

Machine Learning and Climate Indicators as an Approach to Enhance Operational Decisions

Alex Tarter¹, Noemi Guindin¹

¹USDA National Agricultural Statistics Service,
1400 Independence Ave., SW, Washington, DC 20250

Abstract

Forecasting crop yield ahead of a harvest period is a complex problem. In recent years abnormal conditions such as drought, heat waves, freezes, and floods have been observed in major United States crop-producing regions, making forecasting even more challenging. An increase in climate variability and frequency of extreme weather due to climate change is expected to bring additional challenges to crop yield forecasts for different crops in diverse geographies. Our research focuses on using machine learning approaches to develop indicators for critical climate events, with an emphasis on indicators for yield of winter wheat. As an early proof of concept, we develop a random forest model trained on county-level climate data to predict state-level yield in a case study of two U.S. states. Comparison between historic yield forecasting methods, based on survey and remote sensing data, and the random forest techniques indicates that machine learning methodology may be a useful supplement in developing early-season yield estimates.

Key Words: Climate indicators, yield forecasting, random forests, machine learning

1. Introduction

The mission of the National Agricultural Statistics Service (NASS), a statistical agency of the United States Department of Agriculture (USDA), is to provide timely, accurate, and useful statistics in service to U.S. agriculture. One way NASS accomplishes this mission is by publishing a monthly Crop Production Report no later than the twelfth day of each month. These reports assist with the production of Principal Federal Economic Indicators, a key statistical measure of the conditions of the current domestic economy (50 FR 38932, Sep. 25, 1985); as a federal statistical agency, NASS publishes these reports in accordance with federal law. The report includes forecasts of three related quantities for each crop: the harvested acreage total; the projected total production; and the forecasted crop yield, the total production divided by harvested acreage. These statistics are the official consensus of the NASS Agricultural Statistics Board (ASB), a group of statisticians and commodity experts, who review contemporary and historic surveys, administrative-level data, and weather and crop information.

Official agricultural statistics for key crops are important for informing the expectations of commodity markets via the release of the Principal Federal Economic Indicator. NASS therefore has an interest in strengthening its estimation methods by incorporating supplemental relevant data sources to its survey-based forecasting techniques. This paper outlines the early stages of the development of climate indicators, which through machine learning techniques could be incorporated into the in-season crop yield forecasts. The supplemental data sources for climate indicators and the available yield estimates are described in Section 2. Statistical methodology, detailing a random forest model trained on county data per state, is outlined in Section 3. The models under development can produce both county- and state-level yield predictions. Different models are

to be developed for each U.S. state producing the crop of interest. Empirical results for winter wheat in two U.S. states are discussed in Section 4. Discussion and conclusions are provided in Section 5.

2. NASS Crop Yield Estimates and the Agro-Climate Information System

2.1 Establishing Need for Additional Data Sources

NASS publishes forecasts of acreage, production, and yield throughout the growing season for many crops (Vogel, 1998). For example, reports are made public in May, June, and July for the twenty-five U.S. states designated as major contributors to the winter wheat crop, as highlighted in Figure 1. These forecast statistics made available in the Crop Production Report are based on data collected from multiple in-season surveys (which NASS conducts annually for all major crops) along with other sources of information, such as administrative weather data. In recent years, the presence of abnormal weather conditions during critical stages of crop growing seasons has introduced additional complexities to the existing challenges of yield forecasting (Cruze, 2015). For example, in the United States alone, major crop-producing regions have been subject to extreme temperature events, such as heat waves and freezes, and to extreme precipitation events, such as droughts and floods.

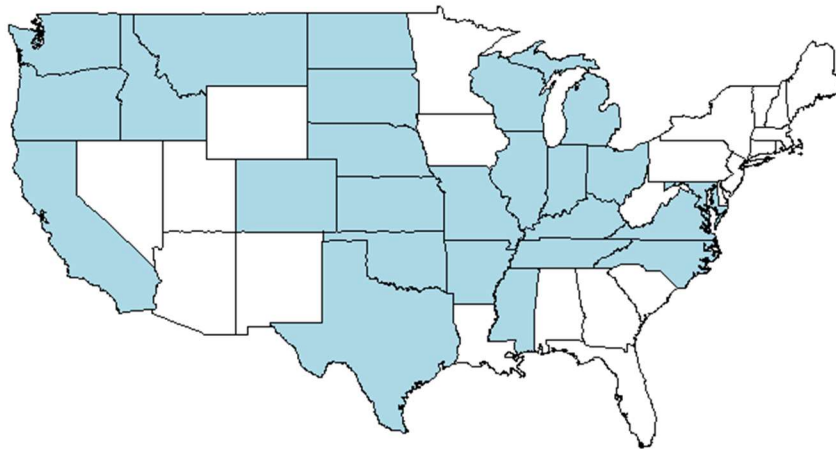


Figure 1: A map of the contiguous United States, with winter wheat producing states in blue.

Extreme weather events can lead to significantly different final actual yield values from the forecasts over the course of the growing season, specifically by lowering the acreage harvested at the end of the season and the number of living plants in the field and therefore the production value, or by a combination of these effects (Harfield & Prueger, 2015; Iizumi, 2015; Mendelsohn, 2007). Globally, nearly one third of crop yield variability can be explained by weather; in fact, over half of variability can be explained by weather in substantial areas of the key global breadbaskets (Lobell and Field, 2007; Ray et al., 2015). For example, an unanticipated reduction in the national corn yield was recorded in the United States in 2020, the most anomalous difference measured since 1993 (Irvin, 2021; Rizzo et al., 2022). Likewise, yield forecasting systems in Europe did not foresee the severe reduction in winter wheat yield recorded in France in 2016 (Ben-Ari et al., 2018). In short, an increase in climate variability and frequency of extreme weather due to climate change is duly expected to bring challenges to yield forecasts for different crops in diverse geographies (Motha, 2011). Since agricultural production is highly sensitive to weather events and climate conditions, NASS has a growing interest in incorporating additional measures of climate information to the current forecasting methods. The variability in published forecasted yields over many growing seasons for these crops can lead to adjustments at later reporting times, as highlighted for recent years of winter wheat yield in Figure 2. Supplementary prediction techniques can assist the yield predictions early in the growing season and, in turn, explain the forecast changes through information about inclement climate events.

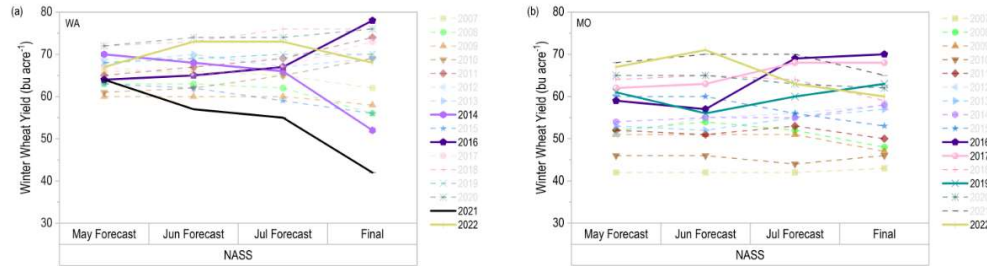


Figure 2: NASS winter wheat yield forecasts from monthly Crop Production Reports and end-of-season yield estimates, 2007-2022, in Washington (a, left) and Missouri (b, right).

2.2 NASS Climate Indicators and Crop Phenology

NASS has entered a co-operative agreement with the Department of Agricultural and Biological Engineering at the University of Florida, leading to the development of the Agro-Climate Information System (CIS), which translates meteorological variables into climate indicators. The climate indicators selected for CIS are relevant to crop growth and development; they are either observed in a sufficient historical context to characterize anomalies and extreme events or can be back-calculated based on the observed variables (Walsh et al., 2020). The CIS then displays the observed value, percentile, and deviation comparisons from the running thirty-year and five-year averages for the period of interest. The climate indicators can be calculated for specific periods, from days to months, and spatially aggregated to different levels according to the user's interest, including state, agricultural statistical district, and county levels.

Indicators have been used for monitoring the potential impact of weather-related events on global crop production within the growing season (Johansson et al., 2015; Ben-Ari et al., 2016; Ceglar et al., 2016). Previous work investigating the impact of weather-related events on winter wheat yields was based on the fall, winter, and spring fixed calendar seasons (Tavakol et al., 2020). However, this method could misrepresent the climate-crop association due to differences in crop management and regional climates (Zhao et al., 2022). To better characterize the potential impacts of weather events on winter wheat yield, crop-critical stages and events that could cause a negative crop yield impact have been identified.

In the U.S., winter wheat is planted in the fall (September and October), goes into dormancy over the winter months, and grows again in the spring. The selected critical stages of this study were from the spring vegetative growth until the mid-reproductive stages. The spring vegetative growth was predicted to start on the date when the 10-day average temperature exceeds 38 °F, which in this study is denoted as the break dormancy date (BDD). For the climate indicators selection, we identified events that have the potential to negatively impact final yield during the selected critical stages. The selected climate indicators are the percentiles of minimum, average, and maximum temperatures, in degrees Fahrenheit; precipitation, in inches; and Agricultural Reference Index for Drought (ARID; Woli et al., 2012), a unitless measure of the proportion of soil evapotranspiration. The climate indicators were derived for three periods: (a) from break dormancy date (BDD) to May 1, (b) from May 1 to June 1, and (c) from June 1 to July 1. These periods are reference points to facilitate the interpretation of the climate indicators for operational decisions by May 1, June 1, and July 1 to adjust in-season forecast.

2.3 NASS Historic Commodity Statistics

NASS stores publicly accessible agricultural data in the Quick Stats database, a comprehensive tool that allows user-specified querying by commodity type and category, geographic location (from county to national levels), and time period. Users can then export the queried data directly through

their web browser to save for external uses. For winter wheat, as with many crops, it is worth noting that there are some caveats to the publication of estimates for acreage, production, and yield. Some counties within a state of interest may not grow a particular crop each year, if ever; in this case, agricultural estimates for the year may be reported with a value of zero units or may simply not be published. In addition, due to regulations regarding the maintenance of producer confidentiality and privacy, NASS has established internal publication standards for statistical disclosure, including (but not limited to) minimum commodity values, minimum harvested crop acreage, and minimum number of commodity reports. In certain cases of an insufficient number of reporting producers within a region, publication of the commodity estimates may be suppressed.

3. Machine Learning-Based Modeling

3.1 County-Based Modeling

The pilot models developed in this project predict how far above or below the running 5-year average the yield will be each year. The difference from the average is predicted as a percentile, to standardize the yield estimates across counties. Once this difference is computed, a linear transformation of the predicted percentile conditioned on the running average yield returns the yield estimate for the region.

The data used in these models include the climate percentiles from those counties with published yield estimates and harvested acreage estimates. Ideally, during every year in which the model should predict the yield for a given county of interest, the estimated yield and acreage from the previous five years ought to be available; however, as mentioned above, given the presence of crop rotations and privacy considerations, these values are sometimes published as zero units or remain unpublished. In this case, provided there is at least one year in the previous five years during which the county-level yield estimate is published, the missing yield values for this model are chosen to be imputed by the estimated yield averaged over the years with available data. Once the list of counties with at least one estimated yield value available in the previous five years is determined from analyzing a Quick Stats query, the climate percentiles for those counties in that year are gathered. This procedure occurs for all years between 1981 (with yields available in 1976) to 2022 since this timeline is when the oldest daily weather data are available for computing the predicted BDD.

3.2 Extending County Models to State Level Predictions

Once the county-level models are generated, the resulting county yield estimates can be combined based on the county-level estimates of the harvested acreage to produce state-level estimates of winter wheat yield. Formally, consider a sequence $(\hat{X}_{t,1}, \dots, \hat{X}_{t,n(t)})$, the unbiased county yield estimates, and $(\hat{Y}_{t,1}, \dots, \hat{Y}_{t,n(t)})$, the corresponding independent unbiased county harvested acreage estimates, available at time t . Then the variable \hat{X}_t defined by

$$\hat{X}_t = \frac{\sum_{k=1}^{n(t)} \hat{X}_{t,k} \hat{Y}_{t,k}}{\sum_{k=1}^{n(t)} \hat{Y}_{t,k}} \quad (1)$$

is an estimator of the state-level yield for the crop at time t . The county sequences of yield and harvested acreage estimates are often incomplete in practice for winter wheat; that is, the number of counties $n(t)$ in the sequences of estimates may be less than the actual number of counties in the state. In this case the numerator of the estimate is a negatively-biased estimator of the total production at the state level, and the denominator of the estimate is a negatively-biased estimator of the total harvested acreage. In years where NASS has published a larger subset of estimated yields from producing counties in each state, the magnitude of bias of both estimators is decreased; when

all producing counties are included, the estimators are unbiased for the total production and total harvested acreage, respectively. Note that, in practice, at least one of the above assumptions (that the county estimates are unbiased and independent) may be violated. For the models in this project, for counties considered during each year, the harvested acreage estimates were developed by simply computing the five-year average harvested acreage, with missing values imputed by a similar process to the yield imputation above.

3.3 Case Study and Sequence of Random Forest Models

As a proof of concept in this pilot study, models are developed for the U.S. states of Washington in the Pacific Northwest and Missouri in the Midwest. The two states were selected due to their differences in geographic region, meaning they have varied climate and soil conditions. Additionally, the causes of winter wheat yield reductions reported by producers contrasted between these states. More specifically, producers stated that lack of moisture in Washington tends to be a critical factor in determining yield reduction, whereas excessive moisture in Missouri is stated as the critical factor. It is therefore reasonable to expect that each climate indicator model produced in this work may identify different weather events that impact production outcomes for each location.

For each state in consideration, a sequence of three models was developed based on the time periods at which the Crop Production Report estimated yields are to be published. The predictor variables in the first model are the 5 climate indicators (for minimum, average, and maximum temperature; precipitation; and ARID) from BDD to May 1, and each subsequent model includes five additional climate indicators from the next chosen period available, or 5 indicators from May 1 to June 1 in the second model and 5 indicators from June 1 to July 1 in the third model.

One of the goals of the CIS project is to incorporate machine learning techniques into the NASS yield estimation methods. With this objective in mind, the earliest stage models developed are random forests, which produce a sequence of gradually improved decision trees. For each state in which a model was developed, the available climate and yield data are randomly split in half for each decision tree in the random forest; half of the data is used for training, and the other half is used for testing and validation. The random forests developed 1000 decision trees in sequence. The data set for Washington consists of 775 rows, and the data set for Missouri consists of 3228 rows. Each row has the county-level percentiles for a given county and year from the CIS and non-zero yield estimates from Quick Stats.

4. Preliminary Case Study Results

In the state of Washington, NASS early-season yield estimates are generally robust to the effects of random weather events; however, as seen in Figure 2(a), in years such as 2014, 2016, and 2021 when crop areas are subject to particularly abnormal climate conditions, the estimated yield can drop somewhat drastically, even in such short time spans as between the monthly Crop Progress Reports. Similar outcomes occurred in the state of Missouri in years such as 2014, 2017, and 2022, as seen in Figure 2(b).

The sequences of models were created for each state in order to generate county-level yield predictions. Alongside the five-year average harvested acreage from each producing county, the forecasted yield was used in Equation (1) to generate a state-level forecast of yield. The final state yield was then compared to the published NASS end-of-season winter wheat yield estimates, considered to be “ground truth” since they occur after the completion of the crop harvest at the end of the season. The series of annual comparisons between the published yield estimates from the historic NASS methodology in the Crop Production Reports in May, June, and July and the predicted yield from the sequence of random forest models in Washington and Missouri, respectively, are displayed in Figures 3 and 4.

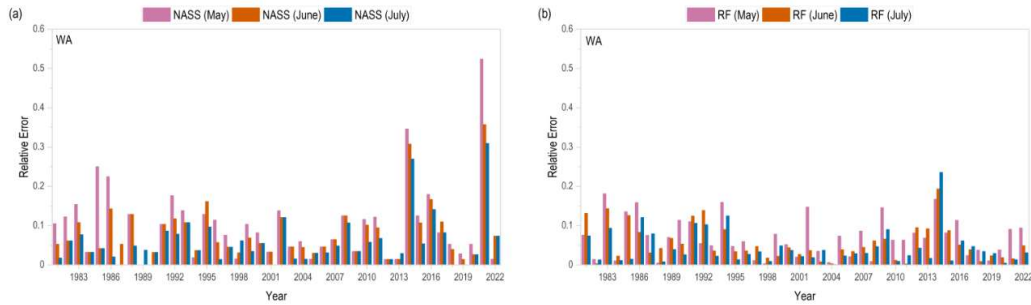


Figure 3: Relative errors of Washington wheat yield forecasts with respect to NASS end-of-season yield estimates in 1981-2022, from historic NASS estimate (a, left) and random forest (RF) techniques (b, right).

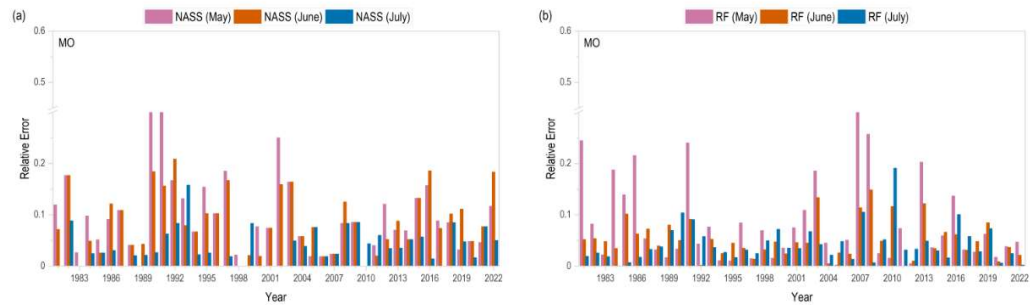


Figure 4: Relative errors of Missouri wheat yield forecasts with respect to NASS end-of-season yield estimates in 1981-2022, from historic NASS estimate (a, left) and random forest (RF) techniques (b, right).

As a summary comparison of the initial random forest models to the NASS standard methodology, the mean absolute percent errors of the wheat yield forecasts from the models and the historic NASS estimates with respect to NASS end-of-season yield estimates were compared for each state under consideration over the entire period of data availability (between 1981 and 2022). In Washington the early-season NASS yield estimates from the Crop Production Report have a 10%, 8.1%, and 6.0% mean absolute error rate in May, June, and July, respectively. Conversely the three new Washington random forest models have 7.4%, 5.4%, and 4.5% mean absolute error rates, respectively. Likewise, in Missouri the early-season NASS yield estimates from the Crop Production Report have a 9.9%, 8.17%, and 3.9% mean absolute error rate in May, June, and July, respectively. Conversely, the three new Missouri random forest models have 8.2%, 5.2%, and 4.3% mean absolute error rates, respectively.

5. Discussion and Conclusions

The results from the initial models of this project, albeit in the early stages of development, demonstrate the potential of using available CIS climate data alongside machine learning techniques to supplement, but not replace, the historic NASS yield estimation methods for winter wheat in the monthly Crop Production Reports. The model has lower mean absolute error rates at all stages of yield prediction (except July in Missouri) as compared with the traditional methods, when measured with respect to the end-of-season yields. This comparison suggests that it may be possible to create climate indicators to determine whether the reported yield estimates will increase or decrease between reports. Though these contemporary models will not replace the historic NASS methodology, the random forest yield estimate may be able to provide improved estimates of

uncertainty in the Crop Production Report statistics on winter wheat yield. Furthermore, in periods of particularly inclement weather, the model may be able to account for the effect that extreme climate events have on final yields.

Future work in model development is multifaceted. At the present the model only includes climate percentile data from the early reproductive stages of winter wheat crop. It is possible to incorporate data from earlier in the planting period, to determine whether any early-season information could improve the yield estimates from the models, and the testing of percentiles from finer time periods (such as May 1 to May 15 and May 15 to June 1, rather than May 1 to June 1) is ongoing to gain a more detailed view of the effects of inclement weather on crop growth. Development of different models for other winter wheat producing states is also underway. Machine learning techniques beyond random forests, such as single- and multi-layer support vector machines, *k*-means clustering methods, and *k*-nearest neighbors algorithms, are also within scope for implementation in the project. Lastly, long term plans include the expansion of the climate indicator methodology toward crops other than winter wheat.

References

- Ben-Ari, T., Adrian, J., Klein, T., Calanca, P., Van der Velde, M., & Makowski, D. (2016). Identifying indicators for extreme wheat and maize yield losses. *Agricultural and Forest Meteorology*, 220, 130-140.
- Ben-Ari, T. et al. (2018). Causes and implications of the unforeseen 2016 extreme yield loss in the breadbasket of France. *Nature Communications*, 9, 1627.
- Ceglar, A., Toreti, A., Lecerf, R., Van der Velde, M., & Dentener, F. (2016). Impact of meteorological drivers on regional inter-annual crop yield variability in France. *Agricultural and forest meteorology*, 216, 58-67.
- Cruze, N.B. (2015). Integrating survey data with auxiliary sources of information to estimate crop yields. In JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association.
- Harfield, J. L., & Prueger, J. H. (2015). Temperature extremes: Effect on plant growth and development. *Weather and Climate Events*, 10, Part A, 4-10.
<https://doi.org/10.1016/j.wace.2015.08.001>
- Iizumi, T., & Ramankutty, N. (2015). How do weather and climate influence cropping area and intensity? *Global Food Security*, 4, 46-50. <https://doi.org/10.1016/j.gfs.2014.11.003>
- Irvin, S. (2021). Was the final USDA estimate of the 2020 U.S. corn yield an outlier? *farmdoc daily*, 11, 30. Department of Agricultural and Consumer Economics, University of Illinois at Urbana-Champaign.
- Johansson, R., Luebehusen, E., Morris, B., Shannon, H., & Meyer, S. (2015). Monitoring the impacts of weather and climate extremes on global agricultural production. *Weather and Climate Extremes*, 10, 65-71.
- Lobell, D. B., & Field, C. B. (2007). Global scale climate–crop yield relationships and the impacts of recent warming. *Environmental research letters*, 2(1), 014002.
- Mendelsohn, R. (2007). What causes crop failure?. *Climatic change*, 81(1), 61-70.
- Motha, R. P. (2011). Chapter 30: The impact of extreme weather events on agriculture in the United States. In S. D. Atti, L. S. Rathore, M. V. Sivakumar, & S. K. Dash, *Challenge and opportunities in agrometeorology* (pp. 397-407). https://doi.org/10.1007/978-3-642-19360-6_30

Office of Management and Budget (1985). Statistical Policy Directive No. 3: Compilation, Release, and Evaluation of Principal Federal Economic Indicators. 50 Fed. Reg. 186.

Ray, D. K., Gerber, J. S., MacDonald, G. K., & West, P. C. (2015). Climate variation explains a third of global crop yield variability. *Nature communications*, 6, 5989.
<https://doi.org/10.1038/ncomms6989>

Rizzo, G., Monzon, J. P., Tenorio, F. A., Howard, R., Cassman, K. G., & Grassini, P. (2022). Climate and agronomy, not genetics, underpin recent maize yield gains in favorable environments. *Proceedings of the National Academy of Sciences*, 119(4), e2113629119.
<https://doi.org/10.1073/pnas.2113629119>

Tavakol, A., Rahmani, V., & Harrington Jr, J. (2020). Probability of compound climate extremes in a changing climate: A copula-based study of hot, dry, and windy events in the central United States. *Environmental research letters*, 15(10), 104058.

Vogel, F. & Bange, G. (1998). *Understanding Crop Statistics*. United States Department of Agriculture. Miscellaneous Publication No. 1554.

Walsh, M. K., P. Backlund, L. Buja, A. DeGaetano, R. Melnick, L. Prokopy, E. Takle, D. Todey, L. Ziska. 2020. *Climate Indicators for Agriculture*. USDA Technical Bulletin 1953. Washington, DC. 70 pages. DOI <https://doi.org/10.25675/10217/210930>.

Woli, P. et al. (2013). Agricultural reference index for drought (ARID). *Agronomy Journal*, 104, 287-300.

Zhao, H., Zhang, L., Kirkham, M. B., Welch, S. M., Nielsen-Gammon, J. W., Bai, G., ... & Lin, X. (2022). US winter wheat yield loss attributed to compound hot-dry-windy events. *Nature communications*, 13(1), 7233.