though station data was available to compute the growing season, we elected to use MetData because of the spatial and temporal inconsistencies in station data. We describe several reasons to justify this decision. First, using station data to compute the growing season would eliminate the any counties without a station from this analysis. MetData is not limited by station location nor does it exhibit the same reporting inconsistencies as station data (Figure S4.1).

Second, Figure S4.1 indicates that temperature data is available from most of the GHCND stations which have relatively broad spatial coverage. While spatial coverage is decent, these data have a number of temporal inconsistencies that led to use MetData when computing growth phases. (1) The number of stations has consistently increased with time and stations are decommissioned frequently, resulting in averages derived from a different spatial subset within a county every year (Menne et al., 2012). (2) Though GHCND reports daily statistics, missing data is prevalent and not every day records data.

Third, the magnitude of station-based county averages are similar to those derived using MetData, so we are likely not obscuring important data. However, we did find that station data often contain a wider range of county averages when compared to MetData. More extreme values occur since a given county may only have one station in the coldest/warmest part of that county, compared to an evenly-spaced 4 km grid within a county using MetData. If a station is located in the coolest/warmest part of a county, its average may not provide an accurate representation of the conditions experienced by crops. Because station location varies from county to county, there is no consistent trend among MetData, ISD, or GHCND (Figure S4.2). In other words, station data is randomly biased as a result of the station location(s) in a county. As a result of these issues, we use MetData to calculate the phase-lengths and growing season bookends.

4.2.7 Experimental Design

Now that we have all of the data, we briefly explain some of the decisions used to create the final suite of models and the data used therein. First, MetData
Table 4.4: Comparison of explained model variance using all variables for each data set in a crop-specific regression. MetData model contains all variables listed in Table 3.1, including year. GHCND model contains all variables listed in Table 4.1, and ISD model contains all variables listed in Table 4.2.

<table>
<thead>
<tr>
<th>Explained variance ($R^2$)</th>
<th>sorghum</th>
<th>corn</th>
<th>soybean</th>
<th>wheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>MetData</td>
<td>0.73</td>
<td>0.86</td>
<td>0.80</td>
<td>0.82</td>
</tr>
<tr>
<td>GHCND</td>
<td>0.46</td>
<td>0.64</td>
<td>0.47</td>
<td>0.41</td>
</tr>
<tr>
<td>ISD</td>
<td>0.43</td>
<td>0.58</td>
<td>0.49</td>
<td>0.48</td>
</tr>
</tbody>
</table>

was used as climate data in the statistical crop models in lieu of station data because the inconsistency and randomly biased county-level values from station data affected the crop models. Model performance using only station data was poor when compared with the values from analogous MetData models (Table 4.4). There were two possible reasons for this poor model performance: (1) the subset for which station data is available was not representative of the entire region, or (2) the random bias introduced by using the station data to compute county-level averages did not adequately represent growing conditions. To test whether the subset or the bias was causing the change, we tested the station data subset using MetData in lieu of station-derived values. This investigation produced results very similar to the MetData models, which suggested that the spatial subset is not causing the problem. It appears as though the climate variables computed from station data are not representative of average growing conditions. Therefore, the remainder of this work used models with MetData as climate data in lieu of station data.

Second, because each dust metric has its own set of uncertainties as well as spatial and temporal variation, we test each metric independently. The extent of spatial variation in the sampled data for each of the dust metrics can be seen in Figure 4.4. The temporal variation between satellite and station-based dust metrics is also significant because satellite data begins in 2002, thus limiting the range of data we can use.

Finally, for each crop, we use the fundamental variables, a single dust metric, and time (year) as predictors (Figure 4.5). We also tested whether the inclusion of mean AOD as a predictor affected model performance and/or yield response to individual variables. Therefore each crop has eight basis models: four dust metrics with and without AOD as a predictor. Every model includes all three growth phases
Figure 4.4: Log plot of the total number of dust events during the growing seasons of corn from 2002 to 2016. (top left) MODIS (top right) GHCND - WT07 reports (bottom left) ISD - NWS Dust Storm Warning (bottom right) ISD - NWS Blowing Dust Advisory.

and the growing season values for all climate predictors and dust over the entire region (except for spatial analysis - Figure 4.5). Because of the varying spatial and temporal extent of each model, we designed a set of experiments that allowed us to utilize as much data as possible.

Figure 4.5 also includes the systematic tests used to detect whether dust affects historic crop yields in the United States. The first group of experiments tests the
Figure 4.5: Experimental structure used to check for a detectable effect of dust events on crop yields.

sensitivity of the model and results to parameters in the random forest. The second and third set of experiments address phase-specific and residual analysis. The final set of experiments tests for the effect of dust on yields on a smaller scale through by partitioning the region into smaller spatial subsets.

Moving forward into the Results section, with four crops, eight basis models, and ten experiments outlined in Figure 4.5, including the results for each individual model is excessive and unnecessary. This is primarily because we found very little variation between models and experiments. If our results identify a unique model or response, we explicitly include those results, otherwise we include a result that is representative of the experiment.

4.3 Results

For the eight basis models for each crop, our initial results reveal that dust event frequency variables are consistently the least important predictors in every model regardless of dust metric (Tables S4.2-S4.5). Following the format of Chapter 3, we have included a variable importance plot for corn using WT07 as an example.
Table 4.5: A representative example of change in R² imparted by removing the dust predictor from a model. In this example, we used the core model variables, the MODIS dust metric and included AOD.

<table>
<thead>
<tr>
<th></th>
<th>Sorghum</th>
<th>Corn</th>
<th>Soybean</th>
<th>Wheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>R² (dust + AOD)</td>
<td>0.711</td>
<td>0.855</td>
<td>0.788</td>
<td>0.819</td>
</tr>
<tr>
<td>R² (AOD)</td>
<td>0.708</td>
<td>0.856</td>
<td>0.787</td>
<td>0.818</td>
</tr>
</tbody>
</table>

(Figure 4.6). Including this figure for other models is uninformative, because there is minimal variation in the importance of dust. The importance of each variable is evaluated using the permutation importance metric, which is calculated as the mean decrease in accuracy as a result of permuting a predictor over all trees. Raw, unscaled values of this metric are considered to be most reliable (Díaz-Uriarte and de Andrés, 2007; Strobl et al., 2007, 2008).

To determine whether the effect of dust was more subtle than we expected, we tested whether removal of the dust metric affected model performance or the yield response to any of the other predictors. We found little impact on model performance or on modeled response curves for the non-dust variables (Table 4.5). These changes were statistically insignificant. We also inspected the yield response to different variables for dust interactions, but we found no interactions. The correlation between AOD and the MODIS dust metric was identified in this inspection, but this was expected.

This negative result led us to question whether we can detect expected dust impact in smaller targeted analyses. Next, we systematically address the robustness of this result through four sets of experiments that focus on (1) the sensitivity of the result to parameters in the random forest, (2) phase-specific analysis, (3) residual analysis, and (4) spatial analysis (Figure 4.5).

4.3.0.1 Sensitivity to RF parameters

Although random forest does not have many tunable parameters, we examine those that may have had an effect on these results. To lower the correlation between trees, not all predictors are used for splitting at each node in every tree of the RF; this parameter is called mtry. A subset is chosen to reduce variance by forcing the model to assess the importance of all the variables in the model, including the weak
predictors. We used the default value for $m_{\text{try}}$; with four phases, six fundamental variables, time, dust, and AOD as predictors $m_{\text{try}}$ was typically equal to ten or eleven. While each predictor has an equal chance of being selected at each node to be included in the subset, there is no guarantee that each predictor will be chosen at the split. Based on the variable importance rankings of each of the dust metrics in Figure 4.6, it was very rare that dust provided the best splitting rule. By reducing $m_{\text{try}}$ at each split, we artificially increased the probability of the model to choose the dust metric more often. By giving the model a smaller subset to choose from, it statistically gives dust a higher probability of being the most informative splitting rule. By reducing $m_{\text{try}}$ in our models, we found some variation in the order of the most important predictors, but there was no overall effect on the importance of dust. It is possible that effect of dust on yields is very small, and the signal cannot be distinguished from noise in the basis models. RF regression typically terminates tree growth when the terminal node size is equal to five, this parameter is called nodeSize. When nodeSize is set to increasingly small values, it effectively forces the model to look through the noisy data to find patterns. Setting nodeSize $= 3$ did not increase the importance of any of the dust metrics.

A well-known bias of RF is its tendency to choose categorical predictors over continuous ones (Strobl et al., 2007, 2008; Boulesteix et al., 2011). Albeit slightly unorthodox, we tested dust as a categorical variable in an attempt to artificially inflate its importance. Instead of considering dust event frequency as a continuous variable, the categorical variable measured 'dust' or 'no dust' during each phase. This was tested for all eight basis models for each crop and we found no significant change in variable importance or model performance.

Splitting criteria in RF preferentially select predictors with more categories than predictors with fewer categories, regardless of their association with the response (Strobl et al., 2007, 2008; Boulesteix et al., 2011). Though dust event frequency is specified as a continuous variable, the nature of frequency data (i.e. discrete integer values) may cause the data to function as a categorical variable. Therefore we tested a separate model that is similar to RF, but uses conditional inference trees instead of classification and regression trees. Compared to the splitting criterion used in RF, trees based on conditional hypothesis testing are split based on each candidate predictor’s association with response. At each split, a p-value is computed for each candidate predictor and that value is used to split the data (Hothorn et al., 2013).
The p-value is conditional, meaning that it reflects the probability to obtain a higher association with the response. By using the p-value to split the data, these trees are not affected by the variable selection bias which preferentially selects predictors with many categories. Conditional inference trees effectively fix the variable importance bias responsible for preferentially choosing correlated predictors in RF (Strobl et al., 2008). To test this, we use the function \texttt{cforest} in the R add-on package \texttt{party}, which implements a conditional random forest. For each crop, we tested the eight basis models and found no change in the importance of dust, meaning it does not have low importance because of the number of unique values, but because of its low association with yield.

For each crop, we tested whether the results were sensitive to the RF parameters ($m_{\text{try}}$ or \texttt{nodeSize}) or splitting criteria (categorical variable or conditional inference trees). We found that the variable importance of dust does not vary with $m_{\text{try}}$ or \texttt{nodeSize}, nor does it vary when the variable was converted to a categorical variable or when a conditional random forest was used instead of a classical random forest.

### 4.3.0.2 Phase-specific analysis

For our next test, we considered whether simultaneously using all three growth stages and growing season values in the models inhibited our ability to detect the impact of dust on crop yields. When predictors are correlated in RF, as the individual growth phases of each predictor are, the importance will be divided among the correlated variables. To test whether we are making it difficult to detect by using correlated dust metrics, we considered two different experiments. The first included only one phase of growth for all predictors, including dust. We tested the results of this experiment for all three growth phases and for the growing season (for each crop and dust metric). We found that dust was still the least important metric in every case. The second experiments included all growth phases for every variable except dust, which only contained one phase. We tested this experiment for all three growth phases and the growing season. Though slightly unorthodox, this experimental design was an attempt to artificially increase the importance of dust relative to the other predictors. Again, there was no significant change in the importance of the dust metric for any of these cases.
4.3.0.3 Residual analysis

We also tested whether the effect of dust on yields was embedded in the residuals of the model without dust. To test this, we created a two-stage regression with the RF tool. First we computed the residuals from models with time and fundamental variables with and without AOD. Second, we used these residuals as the response variable and dust as the only predictor. Because RF with one predictor is uncommon, we also tested a linear regression as well as RF (Table S4.6). We found no significant relation between the residuals of the fundamental variable model and dust event frequency (Figures S4.9-S4.12).

4.3.0.4 Spatial analysis

Because this study covers 18 states and 1631 counties therein, we tested whether the effect of dust on yields will emerge by excluding counties that do not experience any dust events. For each of the dust metrics, and their respective temporal range, we remove all the counties that do not experience any dust events. However, in each of the eight basis models for each crop using this new spatial extent, dust was still the least important variable (Table S4.7) and did not affect yields in any significant manner.

In every experiment, except for two, dust is the least important metric with each of the phases ranking between one and four (Table S4.7). For winter wheat, the model containing the MODIS dust metric without AOD ranks dust events accumulated over the growing season as seven (still very low, but not in the bottom four). We investigated this further and found that because the MODIS metric is computed using AOD, it is correlated to AOD. As a result, the dust metric is representing the variability that was previously explained by AOD. This is a well known bias in the variable importance metric used in RF (Strobl et al., 2008).

Similar to the concept of removing counties without any dust events, it is possible that the effect of dust on yields may be obscured by aggregating over such a large area. Therefore, we also performed this analysis for individual states. We found no effect on variable importance of dust for any of the station-based metrics. However, yields in several states indicated a response to the satellite-based dust metric. In general, we found that dust events tend to have a negative effect on crop yields.
In this state-by-state analysis, we define a variable-phase combination as significant if it is above or equal to rank 10. Sorghum yields in Kansas and Nebraska are negatively affected by increasing dust events primarily during grain filling (Figure 4.7). Corn yields are negatively affected by dust events primarily during the critical window in North and South Dakota. Soybean is negatively affected by dust events over the growing season in North Dakota, but have a larger effect on yields in South Dakota when events occur during grain filling. The effects of dust on soybean are minimal with partial dependence plots attributing no more than 100 kg/ha yield decreases to dust events. Winter wheat yields are negatively affected by increasing dust events over the establishment phase in North Dakota, though these are also minimal. We found no significant impact on the variable importance of dust in any of the other 14 states, even those in which dust events occur frequently (Figure 4.4).

Partial dependence plots indicate a rapid decrease in yields as a result of increasing dust events. In Nebraska, Kansas, and South Dakota is appears that even one dust event is associated with a decrease in average yield. In Chapter 3 we discussed how the edges of the partial dependence plots are not necessarily robust due to noisy data caused by limited data. However, the distribution of dust events predominantly consisted of zeros, so the signal on the low end of the x-axis in Figure 4.7 is robust.

We note that the models consistently identified a phase with a negative yield response (to increasing dust events) to be more informative than those with a positive one. We have made a conscious decision to omit partial dependence plots until now because variables with extremely low importance have crude, unstable partial dependence plots (Figure S4.13).

Despite having found a weak dust signal, it is important to note the caveats of these results. By using MODIS data and state-specific yields, we severely limited the amount of data used in each regression. For instance, using the MODIS data and corn yields, each basis model consisted of 9,984 observations, but the corresponding values for North and South Dakota were 372 and 636, respectively. Though RF is a great model to use with limited observations and a large number of predictors, these results are much less certain than those derived using 10 to 100 times more data. Another significant caveat is the instability of these results. For instance, the functional form of accumulated precipitation over the growing season remains the
same when tested on different spatial and temporal subsets (including individual states), indicating that the yield response is stable and robust. The same conclusion cannot be made with the yield response to dust because it is only detected in seven instances. When a response is stable, it usually does not disappear, which casts a veil of caution on these results.

To address the stability of these results, we assess the effect of aggregating state data. For instance, sorghum yields responded to dust in Kansas and Nebraska. Aggregating the data from these two states and re-running the model, resulted in increased importance of dust events during grain filling and the yield response remained distinctly negative. In fact, aggregating data from five 'dusty' states still indicated that dust events had a potential effect on yields (as their importances were below 10, but not in the bottom four). Including data from all eight states with significant dust events caused the variable importance of dust to drop back down to the lowest four predictors. We found a response of corn to dust during the critical window in North and South Dakota, a response which is maintained when the two states are aggregated. When grouped these two states, dust events during the critical window emerge as the third most important predictor and suggest a rapid decrease in yields with increasing event frequency. However when Kansas, Nebraska, and Colorado are included, the importance decreases back down to the final four. This analysis identified that the number of dust events over the growing season was important to predicting soybean yields in North Dakota, but the grain filling phase was more important in South Dakota. When these states are merged with Nebraska the grain filling phase maintains significance as a predictor (rank 17). Once aggregated with the other seven dusty states, the variable drops the importance down to rank four. We found winter wheat yields are sensitive to dust events during the establishment in North Dakota. As soon as it was aggregated with South Dakota, the importance dropped to the bottom four again. This immediate response to aggregation suggests that the yield response to dust in North Dakota is unstable.

Though this state-by-state analysis identified an effect of dust on yields with a consistent response, it was only found eight times. When the effect was found, we cannot identify whether it is an artifact of breaking up the data by state or if it is a true effect that gets muddled in with the general variance when aggregating more states. While more research is necessary, this provides a direction for additional
analyses.

4.3.1 AOD effect on yields

Inclusion of AOD as a predictor did not improve the model performance for any crop, but it likely influenced sorghum and winter wheat yields between 2002 and 2016 (Tables S4.2 and S4.5). In general, crops responded positively to increased AOD, and given that AOD functions as a proxy for diffuse radiation, this result was expected (Figure 4.8). On overcast days, increasing AOD on has a negative effect on net primary productivity (Cohan et al., 2002; Greenwald et al., 2006), but this was not captured with our model.
Figure 4.6: Variable importance plot for corn using WT07 (GHCND) as the dust metric. Variables above the dashed line are considered important, as they are in the top third of the variables. The model $R^2$ is also included in the bottom left. Note that the exact order of the variables is not necessarily informative, changing different parameters of the model (e.g. $m_{try}$) will shift the metrics around slightly.
Figure 4.7: Partial dependence plot for dust event frequency (derived from MODIS) in the phase identified by the model as being most important in the state-by-state analysis. Y-axis represents average yield in kg/ha. All values are plotted (not the innermost 95th percentile).
Figure 4.8: Partial dependence plot for AOD for each growth phases, including the growing season. Colors represent the growth phase: establishment (blue), critical window (green), grain filling (red), growing season (black). Responses from the 2.5-97.5th quartiles are plotted. These crop models include the MODIS dust metric, but the plots for AOD do not change if the metric is changed or removed.
4.4 Discussion

In general, we were unable to detect a significant effect of dust event frequency on crop yields over the full region of interest shown in Figure 4.1. Using the satellite-derived dust metric and partitioning the data by state, we detected a negative yield effect of increasing dust event frequency in four states, though we cannot identify whether detection occurred due to an actual yield effect or an artifact of partitioning the data. For completeness, we discuss the potential dust effect below.

4.4.1 Potential yield impact of dust events

In general, dust events appear to reduce sorghum yields when they occur during grain filling, corn during the critical window, and soybean and wheat during the grain filling and/or full growing season. Results suggest that dust events during the establishment phase have no effect on yields for any summer crops. We originally suspected that winter wheat was particularly at risk from early-season dust events caused by wind erosion because plants are young and vulnerable when vegetation is low and dust events are frequent. We found a weak negative response of winter wheat to dust events during the establishment phase in North Dakota, but the effect was very small (approximately 40 kg/ha). While our original hypothesis suggested that the effect would be negative, we had not considered the resiliency of winter wheat during this phase. Winter wheat can withstand grazing during establishment without a reduction in yield as a result of heavy tillering (Winter and Thompson, 1986), and likely withstands dust events for the same reason. It is important to note that the winter wheat model in North Dakota only contains 303 observations and a model R² equal to 0.5 (Table S4.8). As a result, these wheat results should be treated with caution.

A reduction in yield as a result of dust events during grain filling can be explained by several factors. A reduction in yield may be caused by reduced plant activity as a result of defoliation and damaged photosynthetic leaf tissue from dust emission caused by wind erosion (Woodruff, 1956; Armbrust, 1968; Michels et al., 1995; Armbrust and Retta, 2000). Defoliation during grain filling is known to reduce yields (Echarte et al., 2006). Leaf damage also causes increased respiration and high-
intensity moisture stress due to impaired stoma control which may lead to decreased yields from moisture stress during grain filling (Armbrust, 1984; Fryrear et al., 1975). Provided that deposited dust is not washed off leaf surfaces via precipitation, dust deposition not only reduces the amount of direct radiation reaching the leaf, but it also increases the canopy and leaf temperature. The average increase in leaf temperature from dust deposition ranges between 1.7° and 4.1°C (Eller, 1970; Hirano et al., 1995; Zia-Khan et al., 2015). Results from Chapter 3 clearly identified temperature thresholds during grain filling for TMAX and TMIN in sorghum, corn and soybean yields; corn yields also decreased linearly when critical window TMAX increased (Figure 3.4). Dust may have a pronounced effect on yields during these phases because of the temperature effect. The original effect of dust may be too small to detect, but if it pushes the plant over a temperature threshold, which has a large effect on yields, the effect may become large enough for us to detect.

Though we primarily focused on the variable importance of dust, we note that the shape of the yield response to dust event frequency was not affected by the inclusion or exclusion of optical depth. Therefore, even if AOD has a positive effect on yields due to the increase in diffuse radiation as a result of suspended dust (Chameides et al., 1999; Greenwald et al., 2006; Xi and Sokolik, 2012), it is likely that the negative effects of dust at the surface dominate the signal. This is primarily because the lifetime of dust after a dust event is typically about 2.3 days in North America (Tanaka and Chiba, 2006). Considering an average growing season of most summer crops at 120 days, and wheat much longer, a period of high diffuse radiation for 2-3 days will likely not affect yields in a significant manner.

### 4.4.2 Inconclusive results

Aside from the aforementioned "blips on the radar", the results of this study were largely inconclusive. We detected no yield response to dust events over the entire region of interest. This result was not sensitive to crop, dust metric, random forest parameters, or phase inclusion in the model. We also tested whether the dust effect was related to the residuals of models without dust. The lack of detection of any yield response to dust was also not sensitive to the exclusion of counties without any dust events. Finally, even in the state-by-state breakdowns, no effect of dust on yield was found using any station-based metrics. We were unable to detect an
effect of dust on crop yields with any of the station-based metrics. However, there are both physical and statistical explanations as to why our results were largely inconclusive.

Detecting the effect of dust on yields was challenging because of the seasonality of dust and the differences in dust metrics. There is a clear discrepancy between the seasonality of events captured by satellite and those by station. Station-based dust records, particularly those based on wind and visibility (NWS Warning and NWS Advisory), indicate more frequent dust events during winter, while the satellite-based dust records indicate a peak in early spring followed by a smaller secondary peak in summer (Figure 4.9). In spite of this discrepancy, our satellite and station records agree with those for AOD and PM2.5 seasonality found in other work over this region (Li et al., 2015). The discrepancy between surface dust event detection by stations and satellite dust event detection is likely caused by seasonal changes in vertical aerosol distribution as a result of a seasonal varying mixing height which results in a particularly strong disconnect between PM2.5 and AOD in winter (i.e. low AOD, but high PM2.5) (Li et al., 2015).

We also may not have observed a dust signal because the growing season for summer crops does not occur when high-wind, low-visibility dust events are frequent (as detected by station data). We expected a pronounced yield response from winter wheat because its growing season overlaps with the peaks in station-derived events, but were unable to detect one with any dust metric over the entire region. The state-by-state analysis highlighted a potential yield impact from dust events (as detected by satellite) in North Dakota.

If a major wind erosion event occurred during the establishment phase it is possible that the crop sustained sufficient damage to require replanting (Fryrear, 1973; Fryrear et al., 1975; Baker, 2007; Zobeck and Van Pelt, 2014). If replanting occurred, it would likely make up for much of the damage and obscure the signal of interest, making it difficult to detect. Additionally, if damaged crops were not replanted and resulted in a total crop failure, a zero-yield may not be included in the harvest survey used by the USDA NASS to compute the county-level yields used in this study.

The majority of dust events occur west of the 100th meridian, a heavily-irrigated agricultural region. We have explicitly omitted counties that reported over 25% irrigation by (harvest) area, to ensure that we capture the rainfed signals. However,
it is possible that the dust signal is already being captured by precipitation. Additionally, winter wheat recovers from abrasive injury caused by dust emission remarkably well when water was applied after the events (Woodruff, 1956). An area of future work is to perform this experiment with irrigated yields to see if a dust effect emerges.

Lastly, a simple, albeit unsatisfying explanation is that there is no signal to detect on this scale. If plants are damaged, the variation in yield may be absorbed and explained by other climate predictors. If a plant is not significantly damaged by a dust event, it may recover. Plants are bred to be resilient, and its possible that dust does not pose a significant threat under these conditions because of the short duration of the stress and the relatively long growing seasons.

From a more statistical perspective, it may be that dust events simply do not occur frequently enough to result in statistically significant yield effects. It may only be possible to detect yield effects from dust events using fine-scale data (i.e. state-scale or smaller).

4.4.3 Potential sources of error in methodology or data

We acknowledge several sources of potential error that could contribute to the inconclusive results of this study. The simple algorithm used to estimate planting and growth phases in each year and county uses only temperature to inform the decision. In reality, a farmer uses many other metrics to inform their decision to plant, like precipitation and local knowledge (Waha et al., 2012). A single temperature threshold is used over the entire region of interest despite some latitudinal variation in this value, which could also affect the accuracy of these phases (Yang et al., 2017). However, other statistical crop modeling studies have found their results to be largely insensitive to the range of the growing season (Lobell et al., 2008). Though this is a potential source of error, our conclusion is robust enough to suggest that this error has a small effect, if any.

MODIS-based dust event detection criteria may be too lax. As with many other satellite-based dust detection efforts, wildfire emissions are a significant source of noise in identifying dust events (Lei and Wang, 2014). Wildfire records could be used to filter these events out. However, we tested several variations of the dust detection criteria and found minimal sensitivity in the model (see Section 4.6.2). It
is also possible that important information is lost because we use 1° MODIS data instead of the 10 km pixels arrays (Ginoux et al., 2010, 2012).

As discussed in Section 4.2.6.2, stations in a given county are not necessarily representative of the average growing conditions for that county (Section 4.2.6.2). Therefore, using dust data from these stations may be representative of conditions experienced by crops, however we operate under the assumption that some data is better than no data.

The number of stations included in the GHCND dataset has increased with time (Menne et al., 2012). For consistently reported variables, this implies increasing coverage. However, this is not the case for reports of WT07 in the US. From 1980 to 2007, the number of stations reporting WT07 is relatively consistent (Figure S4.3). The number of stations reporting WT07 from 2008 and 2010 increases by about 50 stations, followed by a dramatic drop off the number of reporting stations. The associated increase in WT07 reports from 2008-2010 is disproportionately large. Our conclusion does not change when these years are excluded from the analysis (Table S4.1).

In contrast to the issue with WT07 reporting, there is a slight bias in the dust metrics derived from ISD data as well. The number of stations included in the ISD dataset has increased with time and so have the wind and visibility observations. This introduces a slight bias into our data by allowing more dust events to occur in recent years. Though this trend exists in the underlying data, it is important to remember that a station has to lie in a county that (1) meets the conditions for planting a crop and (2) must co-occur with a reported yield in order to be included in our analysis.

Another potential source of error in the ISD data, is that the criteria used to compute the NWS metrics are based on wind and visibility, but optical visibility sensors cannot distinguish between blowing dust, snow, rain, fog, smoke, etc (Lee et al., 2012).
Figure 4.9: Average number of dust events per day over the region of interest and time series. Vertical grey lines are included every 50 days. Black line represents the centered 30-day rolling average. This represents the annual cycle of dust events. Colored segments from the top down denote the average growing season over the region for sorghum, corn, soybean, and winter wheat, respectively. The first (blue) segment denotes the establishment phase, the second (green) denotes the critical window, the third (red) denotes grain filling, and the final short black segment represents drying, which can occur before or after harvest.
4.5 Conclusion:

In this paper, we investigated whether the effect of dust on crop yields in the central US could be detected. We estimated dust event frequency using four different methods from three different sources and merged these data with daily weather information from 1980 to 2016. Using random forest as a diagnostic statistical crop model, we were unable to detect an effect of dust on sorghum, corn, soybean and winter wheat yields from the central US. The ability to detect a dust effect on yields was not sensitive to the random forest parameters, inclusion or exclusion of optical depth, or phase-specific analysis. We also did not detect the effect of dust on the residuals of the models without dust. Only when we partitioned the data by individual states, were we able to detect a possible signal of dust on crop yields. In seven instances in four states (North and South Dakota, Nebraska, and Kansas) the model identified dust as an important predictor. In all seven instances yields were negatively affected by increasing dust events. The summer crops, sorghum, corn, and soybean were typically only affected when the dust events occurred in the later growth stages (i.e. not the establishment phase). We suspect that a yield response to dust was detected during this phase as a result of dust deposition on leaves increasing leaf temperature. It is warmest during the grain filling phase, and the results of Chapter 3 identified a temperature threshold response during grain filling for both TMAX and TMIN. If the increase in leaf temperature as a result of dust deposition occurred during grain filling, the threshold temperature of the plant may be exceeded, and result in amplification of the dust effect, allowing the RF to detect it. However, instances in which dust affected yields disappeared when more than five states were aggregated, indicating that it is not a stable response, which casts a veil of caution on these results. More research is necessary to discern whether these instances of detection are an artifact of breaking up the data by state, or if it is a true yield effect that gets mixed in with the general model variance when aggregating more states. If it is a true response, it would imply that yields in the northern High Plains are potentially more vulnerable to changes in dust event frequency.

Though we detected the negative effects of dust events on crop yields in several states, the results of this study are largely inconclusive as we were unable to detect a stable signal of dust on crop yields. This conclusion should not be interpreted to
mean that dust has no affect on yields, but rather that it cannot yet be detected
given the available data. This conclusion should also not be considered a blanket
statement. The central US has significantly lower dust loadings and fewer dust
events than many countries in sub-Saharan Africa. If the data quality in that region
of the world becomes more reliable, this experiment should be repeated.

Though the results of this experiment were inconclusive on our scale of interest,
it was a worthwhile scientific inquiry as it has highlighted more questions and a
path forward.
4.6 Supplement

Figure S4.1: Maximum temperature averaged from 1980 to 2016 for MetData, GHCND, and ISD. Scale units are °C.

4.6.1 Characteristics of dust storms in this region

The most common type of dust event in the western US is triggered by mesoscale and small-scale systems that generate high winds (Lei and Wang, 2014). However, dust storms caused by fronts are typically much more intense; these events typically
occur in the central west plain region (Novlan et al., 2007; Lei and Wang, 2014). Long-lasting dust events can also be caused by cyclogenesis, but these events are typically difficult to detect with satellites and are associated with low surface dust concentrations (Lei and Wang, 2014).

### 4.6.2 MODIS Methodology

Though we use the criteria outlined in Section 4.2.3, we analyzed multiple variations. We tested the detection (and model performance) without the single scattering albedo condition and by decreasing the threshold for the Angstrom exponent to (a conservative) 0.5. We find that removing the single scattering albedo condition likely allows too many non-dust events, particularly those in Arkansas, Indiana, and Ohio. The number of these events over the growing season decrease when the condition is in place. Tightening the condition on the Angstrom exponent almost removes these events completely and highlights a particularly dusty region in west Texas, the panhandle of Oklahoma, and west/central Kansas. The impact on the regression model did not vary significantly depending on criteria modifications.

### 4.6.3 Station Methodology

ISD data were not as simple to extract as the GHCND data. Each station reported data in UTC and therefore had to be converted to LST in order to ensure accurate, and comparable, daily statistics.

We also considered the "WV07" variable, but it was reported so infrequently that it was not useful. Engelstaedter et al. (2003) defined a dust storm as "an event in which visibility was reduced to <1km due to the presence of dust". We computed events in which visibility dropped below 1km, but we cannot accurately attribute the reduction in visibility to the presence of dust. Therefore, we omitted that variable.

### 4.6.4 Experimental set up

In all the experiments included in the manuscript, we merged one dust metric with the fundamental variables from MetData. However, we did test whether dust became more important when only using station data. We tested this for the
GHCND and ISD data, and found no strong effect. Dust was still in the noise of the model.

Another reason we elected to use MetData in the model in lieu of station data is because the fundamental variables were not available on a large enough scale. Specifically, there was a lack of radiation observation. The station analog for solar radiation (SRAD in MetData) is TSUN, found in the GHCND dataset. When TSUN was included in the variable list only 1644, 2743, 2449, and 2928 instances occur for sorghum, corn, soybean, and wheat over 37 years. In contrast, using MetData allows 17973, 29678, and 27251, 31968.

Figure S4.2: State-averaged maximum temperature for MetData (blue), GHCND (red), and ISD (green) for Texas, Wisconsin, Iowa, and North Dakota. Temperatures are °C.
Figure S4.3: Reporting inconsistencies of WT07. Red line (left y-axis) represents the number of codes issued each year over the region of interest used in this study. Black diamonds (right y-axis) represent the number of stations reporting WT07 each year in the United States.
Table S4.1: R² values for each crop when reports of WT07 between 2008 and 2010 are. Top row includes the model with AOD, the bottom does not. In every instance, dust is still the least important predictor.

<table>
<thead>
<tr>
<th></th>
<th>sorghum</th>
<th>corn</th>
<th>soybean</th>
<th>wheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>AOD</td>
<td>0.714</td>
<td>0.857</td>
<td>0.803</td>
<td>0.812</td>
</tr>
<tr>
<td>no AOD</td>
<td>0.713</td>
<td>0.851</td>
<td>0.812</td>
<td>0.791</td>
</tr>
</tbody>
</table>

Figure S4.4: Seasonality of Deep Blue AOD at 550 nm over the region of interest (1631 counties). Grey polygon bounds the 25th to 75th percentiles. Thick black line denotes 50th percentile. Thick green line denotes 90th percentile, while red points indicate the maximum AOD.
Figure S4.5: Heat map visualization of the correlation matrix computed for all candidate predictors included in basis models for sorghum.
Figure S4.6: Heat map visualization of the correlation matrix computed for all candidate predictors included in basis models for corn.
Figure S4.7: Heat map visualization of the correlation matrix computed for all candidate predictors included in basis models for soybean.
Figure S4.8: Heat map visualization of the correlation matrix computed for all candidate predictors included in basis models for wheat.
Figure S4.9: Heat map visualization of the correlation matrix computed for residuals of the fundamental variable model for sorghum against all dust predictors and AOD.
Figure S4.10: Heat map visualization of the correlation matrix computed for residuals of the fundamental variable model for corn against all dust predictors and AOD.
Figure S4.11: Heat map visualization of the correlation matrix computed for residuals of the fundamental variable model for soybean against all dust predictors and AOD.
Figure S4.12: Heat map visualization of the correlation matrix computed for residuals of the fundamental variable model for winter wheat against all dust predictors and AOD.
Table S4.2: Variable importance for sorghum basis models. Variable importance was calculated using the raw, unscaled permutation importance. Unscaled data should only be compared within single columns (models). Numbers beneath crop names denote R² values for each model.

<table>
<thead>
<tr>
<th>MODIS</th>
<th>MODIS</th>
<th>W 107</th>
<th>W 107</th>
<th>WNS Warning</th>
<th>WNS Warning</th>
<th>WNS Advisory</th>
<th>WNS Advisory</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMAX</td>
<td>Establishment</td>
<td>0.711</td>
<td>0.703</td>
<td>0.712</td>
<td>0.708</td>
<td>0.711</td>
<td>0.710</td>
</tr>
<tr>
<td>TMAX</td>
<td>CriticalWindow</td>
<td>0.714</td>
<td>0.709</td>
<td>0.715</td>
<td>0.709</td>
<td>0.712</td>
<td>0.711</td>
</tr>
<tr>
<td>TMAX</td>
<td>GrainFiling</td>
<td>0.700</td>
<td>0.699</td>
<td>0.702</td>
<td>0.699</td>
<td>0.704</td>
<td>0.704</td>
</tr>
<tr>
<td>TMAX</td>
<td>GrowingSeason</td>
<td>0.716</td>
<td>0.716</td>
<td>0.718</td>
<td>0.717</td>
<td>0.719</td>
<td>0.719</td>
</tr>
<tr>
<td>TMIN</td>
<td>Establishment</td>
<td>0.703</td>
<td>0.703</td>
<td>0.703</td>
<td>0.703</td>
<td>0.703</td>
<td>0.703</td>
</tr>
<tr>
<td>TMIN</td>
<td>CriticalWindow</td>
<td>0.706</td>
<td>0.706</td>
<td>0.706</td>
<td>0.706</td>
<td>0.706</td>
<td>0.706</td>
</tr>
<tr>
<td>TMIN</td>
<td>GrainFiling</td>
<td>0.704</td>
<td>0.704</td>
<td>0.704</td>
<td>0.704</td>
<td>0.704</td>
<td>0.704</td>
</tr>
<tr>
<td>TMIN</td>
<td>GrowingSeason</td>
<td>0.719</td>
<td>0.719</td>
<td>0.719</td>
<td>0.719</td>
<td>0.719</td>
<td>0.719</td>
</tr>
<tr>
<td>PRCP</td>
<td>Establishment</td>
<td>0.741</td>
<td>0.741</td>
<td>0.741</td>
<td>0.741</td>
<td>0.741</td>
<td>0.741</td>
</tr>
<tr>
<td>PRCP</td>
<td>CriticalWindow</td>
<td>0.741</td>
<td>0.741</td>
<td>0.741</td>
<td>0.741</td>
<td>0.741</td>
<td>0.741</td>
</tr>
<tr>
<td>PRCP</td>
<td>GrainFiling</td>
<td>0.741</td>
<td>0.741</td>
<td>0.741</td>
<td>0.741</td>
<td>0.741</td>
<td>0.741</td>
</tr>
<tr>
<td>PRCP</td>
<td>GrowingSeason</td>
<td>0.741</td>
<td>0.741</td>
<td>0.741</td>
<td>0.741</td>
<td>0.741</td>
<td>0.741</td>
</tr>
<tr>
<td>VPD</td>
<td>CriticalWindow</td>
<td>0.777</td>
<td>0.777</td>
<td>0.777</td>
<td>0.777</td>
<td>0.777</td>
<td>0.777</td>
</tr>
<tr>
<td>VPD</td>
<td>GrainFiling</td>
<td>0.778</td>
<td>0.778</td>
<td>0.778</td>
<td>0.778</td>
<td>0.778</td>
<td>0.778</td>
</tr>
<tr>
<td>VPD</td>
<td>GrowingSeason</td>
<td>0.778</td>
<td>0.778</td>
<td>0.778</td>
<td>0.778</td>
<td>0.778</td>
<td>0.778</td>
</tr>
<tr>
<td>SRAD</td>
<td>CriticalWindow</td>
<td>0.777</td>
<td>0.777</td>
<td>0.777</td>
<td>0.777</td>
<td>0.777</td>
<td>0.777</td>
</tr>
<tr>
<td>SRAD</td>
<td>GrainFiling</td>
<td>0.777</td>
<td>0.777</td>
<td>0.777</td>
<td>0.777</td>
<td>0.777</td>
<td>0.777</td>
</tr>
<tr>
<td>SRAD</td>
<td>GrowingSeason</td>
<td>0.777</td>
<td>0.777</td>
<td>0.777</td>
<td>0.777</td>
<td>0.777</td>
<td>0.777</td>
</tr>
<tr>
<td>AODLand</td>
<td>Mean GrainFiling</td>
<td>0.779</td>
<td>0.779</td>
<td>0.779</td>
<td>0.779</td>
<td>0.779</td>
<td>0.779</td>
</tr>
<tr>
<td>AODLand</td>
<td>GrowingSeason</td>
<td>0.781</td>
<td>0.781</td>
<td>0.781</td>
<td>0.781</td>
<td>0.781</td>
<td>0.781</td>
</tr>
<tr>
<td>DustEventFrequencyEstablishment</td>
<td>0.792</td>
<td>0.792</td>
<td>0.792</td>
<td>0.792</td>
<td>0.792</td>
<td>0.792</td>
<td></td>
</tr>
<tr>
<td>DustEventFrequencyCriticalWindow</td>
<td>0.838</td>
<td>0.838</td>
<td>0.838</td>
<td>0.838</td>
<td>0.838</td>
<td>0.838</td>
<td></td>
</tr>
<tr>
<td>DustEventFrequencyGrainFiling</td>
<td>0.838</td>
<td>0.838</td>
<td>0.838</td>
<td>0.838</td>
<td>0.838</td>
<td>0.838</td>
<td></td>
</tr>
<tr>
<td>DustEventFrequencyGrowingSeason</td>
<td>0.838</td>
<td>0.838</td>
<td>0.838</td>
<td>0.838</td>
<td>0.838</td>
<td>0.838</td>
<td></td>
</tr>
</tbody>
</table>
Table S4.3: Variable importance for corn basis models. Variable importance was calculated using the raw, unscaled permutation importance. Unscaled data should only be compared within single columns (models). Numbers beneath crop names denote R² values for each model.

<table>
<thead>
<tr>
<th></th>
<th>MODIS</th>
<th>MODIS</th>
<th>W107</th>
<th>W107</th>
<th>NWS Warning</th>
<th>NWS Warning</th>
<th>NWS Advisory</th>
<th>NWS Advisory</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMAX Establishment</td>
<td>276209</td>
<td>307420</td>
<td>293823.84</td>
<td>284138.98</td>
<td>291450.35</td>
<td>286242.45</td>
<td>293998.44</td>
<td>275334.42</td>
</tr>
<tr>
<td>TMAX CriticalWindow</td>
<td>817148</td>
<td>882885</td>
<td>809133.76</td>
<td>672412.44</td>
<td>797352.25</td>
<td>705438.79</td>
<td>761194.4</td>
<td>67934.17</td>
</tr>
<tr>
<td>TMAX GrainFilling</td>
<td>1107766</td>
<td>1235800</td>
<td>1112424.39</td>
<td>871987.38</td>
<td>1108436.59</td>
<td>986001.47</td>
<td>108011.17</td>
<td>963054.23</td>
</tr>
<tr>
<td>TMAX GrowingSeason</td>
<td>246982</td>
<td>294968</td>
<td>265480.95</td>
<td>262163.87</td>
<td>260641.19</td>
<td>254962.24</td>
<td>231686.94</td>
<td>239154.14</td>
</tr>
<tr>
<td>TMIN Establishment</td>
<td>832886</td>
<td>884545</td>
<td>166292.98</td>
<td>398932.96</td>
<td>386545.94</td>
<td>384379.89</td>
<td>388020.72</td>
<td>397442.14</td>
</tr>
<tr>
<td>TMIN CriticalWindow</td>
<td>441818</td>
<td>372519</td>
<td>371918.91</td>
<td>423795.71</td>
<td>369372.95</td>
<td>412539.91</td>
<td>356475.73</td>
<td>417016.07</td>
</tr>
<tr>
<td>TMIN GrainFilling</td>
<td>1262477</td>
<td>1347226</td>
<td>1344142.92</td>
<td>1062998.31</td>
<td>1249152.08</td>
<td>1059588.73</td>
<td>1225982.02</td>
<td>1037682.24</td>
</tr>
<tr>
<td>TMIN GrowingSeason</td>
<td>315999</td>
<td>359120</td>
<td>531329.94</td>
<td>596130.29</td>
<td>509162.04</td>
<td>584764.32</td>
<td>517993.51</td>
<td>576651.96</td>
</tr>
<tr>
<td>PRCP Establishment</td>
<td>258980</td>
<td>31713</td>
<td>252640.48</td>
<td>291834.08</td>
<td>274992.22</td>
<td>294912.36</td>
<td>271838.14</td>
<td>294510.10</td>
</tr>
<tr>
<td>PRCP CriticalWindow</td>
<td>385157</td>
<td>378866</td>
<td>373503.47</td>
<td>388808.13</td>
<td>375928.62</td>
<td>375909.92</td>
<td>374285.14</td>
<td>388029.73</td>
</tr>
<tr>
<td>PRCP GrainFilling</td>
<td>851062</td>
<td>458956</td>
<td>301322.84</td>
<td>341812.91</td>
<td>333124.39</td>
<td>321141.96</td>
<td>317218.04</td>
<td>319661.36</td>
</tr>
<tr>
<td>PRCP GrowingSeason</td>
<td>1588934</td>
<td>1906475</td>
<td>1670559.04</td>
<td>1279678.98</td>
<td>179446.23</td>
<td>1411099.59</td>
<td>167880.62</td>
<td>1391941.94</td>
</tr>
<tr>
<td>SHAD Establishment</td>
<td>56923</td>
<td>58247</td>
<td>56399.42</td>
<td>323449.78</td>
<td>574359.19</td>
<td>324891.74</td>
<td>592851.76</td>
<td>325207.79</td>
</tr>
<tr>
<td>SHAD CriticalWindow</td>
<td>321440</td>
<td>375287</td>
<td>328531.52</td>
<td>262161.62</td>
<td>327121.32</td>
<td>306317.19</td>
<td>309041.38</td>
<td>256391.106</td>
</tr>
<tr>
<td>SHAD GrainFilling</td>
<td>248076</td>
<td>310969</td>
<td>262519.95</td>
<td>322215.3</td>
<td>245206.43</td>
<td>319782.98</td>
<td>285810.15</td>
<td>302274.67</td>
</tr>
<tr>
<td>SHAD GrowingSeason</td>
<td>290631</td>
<td>368842</td>
<td>407439.75</td>
<td>265041.41</td>
<td>304933.82</td>
<td>243507.25</td>
<td>304123.75</td>
<td>261863.06</td>
</tr>
<tr>
<td>VPD Establishment</td>
<td>833994</td>
<td>941004</td>
<td>866361.49</td>
<td>547221.27</td>
<td>806998.4</td>
<td>348007.74</td>
<td>401526.14</td>
<td>31187.468</td>
</tr>
<tr>
<td>VPD CriticalWindow</td>
<td>1187416</td>
<td>1147186</td>
<td>1258346.42</td>
<td>1375925.52</td>
<td>121287.92</td>
<td>1323146.29</td>
<td>1243470.39</td>
<td>138461.109</td>
</tr>
<tr>
<td>VPD GrainFilling</td>
<td>402743</td>
<td>490914</td>
<td>408030.8</td>
<td>787428.49</td>
<td>448347.89</td>
<td>75070.76</td>
<td>372084.68</td>
<td>75278.763</td>
</tr>
<tr>
<td>VPD GrowingSeason</td>
<td>501589</td>
<td>643412</td>
<td>544076.85</td>
<td>776543.49</td>
<td>492114.82</td>
<td>741175.41</td>
<td>613745.81</td>
<td>703065.58</td>
</tr>
<tr>
<td>EDD Establishment</td>
<td>304408</td>
<td>455359</td>
<td>335789.07</td>
<td>385178.97</td>
<td>338519.98</td>
<td>357624.36</td>
<td>383116.63</td>
<td>372047.382</td>
</tr>
<tr>
<td>EDD CriticalWindow</td>
<td>388311</td>
<td>1098050</td>
<td>1021389.23</td>
<td>102166.12</td>
<td>1081180.89</td>
<td>706300.98</td>
<td>106293.28</td>
<td>166064.867</td>
</tr>
<tr>
<td>EDD GrainFilling</td>
<td>514454</td>
<td>509635</td>
<td>51557.43</td>
<td>85269.92</td>
<td>579682.64</td>
<td>484892.95</td>
<td>552491.28</td>
<td>399995.32</td>
</tr>
<tr>
<td>EDD GrowingSeason</td>
<td>1323894</td>
<td>1617147</td>
<td>1331499.62</td>
<td>1230194.32</td>
<td>1380545.68</td>
<td>1169969.69</td>
<td>132168.18</td>
<td>1230406.525</td>
</tr>
<tr>
<td>year</td>
<td>1437494</td>
<td>1870994</td>
<td>1461479.366</td>
<td>3923238.163</td>
<td>1495286.988</td>
<td>3943579.431</td>
<td>149359.72</td>
<td>3947068.062</td>
</tr>
<tr>
<td>AODLandMean Establishment</td>
<td>435748</td>
<td>461380.897</td>
<td>438933.004</td>
<td>186.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AODLandMean CriticalWindow</td>
<td>204755</td>
<td>227071.502</td>
<td>222738.302</td>
<td>-203.42</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AODLandMean GrainFilling</td>
<td>326888</td>
<td>32862.198</td>
<td>128515.495</td>
<td>462.96</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AODLandMean GrowingSeason</td>
<td>246946</td>
<td>246400.062</td>
<td>242711.846</td>
<td>-900.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DstnEventFrequency Establishment</td>
<td>21149</td>
<td>20107</td>
<td>392.3</td>
<td>9146</td>
<td>401.194</td>
<td>183.017</td>
<td>45450.89</td>
<td>95.867</td>
</tr>
<tr>
<td>DstnEventFrequency CriticalWindow</td>
<td>131576</td>
<td>123332</td>
<td>61.638</td>
<td>-68.911</td>
<td>100.565</td>
<td>85.573</td>
<td>217562.72</td>
<td>71.555</td>
</tr>
<tr>
<td>DstnEventFrequency GrainFilling</td>
<td>22802</td>
<td>43824</td>
<td>-566.722</td>
<td>-90.68</td>
<td>98.669</td>
<td>199.017</td>
<td>312225.66</td>
<td>209.195</td>
</tr>
<tr>
<td>DstnEventFrequency GrowingSeason</td>
<td>97075</td>
<td>106631</td>
<td>587.902</td>
<td>72.568</td>
<td>262.559</td>
<td>-39.123</td>
<td>254182.17</td>
<td>111.017</td>
</tr>
</tbody>
</table>
Table S4.4: Variable importance for soybean basis models. Variable importance was calculated using the raw, unscaled permutation importance. Unscaled data should only be compared within single columns (models). Numbers beneath crop names denote $R^2$ values for each model.

<table>
<thead>
<tr>
<th>MODES</th>
<th>MODES</th>
<th>W107</th>
<th>W107</th>
<th>NWS Warning</th>
<th>NWS Warning</th>
<th>NWS Advisory</th>
<th>NWS Advisory</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMAX Establishment</td>
<td>17634.98</td>
<td>23012.65</td>
<td>18294.742</td>
<td>19400.07</td>
<td>17022.943</td>
<td>18894.294</td>
<td>17836.221</td>
</tr>
<tr>
<td>TMAX CriticalWindow</td>
<td>35705.99</td>
<td>47096.59</td>
<td>38914.582</td>
<td>32165.82</td>
<td>32984.746</td>
<td>32863.403</td>
<td>33516.831</td>
</tr>
<tr>
<td>TMAX GrainFilling</td>
<td>15227.04</td>
<td>186519.41</td>
<td>157781.38</td>
<td>18439.69</td>
<td>154159.217</td>
<td>193595.312</td>
<td>158112.442</td>
</tr>
<tr>
<td>TMAX GrowingSeason</td>
<td>34381.83</td>
<td>47050.33</td>
<td>43293.66</td>
<td>44889.97</td>
<td>41450.769</td>
<td>46063.306</td>
<td>35131.443</td>
</tr>
<tr>
<td>TMN Establishment</td>
<td>26970.36</td>
<td>25970.83</td>
<td>24666.43</td>
<td>26243.01</td>
<td>22617.008</td>
<td>28010.904</td>
<td>22599.748</td>
</tr>
<tr>
<td>TMN CriticalWindow</td>
<td>42540.7</td>
<td>48776.62</td>
<td>42324.38</td>
<td>39265.15</td>
<td>39961.732</td>
<td>37232.018</td>
<td>42535.921</td>
</tr>
<tr>
<td>TMN GrainFilling</td>
<td>162268.69</td>
<td>176153.48</td>
<td>164348.066</td>
<td>169960.81</td>
<td>169162.556</td>
<td>197634.259</td>
<td>168886.082</td>
</tr>
<tr>
<td>TMN GrowingSeason</td>
<td>99348.34</td>
<td>108216.66</td>
<td>99360.40</td>
<td>78743.31</td>
<td>101746.337</td>
<td>79099.297</td>
<td>100621.03</td>
</tr>
<tr>
<td>PRCP Establishment</td>
<td>17584.12</td>
<td>22607.55</td>
<td>19651.15</td>
<td>20069.29</td>
<td>18850.938</td>
<td>19495.447</td>
<td>18116.993</td>
</tr>
<tr>
<td>PRCP CriticalWindow</td>
<td>14890.21</td>
<td>14660.66</td>
<td>14705.26</td>
<td>14688.74</td>
<td>15653.241</td>
<td>16289.973</td>
<td>15251.928</td>
</tr>
<tr>
<td>PRCP GrainFilling</td>
<td>12609.32</td>
<td>13016.71</td>
<td>12626.88</td>
<td>15886.16</td>
<td>12418.945</td>
<td>15587.054</td>
<td>115840.928</td>
</tr>
<tr>
<td>PRCP GrowingSeason</td>
<td>37261.25</td>
<td>68562.18</td>
<td>40031.96</td>
<td>39225.06</td>
<td>45639.928</td>
<td>38017.09</td>
<td>40686.829</td>
</tr>
<tr>
<td>SHAD Establishment</td>
<td>19380.17</td>
<td>23776.67</td>
<td>19823.14</td>
<td>19148.22</td>
<td>19911.287</td>
<td>19631.362</td>
<td>19292.359</td>
</tr>
<tr>
<td>SHAD CriticalWindow</td>
<td>25723.19</td>
<td>27232.05</td>
<td>24624.62</td>
<td>21626.66</td>
<td>25902.429</td>
<td>21948.481</td>
<td>24116.175</td>
</tr>
<tr>
<td>SHAD GrainFilling</td>
<td>30826.01</td>
<td>48473.38</td>
<td>38134.99</td>
<td>41356.38</td>
<td>37477.088</td>
<td>43890.591</td>
<td>36608.418</td>
</tr>
<tr>
<td>SHAD GrowingSeason</td>
<td>28549.09</td>
<td>35486.62</td>
<td>36783.80</td>
<td>36246.79</td>
<td>36779.947</td>
<td>34974.757</td>
<td>31384.727</td>
</tr>
<tr>
<td>VPD Establishment</td>
<td>43140.16</td>
<td>39398.33</td>
<td>43044.86</td>
<td>28727.21</td>
<td>24116.832</td>
<td>22670.409</td>
<td>24091.074</td>
</tr>
<tr>
<td>VPD CriticalWindow</td>
<td>22629.01</td>
<td>28270.96</td>
<td>21842.92</td>
<td>29570.21</td>
<td>21473.478</td>
<td>28588.355</td>
<td>21935.281</td>
</tr>
<tr>
<td>VPD GrainFilling</td>
<td>55655.37</td>
<td>59199.87</td>
<td>52984.99</td>
<td>89417.69</td>
<td>54102.224</td>
<td>80917.286</td>
<td>52720.266</td>
</tr>
<tr>
<td>VPD GrowingSeason</td>
<td>23965.03</td>
<td>29686.48</td>
<td>23860.44</td>
<td>32268.71</td>
<td>25459.149</td>
<td>32905.174</td>
<td>24823.901</td>
</tr>
<tr>
<td>EOD Establishment</td>
<td>21680.76</td>
<td>33684.51</td>
<td>22690.11</td>
<td>36431.34</td>
<td>23449.715</td>
<td>30209.912</td>
<td>22182.03</td>
</tr>
<tr>
<td>EOD CriticalWindow</td>
<td>20814.17</td>
<td>31030.55</td>
<td>24104.35</td>
<td>33660.82</td>
<td>24801.916</td>
<td>32967.489</td>
<td>25473.026</td>
</tr>
<tr>
<td>EOD GrainFilling</td>
<td>87699.26</td>
<td>90414.17</td>
<td>87469.74</td>
<td>86040.76</td>
<td>83344.735</td>
<td>85483.751</td>
<td>85076.72</td>
</tr>
<tr>
<td>EOD GrowingSeason</td>
<td>47419.74</td>
<td>55103.11</td>
<td>49047.89</td>
<td>65352.11</td>
<td>47812.784</td>
<td>61093.769</td>
<td>50972.774</td>
</tr>
<tr>
<td>year</td>
<td>194979.32</td>
<td>213155.52</td>
<td>197090.14</td>
<td>336479.87</td>
<td>193641.43</td>
<td>340208.728</td>
<td>198790.976</td>
</tr>
<tr>
<td>ADOLandMean Establishment</td>
<td>62684.2</td>
<td>62926.99</td>
<td>63926.63</td>
<td>62894.533</td>
<td>62894.533</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADOLandMean CriticalWindow</td>
<td>14103.7</td>
<td>13746.40</td>
<td>12881.48</td>
<td>13569.134</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADOLandMean GrainFilling</td>
<td>63789.2</td>
<td>62675.78</td>
<td>62831.80</td>
<td>63099.659</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADOLandMean GrowingSeason</td>
<td>16667.51</td>
<td>16645.17</td>
<td>16612.95</td>
<td>15888.144</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dust Event Frequency Establishment</td>
<td>1358.7</td>
<td>1775.41</td>
<td>9.0381</td>
<td>45.35</td>
<td>18.40</td>
<td>19.986</td>
<td>154.3</td>
</tr>
<tr>
<td>Dust Event Frequency CriticalWindow</td>
<td>438.91</td>
<td>530.81</td>
<td>20.968</td>
<td>0.17</td>
<td>7.0421</td>
<td>21.48</td>
<td>-1.452</td>
</tr>
<tr>
<td>Dust Event Frequency GrainFilling</td>
<td>3790.86</td>
<td>6742.8</td>
<td>49.46</td>
<td>11.67</td>
<td>-12.738</td>
<td>28.606</td>
<td>31.838</td>
</tr>
<tr>
<td>Dust Event Frequency GrowingSeason</td>
<td>3066.7</td>
<td>3334.61</td>
<td>23.18</td>
<td>34.37</td>
<td>-31.493</td>
<td>63.96</td>
<td>117.054</td>
</tr>
</tbody>
</table>
Table S4.5: Variable importance for wheat basis models. Variable importance was calculated using the raw, unscaled permutation importance. Unscaled data should only be compared within single columns (models). Numbers beneath crop names denote $R^2$ values for each model.

<table>
<thead>
<tr>
<th>Variable</th>
<th>MODIS W</th>
<th>MODIS W</th>
<th>W/T0</th>
<th>W/T0</th>
<th>NWS W</th>
<th>NWS W</th>
<th>NWS A</th>
<th>NWS A</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMAX Establishment</td>
<td>6357.4</td>
<td>1735.8</td>
<td>6550.8</td>
<td>5986.9</td>
<td>6752.7</td>
<td>9466.6</td>
<td>3365.1</td>
<td>3341.3</td>
</tr>
<tr>
<td>TMAX CriticalWindow</td>
<td>8839.7</td>
<td>17816.8</td>
<td>92495.6</td>
<td>9865.7</td>
<td>9719.6</td>
<td>90732.6</td>
<td>90614.9</td>
<td>91750.8</td>
</tr>
<tr>
<td>TMAX GrainFilling</td>
<td>4283.4</td>
<td>7848.2</td>
<td>42914.6</td>
<td>6159.1</td>
<td>42679.4</td>
<td>59617.0</td>
<td>40333.6</td>
<td>50864.5</td>
</tr>
<tr>
<td>TMAX GrowingSeason</td>
<td>7350.1</td>
<td>15229.8</td>
<td>82823.9</td>
<td>141572.9</td>
<td>74206.4</td>
<td>14896.0</td>
<td>58398.0</td>
<td>142275.0</td>
</tr>
<tr>
<td>TMIN Establishment</td>
<td>5013.4</td>
<td>9941.9</td>
<td>55729.7</td>
<td>97014.8</td>
<td>95622.6</td>
<td>94168.1</td>
<td>54192.8</td>
<td>94243.18</td>
</tr>
<tr>
<td>TMIN CriticalWindow</td>
<td>3790.2</td>
<td>9990.0</td>
<td>-391.34</td>
<td>63832.5</td>
<td>37750.7</td>
<td>59253.7</td>
<td>39715.9</td>
<td>60836.11</td>
</tr>
<tr>
<td>TMIN GrainFilling</td>
<td>46732.7</td>
<td>70173.5</td>
<td>47210.8</td>
<td>57422.8</td>
<td>45772.5</td>
<td>59984.1</td>
<td>46942.6</td>
<td>54784.45</td>
</tr>
<tr>
<td>TMIN GrowingSeason</td>
<td>8094.2</td>
<td>17194.6</td>
<td>82631.7</td>
<td>101363.2</td>
<td>76981.9</td>
<td>104874.6</td>
<td>84118.8</td>
<td>103734.9</td>
</tr>
<tr>
<td>PRCP Establishment</td>
<td>48615.1</td>
<td>67469.5</td>
<td>45175.5</td>
<td>40497.8</td>
<td>46756.8</td>
<td>39863.6</td>
<td>48750.7</td>
<td>86030.92</td>
</tr>
<tr>
<td>PRCP CriticalWindow</td>
<td>82679.9</td>
<td>113167.0</td>
<td>82895.8</td>
<td>77800.0</td>
<td>84896.7</td>
<td>75114.1</td>
<td>83086.9</td>
<td>74175.61</td>
</tr>
<tr>
<td>PRCP GrainFilling</td>
<td>5308.3</td>
<td>7820.1</td>
<td>53258.6</td>
<td>82864.5</td>
<td>54150.9</td>
<td>84395.8</td>
<td>51130.1</td>
<td>84769.93</td>
</tr>
<tr>
<td>PRCP GrowingSeason</td>
<td>39984.5</td>
<td>49506.9</td>
<td>40541.1</td>
<td>321356.8</td>
<td>83900.0</td>
<td>34453.9</td>
<td>36378.5</td>
<td>45225.91</td>
</tr>
<tr>
<td>SHAD Establishment</td>
<td>40868.8</td>
<td>51094.3</td>
<td>38991.9</td>
<td>72062.0</td>
<td>40722.6</td>
<td>79856.9</td>
<td>39914.7</td>
<td>72430.85</td>
</tr>
<tr>
<td>SHAD CriticalWindow</td>
<td>80724.4</td>
<td>14877.3</td>
<td>82340.5</td>
<td>154019.4</td>
<td>78901.3</td>
<td>19076.8</td>
<td>79576.4</td>
<td>150426.89</td>
</tr>
<tr>
<td>SHAD GrainFilling</td>
<td>34523.2</td>
<td>59924.1</td>
<td>60268.2</td>
<td>74995.5</td>
<td>44331.2</td>
<td>71345.1</td>
<td>45985.7</td>
<td>70959.99</td>
</tr>
<tr>
<td>SHAD GrowingSeason</td>
<td>41069</td>
<td>70448.5</td>
<td>43632.9</td>
<td>92445.1</td>
<td>45067.5</td>
<td>88888.1</td>
<td>43826.15</td>
<td>90428.65</td>
</tr>
<tr>
<td>VPD Establishment</td>
<td>63128</td>
<td>130225.8</td>
<td>63134.9</td>
<td>168090.9</td>
<td>8584.7</td>
<td>18022.5</td>
<td>60121.76</td>
<td>180949.74</td>
</tr>
<tr>
<td>VPD CriticalWindow</td>
<td>77847.6</td>
<td>107107.9</td>
<td>78372.4</td>
<td>19289.31</td>
<td>75389.8</td>
<td>8653.215</td>
<td>76488.81</td>
<td>83819.89</td>
</tr>
<tr>
<td>VPD GrainFilling</td>
<td>3038.8</td>
<td>130422.2</td>
<td>22403.8</td>
<td>103581.1</td>
<td>96241.4</td>
<td>19349.17</td>
<td>52886.97</td>
<td>101043.76</td>
</tr>
<tr>
<td>VPD GrowingSeason</td>
<td>6159.5</td>
<td>11278.4</td>
<td>64171.9</td>
<td>114198.9</td>
<td>65211.3</td>
<td>11448.8</td>
<td>60213.55</td>
<td>108807.46</td>
</tr>
<tr>
<td>EDD Establishment</td>
<td>84401.8</td>
<td>69954.3</td>
<td>43597.8</td>
<td>88105.9</td>
<td>36406.9</td>
<td>85092.38</td>
<td>43729.29</td>
<td>88379.74</td>
</tr>
<tr>
<td>EDD CriticalWindow</td>
<td>43076.4</td>
<td>61204.8</td>
<td>45871.6</td>
<td>52276.8</td>
<td>45249.5</td>
<td>51978.46</td>
<td>43579.24</td>
<td>52011.58</td>
</tr>
<tr>
<td>EDD GrainFilling</td>
<td>41014.8</td>
<td>7062.1</td>
<td>46321.95</td>
<td>108746.9</td>
<td>41773.4</td>
<td>68680.09</td>
<td>43818.17</td>
<td>68427.19</td>
</tr>
<tr>
<td>EDD GrowingSeason</td>
<td>35659.9</td>
<td>65369.3</td>
<td>37133.3</td>
<td>67170.7</td>
<td>49408.82</td>
<td>76330.82</td>
<td>36641.21</td>
<td>76786.61</td>
</tr>
<tr>
<td>year</td>
<td>220891.6</td>
<td>899928.9</td>
<td>237347.1</td>
<td>99272.7</td>
<td>226783.92</td>
<td>665104.09</td>
<td>219854.37</td>
<td>670348.46</td>
</tr>
<tr>
<td>AODLandMean Establishment</td>
<td>156434.9</td>
<td>154896.9</td>
<td>158639.54</td>
<td>156970.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AODLandMean CriticalWindow</td>
<td>264799.4</td>
<td>25993.19</td>
<td>258866.492</td>
<td>256835.818</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AODLandMean GrainFilling</td>
<td>62126</td>
<td>63870.162</td>
<td>99900.744</td>
<td>60024.371</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AODLandMean GrowingSeason</td>
<td>34999.3</td>
<td>33190.941</td>
<td>345229.62</td>
<td>352231.34</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DustEventFrequency Establishment</td>
<td>9583.8</td>
<td>3041.4</td>
<td>46.936</td>
<td>260.854</td>
<td>36.548</td>
<td>937.73</td>
<td>514.854</td>
<td></td>
</tr>
<tr>
<td>DustEventFrequency CriticalWindow</td>
<td>4304.8</td>
<td>4648.4</td>
<td>14.464</td>
<td>51.095</td>
<td>21.437</td>
<td>22.749</td>
<td>97.689</td>
<td></td>
</tr>
<tr>
<td>DustEventFrequency GrainFilling</td>
<td>2780.9</td>
<td>5663.1</td>
<td>-41.852</td>
<td>22.333</td>
<td>10.36</td>
<td>27.167</td>
<td>15.521</td>
<td>68.296</td>
</tr>
<tr>
<td>DustEventFrequency GrowingSeason</td>
<td>12956.4</td>
<td>62493.1</td>
<td>129801.8</td>
<td>238592</td>
<td>184183</td>
<td>609047</td>
<td>360371</td>
<td></td>
</tr>
</tbody>
</table>

Table S4.6: $R^2$ values for the two-stage regression with the MODIS dust metric. The first regression is performed without dust (with and without AOD), while the second regression is performed using the residuals of the first and the dust metric as the only predictor. We test both random forest (RF) and a linear model (LM).

<table>
<thead>
<tr>
<th>Reg.</th>
<th>Type</th>
<th>w/ AOD</th>
<th>w/o AOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sorghum</td>
<td>RF</td>
<td>-0.0100</td>
<td>0.0093</td>
</tr>
<tr>
<td></td>
<td>LM</td>
<td>0.0019</td>
<td>0.0002</td>
</tr>
<tr>
<td>Corn</td>
<td>RF</td>
<td>0.0007</td>
<td>-0.0009</td>
</tr>
<tr>
<td></td>
<td>LM</td>
<td>0.0035</td>
<td>0.0030</td>
</tr>
<tr>
<td>Soybean</td>
<td>RF</td>
<td>0.0037</td>
<td>-0.0012</td>
</tr>
<tr>
<td></td>
<td>LM</td>
<td>0.0026</td>
<td>0.0012</td>
</tr>
<tr>
<td>Wheat</td>
<td>RF</td>
<td>-0.0013</td>
<td>0.0013</td>
</tr>
<tr>
<td></td>
<td>LM</td>
<td>0.0000</td>
<td>0.0083</td>
</tr>
</tbody>
</table>
Figure S4.13: Illustration of the instability of partial dependence plots when variable importance is extremely low. The regression in each panel uses a different seed for the RF.
Table S4.7: We ranked the variable importance of each dust metric during each phase for several different experiments. In each cell, we include the rank for the dust metric during establishment, critical window, grain filling, and growing season, respectively. For each crop and dust metric, there is a small table of four cells. We perform the experiment without AOD included (left column), with AOD included (right column), and for all counties with data (top row) and with no-dust counties removed (bottom row). When AOD is omitted there are 29 variables in the model, and when AOD is included there are 33 variables in the model.

<table>
<thead>
<tr>
<th></th>
<th>MODIS</th>
<th>WT07</th>
<th>NWS Warning</th>
<th>NWS Advisory</th>
</tr>
</thead>
<tbody>
<tr>
<td>sorghum</td>
<td>1,2,3,4</td>
<td>1,2,3,4</td>
<td>2,3,4,1</td>
<td>1,2,3,4</td>
</tr>
<tr>
<td></td>
<td>1,2,3,4</td>
<td>1,2,3,4</td>
<td>2,3,4,1</td>
<td>1,2,3,4</td>
</tr>
<tr>
<td>corn</td>
<td>1,4,2,3</td>
<td>1,4,2,3</td>
<td>4,2,3,1</td>
<td>2,1,4,3</td>
</tr>
<tr>
<td></td>
<td>1,4,2,3</td>
<td>1,4,2,3</td>
<td>4,2,3,1</td>
<td>2,1,4,3</td>
</tr>
<tr>
<td>soybean</td>
<td>2,1,4,3</td>
<td>2,1,4,3</td>
<td>1,2,3,4</td>
<td>3,1,2,4</td>
</tr>
<tr>
<td></td>
<td>2,1,4,3</td>
<td>2,1,4,3</td>
<td>1,2,3,4</td>
<td>3,1,2,4</td>
</tr>
<tr>
<td>winter wheat</td>
<td>3,1,2,7</td>
<td>3,2,1,4</td>
<td>3,1,2,4</td>
<td>4,2,1,3</td>
</tr>
<tr>
<td></td>
<td>3,1,2,7</td>
<td>3,2,1,4</td>
<td>3,1,2,4</td>
<td>4,2,1,3</td>
</tr>
</tbody>
</table>
Table S4.8: Salient results from the state-by-state analysis.

| Crop  | KS (w AOD) | KS (wo AOD) | NE (w AOD) | NE (wo AOD) | ND (w AOD) | ND (wo AOD) | SD (w AOD) | SD (wo AOD) | ND (w AOD) | ND (wo AOD) | SD (w AOD) | SD (wo AOD) | ND (w AOD) | ND (wo AOD) | SD (w AOD) | SD (wo AOD) |
|-------|-------------|-------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| wheat  | 303 | 303 | 504 | 504 | 504 | 504 | 303 | 303 | 504 | 504 | 504 | 504 | 303 | 303 | 504 | 504 |

<table>
<thead>
<tr>
<th></th>
<th>observations</th>
<th>non-zero dust events</th>
<th>R²</th>
<th>phase (rank)</th>
<th>dominant phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>sorghum</td>
<td>907</td>
<td>205</td>
<td>0.7574</td>
<td>GS (15)</td>
<td>GF</td>
</tr>
<tr>
<td>corn</td>
<td>907</td>
<td>205</td>
<td>0.7719</td>
<td>GF (19)</td>
<td>CW</td>
</tr>
<tr>
<td>wheat</td>
<td>303</td>
<td>109</td>
<td>0.7015</td>
<td>GF (28)</td>
<td>GS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.7086</td>
<td>GF (21)</td>
<td>GS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.7305</td>
<td>GS (31)</td>
<td>CW (20)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.7961</td>
<td>GS (25)</td>
<td>CW (29)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.8453</td>
<td>GS (21)</td>
<td>CW (19)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.8194</td>
<td>ES (18)</td>
<td>ES (26)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.8501</td>
<td>GS (15)</td>
<td>GS (23)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.7401</td>
<td>ES (18)</td>
<td>ES (26)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.7501</td>
<td>ES (10)</td>
<td>ES</td>
</tr>
</tbody>
</table>

106
Chapter 5  
Summary and Conclusion

In this dissertation, we developed data and analysis methods to estimate the effect of dust on crop yields. This research began with a straightforward question: does dust affect crop yields? Much of the existing research estimating the effects of dust on plants has been compartmentalized, focusing on the three components of wind erosion independently. A comprehensive approach to estimating the effect of historical dust events on yields had yet to be undertaken until now. A significant hurdle to answering this question on a large scale was the lack of adequate dust and yield records as well as unsuitable crop models.

Chapter 2 is the foundation of this research and developed the use of random forest (RF) as a diagnostic crop model (Breiman, 2001). While dust is influenced by other climate variables, it can also directly (or indirectly) affect them, rendering the use of standard linear models that had been used in statistical crop modeling over the latter half of the last decade ineffective (Lobell and Field, 2007; Lobell et al., 2008). Although RF had been used in bioinformatics and ecology, it had never been tested as a diagnostic crop model. This work sought to improve understanding of the factors that drive food insecurity, so in Chapter 2, we applied RF as a diagnostic crop model in sub-Saharan Africa (Hoffman et al., 2017). We found that technology was a primary driver of yields in sub-Saharan Africa and, even with coarse data resolution, the RF identified a nonlinear yield response to precipitation in line with agronomic expectations (Figures 2.2 and 2.4b, respectively). While the results of this work were undoubtedly important, the process of creating a climate-crop dataset capable of representing the growing season and growing areas comprised a vast majority of the work and was fundamental to this research.

We expected that creating the data and applying RF as a diagnostic crop model
in sub-Saharan Africa would provide an application to study the effect of dust on yields. However, we determined that better data are required before applying this technique to understand this smaller, more nuanced yield response. Using aggregated data (i.e. country-level yields and growing season averages) was not going to permit the dust effect to emerge. We briefly explored the use of satellite data as a proxy for detecting high-resolution yields, but the technology has not advanced to where this methodology can be applied in potentially intercropped areas.

Based on results in Chapter 2, in Chapter 3, we shifted to a region with high quality and high resolution data. We chose the central region of the United States because data was readily available and the High Plains region is a large anthropogenic dust source (Ginoux et al., 2012). We used daily data to develop crop-specific algorithms to estimate the length of individual growth phases and full growing season for every year. Based on daily temperature data and crop physiology, we estimate the length of the establishment period, critical window, and grain filling period (as well as the entire growing season). The phase lengths were then used to compute the average climate during each phase to identify whether yields were sensitive to different variables during various phases of growth. Though we tested a large number of climate predictors, we found that time (year) combined with the fundamental variables, TMIN, TMAX, PRCP, SRAD, VPD, and EDD explain the majority of variability in yields. This analysis also identified optimal grain filling temperature ranges for sorghum, corn, and soybean (Figure 3.4).

Extensive data synthesis was required for this chapter, particularly with regards to filtering and quality controlling crop yields. Although the USDA NASS yields are more reliable than those from the FAO, inconsistencies in reporting practice limit data availability and usability. I contacted various yield specialists from the National Agricultural Statistics Service at the USDA to verify reporting inconsistencies and our yield estimates before we could continue. Through this work, we identified unique phase-specific responses and further refined the methods that would be used to detect the effect of dust on yields.

In Chapter 4, we used dust as a predictor in the diagnostic crop models. Based on previous work and the results from Chapters 2 and 3, we expected RF to identify a weak but significant effect of dust on yields over the central portion of the United States. Unfortunately, we were unable to detect a yield effect related to dust
over the entire central US region, but when broken down by individual states, we
detected a weak signal in Kansas, Nebraska, South Dakota, and North Dakota.
These responses indicated that counties that experienced fewer dust events (as
detected by satellite data) consistently had higher yields (Figure 4.7). However,
we have low confidence in this result because this signal may be an artifact of
partitioning the data by state or the yield response might be obscured when data
is spatially aggregated. We suspect that the variability in yields caused by dust
events is absorbed by a variable with more explanatory power, like precipitation
or vapor pressure deficit. Despite their low confidence, the weak negative yield
responses to dust in the High Plains provides a direction forward.

Although the results of this work are the primary focus, the amount of data
synthesis in the final chapter is also noteworthy. In addition to the gridded weather
data and county-level yields used in Chapter 3, we included county-level data in
Chapter 4 for over 70 climate variables from satellite and station data for sorghum,
corn, soybean, and winter wheat. A vast amount of data has yet to be explored,
particularly for station-based dust information and weather data in each county.

5.1 Future work

While this work is comprehensive, the results leave numerous questions open that
merit further research.

With this research, we have validated the use of RF as an informative diagnostic
crop model to detect functional responses between yield and individual climate
variables. While this was not the primary goal of the work, this tool can be applied
in other locations, to other crops, and potentially other climate variables. Using the
results of this work, particularly information regarding optimal growing conditions,
it is possible to estimate the effect of climate on future yields. Future work will
predict optimal growing conditions in various future climate scenarios out to 2050.
Early identification of regions with optimal future growing conditions may allow
for infrastructure to be developed early, effectively mitigating the possible negative
yield effects of an altered climate. Furthermore, dust emission may change as a
result of future land use and land cover changes, as well as changes in agricultural
practices and technologies. Assessing the sensitivity of these results to future
scenarios with more dust is important.
We also would like to repeat the experiment in Chapter 4 using irrigated yields because we were unable to detect a response of rainfed crop yields to dust over the entire region. We suspect that the effect of dust on a large scale is absorbed into more powerful predictors when using rainfed yields. If that is the case, using irrigated yields may control for that effect.

We have thoroughly examined the use of standard multivariate linear regression and RF as possible tools. From this work, we learned that the yield effects of dust are much smaller than those of other predictors like time, temperature, and precipitation. Because RF does not force trees to split using dust predictors, dust is rarely used to split nodes, and the resulting yield responses to dust are characterized by low confidence. Repeating this experiment with a technique not based on decision trees could be used to verify the results of this dissertation. We suspect that multivariate adaptive regression splines (MARS) (Friedman, 1991) could provide interesting, and complementary information (as done in Lobell et al. (2014)). MARS is a non-parametric regression procedure (as is RF) that uses the data to create a functional relationship between the predictor and the outcome variable as a multiple piecewise linear regression using ‘basis functions’. The breakpoints in each segment of the piecewise regression are determined from the data, and much like RF, MARS can be used to estimate variable importance based on the basis functions that contribute to the prediction (Hastie et al., 2001).

We were unable to detect a yield response to dust using any of the station-based metrics, even at the individual state level. Though this research focused on observations, another area of possible research could involve simulating dust events using the Weather Research and Forecasting model coupled to model chemistry (WRF-Chem) to simulate counties where dust events may have occurred when station data is unavailable (Grell et al., 2005). WRF-Chem uses the GOCART dust emission scheme from Ginoux et al. (2001) and could be used to ameliorate the spatial inconsistencies with the station-based dust data.

For those that want to carry the torch, it may be wise to focus on a single state, like Nebraska, Kansas, or Oklahoma. Accurate yield and irrigation data are critical - as they represent a considerable source of uncertainty in this type of work. While gridded weather products are user-friendly, crop-relevant, field-scale conditions can be more accurately obtained from proximally located weather stations. Complex, ground-based sensors can be used to distinguish dust events from those with similar
optical properties. The central US has significantly lower dust loadings and fewer dust events than many countries in sub-Saharan Africa, so if the data quality in that region of the world becomes more reliable, this experiment should absolutely be repeated.

The goal of this work was to determine and quantify the effect of dust on crop yields so that we could confidently include or exclude the effect from food security analyses and risk mitigation strategies. Given the available data, the final result of this work suggests a minimal effect of dust on yields. Throughout this work, we uncovered phase-specific yield responses to different climate predictors in a broader geographic and climatic range than previously identified. The next step is to assess the potential impact of dust on irrigated crops, as these crops are potentially the most affected in drought and dust-prone areas.


Food and Agriculture Organization of the United Nations, 2014: *Agriculture and food security statistics of the least developed countries, land locked developing countries and small island developing states*. Rome.


Prasad, P. V. V., K. J. Boote, and L. H. A. Jr., 2006: Adverse high temperature effects on pollen viability, seed-set, seed yield and harvest index of grain-sorghum [Sorghum bicolor (L.) Moench] are more severe at elevated carbon dioxide due to higher tissue temperatures. *Agricultural and Forest Meteorology, 139* (3-4), 237–251, doi:https://doi.org/10.1016/j.agrformet.2006.07.003.


Rougier, J., 2008: Comment on article by Sansó et al. [MR2383247]. *Bayesian Analysis, 3* (1), 45–56, doi:10.1214/08-BA301B.


Tanner, C. and T. Sinclair, 1983: *Efficient water use in crop production: research or re-search*, chap. Limitations to Efficient Water Use in Crop Production, 1–27. ASA, Madison, WI.


Vita
Alexis L Hoffman
alexis.l.hoffman@gmail.com
(310) 261-6964

EDUCATION

Doctor of Philosophy 2018
Meteorology & Atmospheric Science
The Pennsylvania State University, University Park, PA

Master of Science 2013
Meteorology
The Pennsylvania State University, University Park, PA

Bachelor of Art 2011
Earth & Planetary Science
Washington University in St Louis, St Louis, MO

TECHNICAL SKILLS

Extensive Programming in: R, NCAR Command Language, LaTeX, Shell Script
Familiar Programming in: Python, Matlab, Fortran 77/90, SQL
Computing Tools & Software: MS Office, Unix
Operating Systems: Linux
Languages: Spanish (basic)

PROFESSIONAL EXPERIENCE

Graduate Research Assistant 2013 - Present
The Pennsylvania State University, Department of Meteorology, University Park, PA

Member of Technology Staff - Casual Employee 2016 - Present
The Aerospace Corporation, Electronics & Sensors Division, Chantilly, VA

Member of Technology Staff - Graduate Summer Intern 2016
The Aerospace Corporation, Electronics & Sensors Division, Chantilly, VA

Graduate Research & Teaching Assistant 2011 - 2013
The Pennsylvania State University, Department of Meteorology, University Park, PA

AFFILIATIONS

• American Geophysical Union - Member (National Society)
• Graduate Advisory Committee - President & Officer (Department of Meteorology)