On The Application of Support Vector Machines to Flash Drought Forecasting

Douglas Schumacher

Undergraduate, Department of Atmospheric and Oceanic Sciences

University of Wisconsin-Madison

Senior Thesis, Spring 2020

Abstract

Flash droughts are rapid onset drought events, typically defined as a two or more categorical increase in drought severity over a six-week period as indicated by the U.S. Drought Monitor (USDM). While long-term seasonal drought prediction has been a common area of research, there is much work to be done on flash drought prediction. Automated flash drought guidance on a weekly basis would be one way to help stakeholders, especially agriculture, prepare for these events in the short-term. This research investigates the feasibility of using a commonly-available machine learning algorithm, support vector machines (SVM), to assist the short-term forecasting of flash drought events. It finds that the use of an SVM trained with precipitation, evapotranspiration, and soil moisture datasets has the ability to classify gridpoints that will enter flash drought criteria over the following six week period with reasonable accuracy.

Introduction

In the last decade, there has been significant discussion in the scientific literature about flash droughts. A good number of the definitions to delineate flash droughts from seasonal-style droughts focus on the period of time for intensification. This paper will use the definition of flash drought as a two or more category increase in drought intensity in the U.S. Drought Monitor within a six week period. The human-produced U.S. Drought Monitor (USDM) product is a weekly analysis of drought contributions across the United States, produced as a collaboration between the National Drought Mitigation Center (NDMC), the Department of Agriculture (USDA), and the National Oceanic and Atmospheric Administration (NOAA). The USDM analysis is treated as the gold-standard of drought analysis, and can be used in this context to determine regions of ongoing flash drought on a week to week basis. While this is a useful monitoring tool, the USDM is not a forecast, and while there is much effort put into seasonal drought forecasting, there is less guidance available for those forecasting flash drought events.

This research seeks to investigate the novel approach of treating flash drought forecasting as a classification problem. This was done by examining the potential usefulness of using a modern machine-learning technique, supervised support vector machines (SVM), to assist in the forecasting of flash drought events in the continental United States. Using a collection of precipitation, evapotranspiration, and soil moisture products from the year 2000 to 2017, SVM models were trained on a randomly stratified set of weekly data points, and then accuracy-tested using a stratified random sample of weekly data points that were set aside.

Literature Review on Flash Drought Prediction

The term "flash drought" was first used in the scientific literature in 2002 by Svoboda et al. to describe rapid-intensification drought events that were being identified in the newly-minted U.S. Drought Monitor (USDM) weekly analysis product. Research on the phenomenon was limited until the widespread flash drought event across the Great Plains and Midwest of 2012 brought the topic to the forefront of the drought community. Otkin et al. (2018) reaffirms the definition of a flash drought in the context of the USDM as "a two-category increase in drought severity over a 6-week period." Common datasets used in the identification and analysis of flash drought events include the Standardized Precipitation Index (SPI), the Evaporative Stress Index (ESI), and soil moisture datasets.

As for the development of predictive methods for forecasting flash drought forecasting, Otkin et al. (2014) describes a probabilistic approach to prediction through the development of the Rapid Change Index (RCI), which is designed to highlight regions undergoing rapid changes in moisture stress as inferred from weekly changes of the ESI. Chen et al. (2018) continues the idea that the identification of rapid changes to evapotranspiration, precipitation, and soil moisture can provide early warnings of flash drought development. They describe an experimental tool developed for the Climate Prediction Center (CPC) for flash drought prediction, in use at the center since April 2018. This product uses the RCI method overlaid with precipitation and soil moisture anomaly thresholds to produce an human-interpretable operational map for forecast support. However, the paper also states that the lack of incorporation of future forecasted precipitation data limits the effectiveness of their method except in circumstances where atmospheric patterns persist during the intervening time period.

With that said, there is a recognition in the literature that there are limited operational products today that synthesize the available precipitation, evapotranspiration, and soil moisture datasets into a flash drought prediction tool. A recent paper from Chen et al. (2019) expands on the team's earlier work described above, examining the onset of flash droughts and stating that their team from NOAA is planning to use their results in an effort to develop a more robust "flash drought prediction tool." As of the time of this writing, the author is unaware of any published literature in the atmospheric sciences examining the use of modern machine learning (ML) techniques on the forecasting of flash drought events, although ML methods are starting to be examined in applications in other parts of the science. As such, the research in this paper seeks to serve as an initial investigation into the effectiveness of one particular type of machine learning technique, SVMs, on the short-term forecasting of flash drought events.

Methods

Datasets and Statistics

To investigate the usefulness of support vector machines to the near-term forecasting of flash drought events, a collection of six datasets were used as part of the training and testing data. These datasets were rasterized onto a standard 4 kilometer grid across the continental United States, as shown in *Figure 1*, with a complete dataset spanning the years 2000 to 2017. From this time frame, data was extracted for the goal of producing a weekly forecast for the warm growing season. Weeks 12 through 47 were selected as forecast dates, corresponding to late March through mid-November. These forecasts were produced using the observational and model data from 6 weeks prior to the forecast data (early February through early October), and the one week change from the week prior to the 6 week initial condition. The datasets used were:

1. *Standardized Precipitation Index (SPI)*

   The Standardized Precipitation Index is a normalized, probability-based index for observed precipitation over a given medium-to-long range time scale. Positive anomalies indicate wetter-than-normal conditions, while negative anomalies indicate drier-than-normal conditions. SPI is often used as a key product in the analysis and forecasting of drought events, given its observational flexibility, simplistic interpretation, and historical context.

2. *Evaporative Stress Index (ESI)*

The Evaporative Stress Index is a satellite-derived, normalized index of the ratio of observed evapotranspiration (ET) by plants to the maximum potential ET. Thermal-IR data from geostationary satellites is fed into a land-atmosphere model to calculate surface energy and moisture fluxes on a daily basis, as described by Anderson et al. (2007a). These fluxes are composited and standardized through the method outlined by Anderson et al. (2007b) over a multi-week period to create a robust measurement of ET responses. ESI has received particular attention in the literature as a leading indicator to the onset of flash drought events, as plants will often experience a period of high ET during the onset of hot, dry periods before their available moisture runs out, and then experience a rapid decrease of ET. This rapid decrease, shown by the ESI as a large negative anomaly, has been shown to proceed the analysed drought development by up to six weeks.

3. *NLDAS Soil Moisture*

Soil moisture data was provided through the use of the North American Land Data Assimilation System (NLDAS) model, providing interval-averaged moisture anomalies for three columns of soil:

   a. Surface-0.1m
   b. Surface-1.0m
   c. Surface-2.0m

It is important to note that these three datasets are not independent of each other, as the deeper layer analysis includes top soil moisture. Nonetheless, all three were included in this research to determine if there was an advantage to the signals provided from any of the layers to predicting flash drought.

4. *U.S. Drought Monitor (USDM)*

The U.S. Drought Monitor is an expert-produced weekly analysis of drought conditions across the United States, published by the National Drought Mitigation Center (NDMC), in collaboration with the USDA and NOAA. It takes into account precipitation and soil moisture data, streamflow and soil moisture data, agricultural and socioeconomic impacts, and citizen reports on the ground, incorporating droughts of many different time scales that may be occurring simultaneously. Drought conditions range on a 6-part intensity scale starting with no drought (encoded as -1 in this dataset) and D0-Abnormally Dry (encoded as 0 in this dataset) to D4-Exceptional Drought (encoded as 4 in this dataset). The USDM category at each gridpoint was used in two ways in this research. One, the USDM category at the time of the production of the 6-week forecast was used as one of the 41 features fed into the model at each gridpoint. Then, the USDM category analysed at the valid forecast week (6 weeks into the future) was used as a verification of the forecast, as if the category increased by 2 or more over the forecast period of six weeks, it was classified as a flash drought grid point. This allows for the supervised training of the support vector machines to occur in the training set and for the calculation of forecast accuracies using the test set.

Support Vector Machine Design

Support vector machines (SVM) are a type of supervised machine learning algorithm that is suited for classification tasks. On the whole, the SVM algorithm takes each data point in the training set, and maps out a representative point in hyperspace, with each feature of the data point as a separate axis. A hyperplane is generated that best separates the two categories of data points (in this case, flash drought develops in the following six week period or it does not), as calculated as the maximum margin between the line and the nearest opposite-class data points on each side of the line. SVMs can be applied to any classification task in which the data set can become linearly separable. In order to achieve linearly separability, the SVM applies an exponential kernel function that remaps all features onto their own uniform coordinate axis.

For this research, the Python module scikit-learn was used to train and evaluate the accuracy of using support vector machines. Scikit-learn is an openly available, "off-the-shelf" style approach to machine learning algorithms, which makes it well-suited for research application to large atmospheric science datasets.

Encoding

For each time step, the entire unfiltered domain, including land and ocean surfaces of and adjacent to the continental United States, comprised 910,000 grid points. However, the full dataset was reduced through a series of quality-control functions, to eliminate points with invalid data across any of the 42 variables extracted. This process first eliminated any point that had a variable with a null value (missing data). Next, it searched through all standardized anomaly-based variables, such as the SPI, ESI, and soil moisture, and eliminated grid points where one or more of the fields had a raw dataset value less than -10 or greater than +10. As these datasets are standardized, this is functionally equivalent to eliminating values beyond 10 standard deviations of their respective mean, and allowed for a masking of all water-based grid points, in addition to locations where satellite measurements could not make observations over that period (such as snow cover in the northern U.S.). This resulted in an average of 434,000 valid data points for any given week, with an average of 17,000 of these points matching flash drought criteria, about 4% of valid points. *Figure 2* shows the number of valid flash drought grid points as a function of week for each year included in the analysis, highlighting the differences of seasonality and geographic extent between years. Two years are highlighted in this figure. In the summer of 2012, a widespread flash drought impacted a large swath of the central United States, an event that stimulated discussion in the drought community about rapid-onset, high impact events. Additionally, in the summer of 2017, a localized flash drought impacted eastern Montana and the western Dakotas. These two events were pulled out as case studies for examining the practicality of using a CONUS-wide model on large and small scale events prior to using the entire 18 year archive.

As the goal of this research was to train a model to forecast flash droughts six weeks into the future, data was extracted for the goal of producing forecast data for weeks 12 through 47, using data from 6 weeks prior to the forecast data, and the one week change from the week prior to the 6 week initial condition. For example, this means that the week 12 USDM category was matched with the week 6 SPI, ESI, and soil moisture data and their changes (also referred to here as deltas) between week 5 and 6.

The SVMs used for this research are binary classifiers that use discrete values for each feature. Thus, thresholds were devised to encode the continuous data points into a handful of classes. USDM classes were left unchanged as they were already integers from -1 to 4. The SPI, ESI, and soil moisture data, already encoded as standardized anomalies, were run through a custom encoding function. This function, outlined in *Table 1*, returned a positive class “p” for values greater than or equal to 0.1; a negative class “n” for values less than or equal to -0.1; and a zero class “z” for values in the middle. This scheme was straightforward and easy on compute power, allowing for preprocessed datasets to be used as inputs into SVM models.

Range of Discrete Anomaly Value	**Encoded Class**
x >= 0.1	p
-0.1 < x < 0.1	z
x <= -0.1	n

Table 1: *Encoding Scheme for turning continuous anomaly values into three discrete classes for use in the SVM.*

Results

2017 Case Study

As a precursor to the SVM training, the relative importance of variables as a predictor of future flash drought was examined through the use of a basic linear regression technique. This analysis used data to determine the relationship of data from week 24 of 2017 to the actual flash drought status of the grid points 6 weeks later, at week 30. *Figure 3* shows the progression of the USDM drought severity from week 24 to week 30, with a flash drought developing over this time period across parts of eastern Montana and the western Dakotas. The predictive analysis was run over the entire domain to determine the correlation between each of the 41 predictive variables (excludes final USDM category) at every gridpoint and the flash drought status of the gridpoint at week 30. *Figure 4* shows the correlation of each product for this single timestep forecast to the binary development/non-development of flash drought, separated by the number of weeks the given data point is averaged over, with the predictiveness of the initial USDM category used as a baseline. It would be expected that this type of correlation would pose a considerable challenge to our SVM model, given that most of the dataset does not experience a flash drought, but may experience a wide variety of changes to SPI, ESI, and soil moisture. Looking at the standard variables, there is indication that the Standardized Precipitation Index (SPI) is the most valuable variable to flash drought prediction. The SPI is one of the primary datasets used by the authors of the USDM to produce their drought analysis, so there is confidence that the approach outlined above works well to determine relative importance of drought prediction variables. The 2 and 4-week composite Evaporative Stress Index (ESI) and 0-10 cm soil moisture analysis are also among the top predictors. Beyond the 8-week composite, only SPI has an advantage over the predictiveness of the initial USDM category. There is an overall trend that the short-term composite products are more predictive than the long-term composite products, which may be due to a stronger signal from current conditions developing during the warm season.

This analysis was also extended to the 1-week change of each variable, in order to introduce an aspect of progression into the forecast. There is a remarkable stability in the correlation of the 4, 8, and 12 week average soil moisture delta datasets, with all three values being more predictive than the initial USDM category. This is evidence that a rapid change to soil moisture content is a good predictor of flash drought onset. On the other hand, weekly changes of long-term SPI and ESI composites are more predictive than their short-term deltas, indicating that large-magnitude changes in these indices are signals to pay attention to when forecasting flash droughts.

A single SVM was trained on this sample dataset as an initial indicator of the effectiveness of the model on a well-known flash drought event, using all 41 predictive variables investigated above. Of the valid CONUS grid points available for week 24, a stratified random sample of 8% of those points were used to train the model, and a separate grouping of 2% was used to test the model. This approach yielded promising results on the testing set, as summarized in *Table 2*:

Portion of Test Sample	**Raw # of grid points**	**Percentage of grid points**	**Categorical (FD/NFD) Accuracy Scores**
Overall Accuracy	9822	100.0%	96.0%
True Positive (FD forecasted as FD)	342	3.5%	81.1%
False Positive (FD forecasted as non-FD)	27	0.3%	
True Negative (Non-FD forecasted as non-FD)	9097	92.6%	96.1%
False Negative (Non-FD forecasted as FD)	356	3.6%	

Table 2: *A truth table for the 2017 flash drought week 30 case study SVM model's test set.*

2012 Case Study

A single SVM approach identical to the 2017 case study was used again to investigate six-week forecast accuracy for the continental U.S.'s most widespread flash drought event to date in the 21st century, during the summer of 2012. During week 29 of that year, nearly 150,000 grid points had reached flash drought criteria, about 31% of all valid CONUS land grid points; this progression is shown in *Figure 5*. *Table 3* summarizes the SVM's results:

Portion of Test Sample	Raw # of grid points	Percentage of grid points	Categorical (FD/NFD) Accuracy Scores
Overall Accuracy	9552	100%	87.6%
True Positive (FD forecasted as FD)	2134	22.3%	82.7%
False Positive (FD forecasted as non-FD)	446	4.7%	
True Negative (Non-FD forecasted as non-FD)	6232	65.2%	89.4%
False Negative (Non-FD forecasted as FD)	740	7.7%	

Table 3: *A truth table for the 2012 flash drought week 29 case study SVM model's test set.*

Accuracy scores were nearly identical between the two case studies for the grid points that did actually reach flash drought status, in the low-80s. However, this widespread case showed a jump in the percentage of non-flash drought grid points incorrectly classified as flash drought grid points. This additional uncertainty could be due to an underrepresentation of the full extent of flash drought conditions during this high-magnitude event, leading to grid points that were actually experiencing flash drought conditions by the indicators but not in the USDM analysis. Nonetheless, these findings show that widespread events can also be forecast with reasonable skill using the SVM method.

Independent weekly SVM

To evaluate the effectiveness of SVM-based prediction on the broad scale, the same approach that was applied to the 2017 and 2012 case studies was applied for the entirety of the

18 year dataset spanning from the year 2000 to the year 2017. 10 percent of each weekly dataset was randomly pulled out in a stratified manner, such that the ratio of flash drought and no flash drought points remained the same between the entire weekly dataset and the training and testing sets. 8% of the overall set was used as training, and 2% as testing examples. As SVMs are limited to training sets of fewer than 50,000 examples, these percentages enabled feasible training times on the order of minutes for each individual model on workstation-grade hardware. Overall, the average accuracy of the weekly models was 96.6%. The total average model performance is summarized in *Table 4*:

Averages for Each Test Sample	**Average Categorical (FD/NFD) Accuracy Scores**
Overall Accuracy	87.6%
True Positive (FD forecasted as FD)	82.7%
False Positive (FD forecasted as non-FD)	9.2%
True Negative (Non-FD forecasted as non-FD)	97.5%
False Negative (Non-FD forecasted as FD)	2.5%

Table 4: *A truth table for the overall averages of the independent weekly SVM models for the entirety of the 2000-2017 period.*

These accuracy scores appear to be quite high, given that this method is using three classes of state variables and three classes of one-week rate of change variables to forecast a time rate of change in the USDM category, while treating it as a binary classification problem. This data can be examined further on a weekly timescale, with *Figure 6* showing the week-to-week accuracies for each year of data, averaged by tertiles based on the cumulative number of flash drought grid points that year. Ordered by most cumulative yearly flash drought points to least number of points, Tertile 1 includes the six years with the greatest cumulative spatial extent of flash drought: 2012, 2000, 2007, 2006, 2002, and 2003. Tertile 2 includes the middle of the pack: 2001, 2011, 2016, 2017, 2005, and 2015. Tertile 3 includes the years with the smallest cumulative geographic extent of flash drought: 2010, 2013, 2008, 2009, 2014, and 2004. There is a small but clear trend to overall accuracy scores for the SVM in that higher-extent flash drought years have slightly lower scores overall, and the difference is maximized in mid-summer, when flash drought grid points numbers are typically highest.

A common trend across all models trained was that the percentage of flash drought points misclassified was nearly always higher than the percentage of non-flash drought points misclassified. This systemic rate may mean that there is some aspect of flash drought development that is not properly captured in the variables used in this analysis. Nonetheless, there is an effectiveness uncovered in this straightforward approach to applying machine learning to flash drought forecasting.

Conclusion

In conclusion, this research goes to show that there is considerable effectiveness in the application of support vector machines to forecasting the onset of flash drought events in the short term. There is potential in a future study on applying the SVM method in an ensemble fashion, with multiple classifiers trained on fully random data and combined into a single predictive model, much in the same way that numerical weather prediction ensemble models are used in the operational space today. Regional and seasonal patterns could be examined as a potential avenue of improving specific model improvements. Additionally, the methods investigated in this paper could feasibly be turned into an operational weekly product, allowing for real-time flash drought forecasting guidance.

Acknowledgements

With the completion of this senior thesis, there are many thanks to be given. First and foremost, this project would not have been possible without the efforts of my research advisor, Jason Otkin, Associate Scientist at the Cooperative Institute for Meteorological Satellite Studies. Jason has worked with me since he first hired me as a summer research assistant to work on building tools to analyze citizen science condition reports for flash drought events in May of 2019. This thesis was an extension of that summer project, applying the conventional gridded datasets to the application of forecasting, which is the topic of my greatest interest in the atmospheric sciences. Jason has provided tremendous guidance, contacts, and resources for this undertaking, and has been a patient, understanding, and flexible advisor over the last two semesters, especially as the COVID-19 pandemic has upended all our daily lives during the spring of 2020. For all this I am very grateful.

Additionally, I'd like to thank Dr. Michael Morgan, Professor in the Atmospheric and Oceanic Sciences department at UW-Madison, for serving as the department liaison and officially signing off on this project. Finally, I want to thank David Lorenz of the Center for Climatic Research at UW-Madison for graciously providing all of the gridded datasets that were analyzed in this research.

Figures

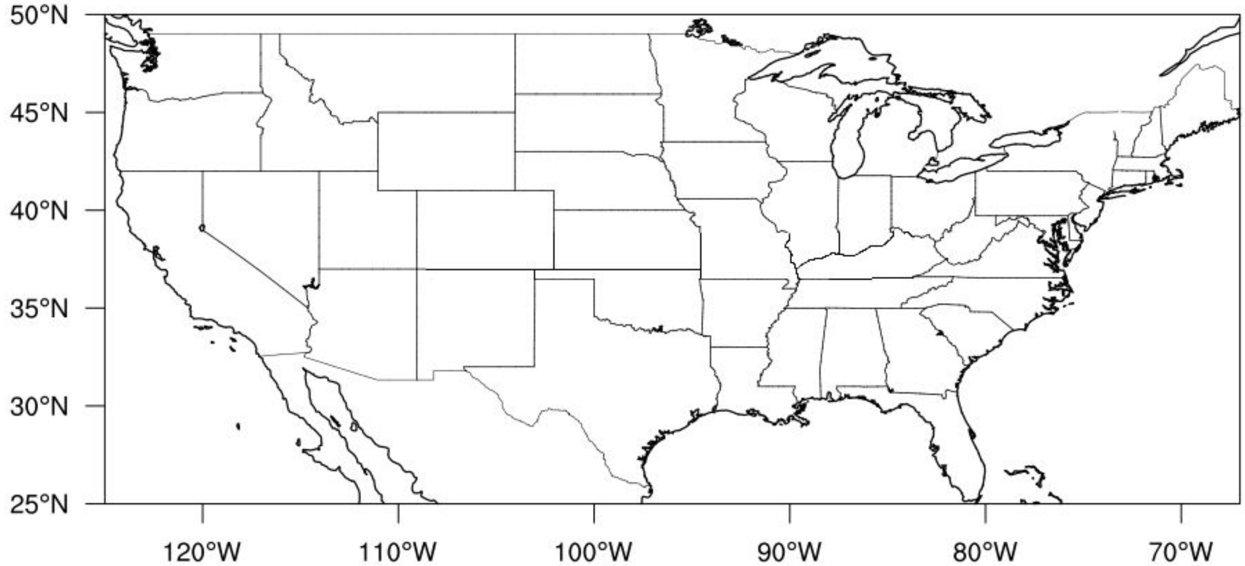

Figure 1: *Extent of data domain across the continental United States analyzed in this project.*

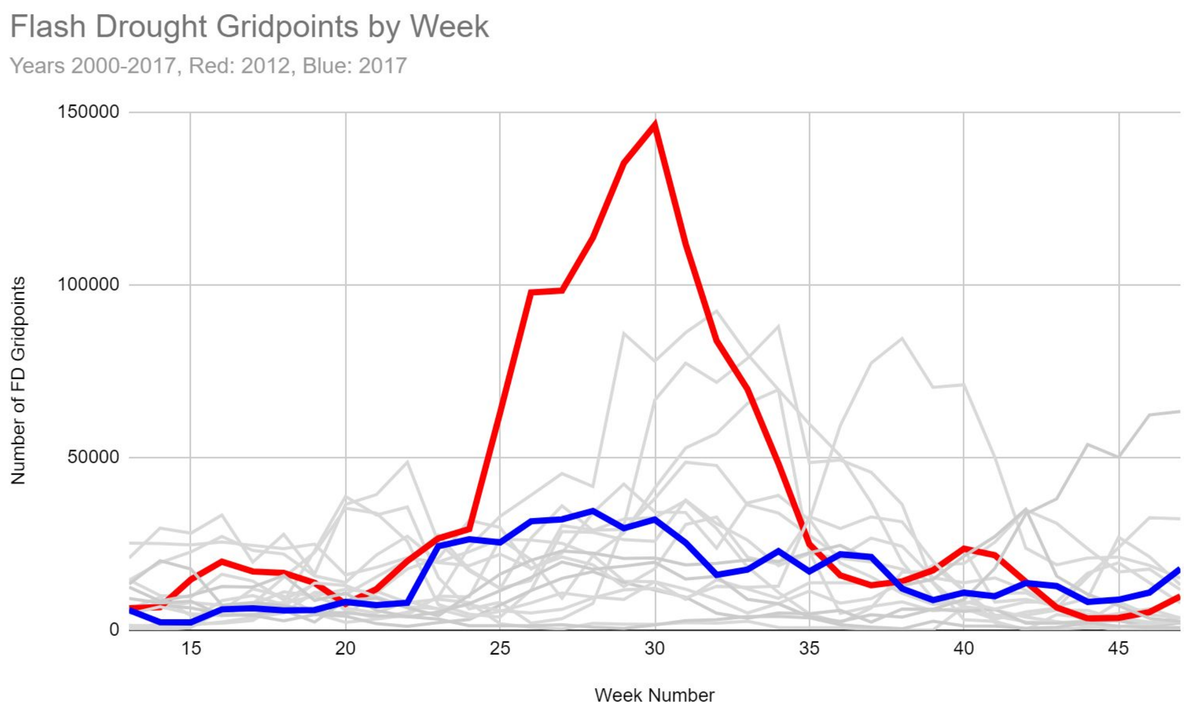

Figure 2: *Number of valid flash drought grid points by week for the years 2000-2017. Case study years highlighted; 2012: Red, 2017: Blue*

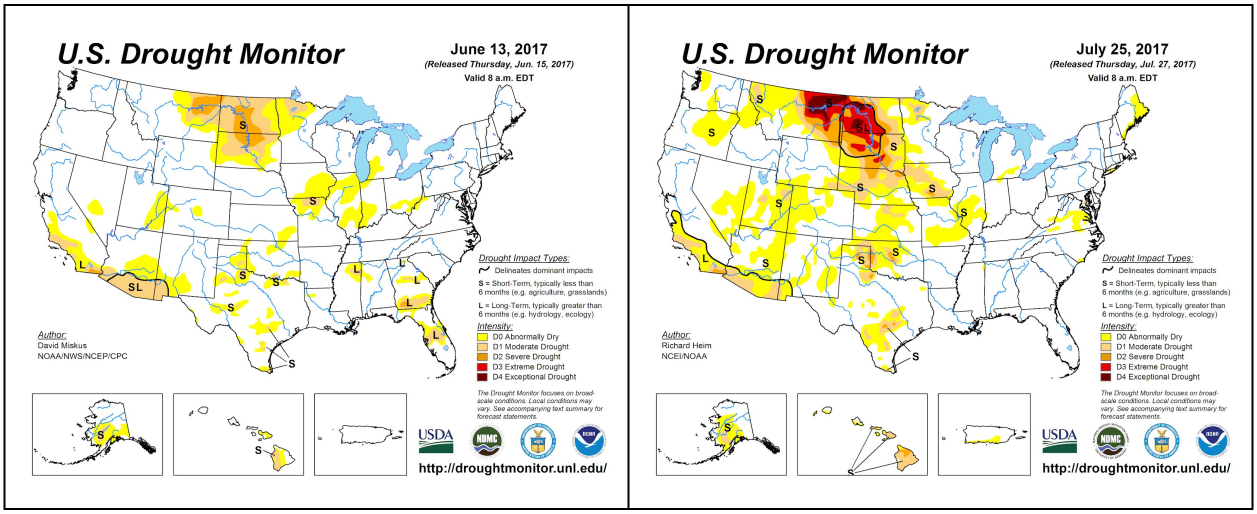

Figure 3*: U.S. Drought Monitor analysis for week 24 (left) and week 30 (right) of 2017. Note the severe flash drought development across eastern Montana and the western Dakotas, and the less intense flash drought development over Nebraska and Iowa over this 6-week time frame.*

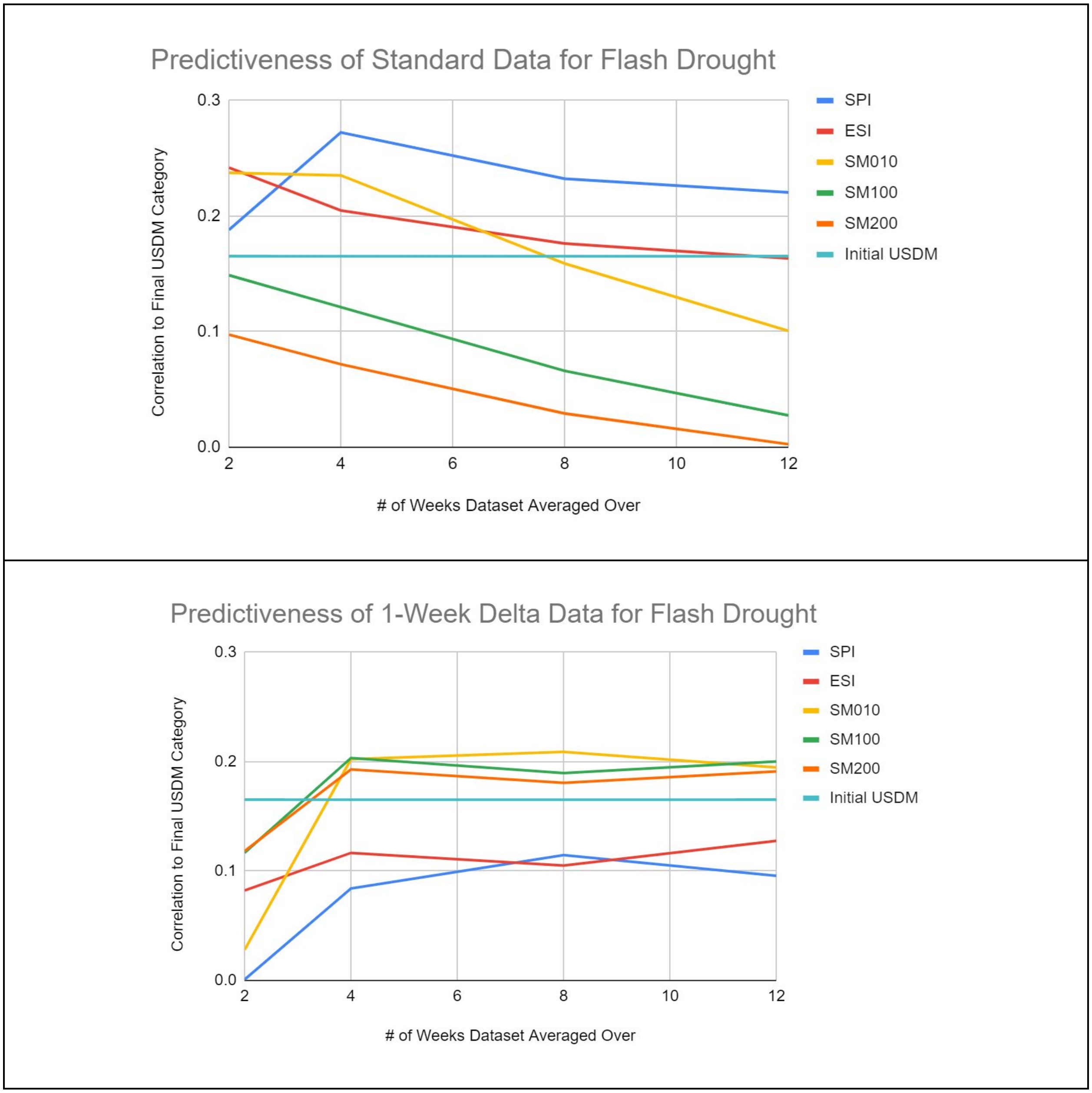

Figure 4: *Line charts showing the correlation of the 41 predictive variables (week 24) to the final USDM category six weeks later (week 30) for the 2017 case study. Each dataset used came in 4 flavors: 2, 4, 8, and 12 week averages, as shown on the x-axis. On top, the standard variables plotted alongside the initial USDM category as a baseline of effective prediction. 2 and 4 week SPI and ESI at the time of forecast creation are the most predictive. On bottom, the predictiveness of one-week changes between week 23 and 24 for all the variables. Changes of 4, 8, and 12 week column soil moisture are most predictive.*

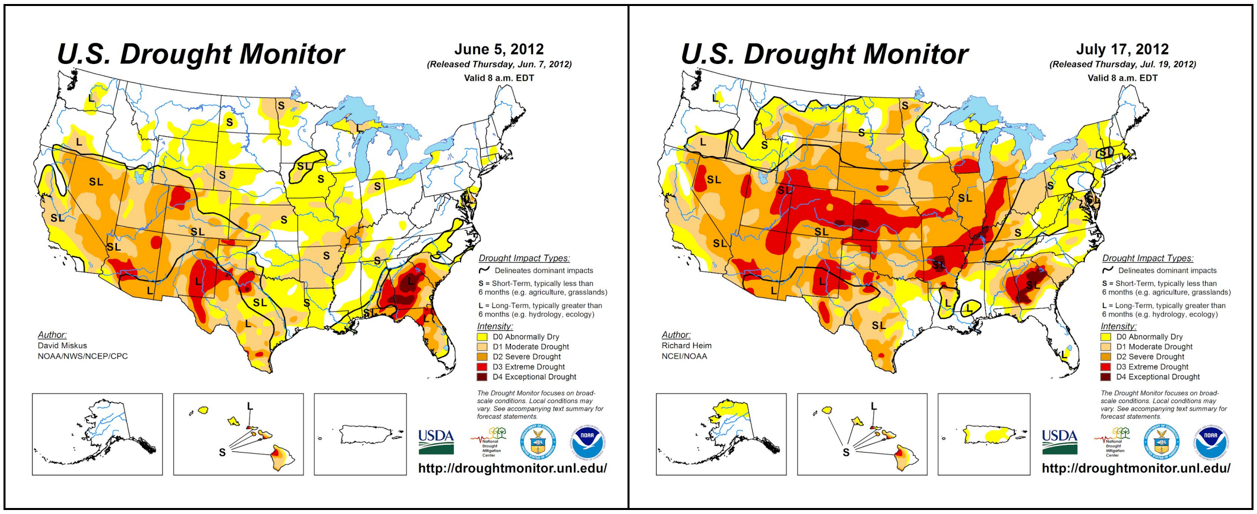

Figure 5: *U.S. Drought Monitor analysis for week 23 (left) and week 29 (right) of 2012. Note the widespread drought intensification across the central Plains and Upper Midwest.*

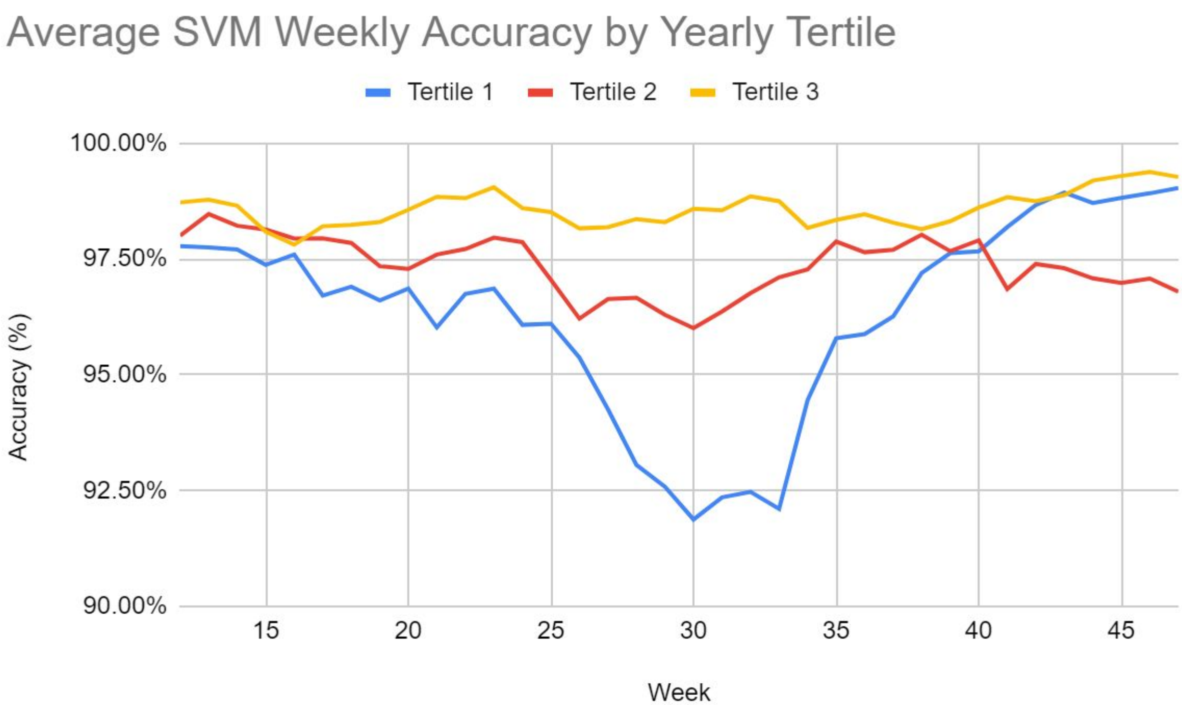

Figure 6: *Overall average accuracy for SVMs trained and tested as weekly forecasts for each year of the entire dataset from 2000 to 2017. Split and averaged into tertiles based on the cumulative number of flash drought grid points.*
Tertile 1: 2012, 2000, 2007, 2006, 2002, 2003.
Tertile 2: 2001, 2011, 2016, 2017, 2005, 2015.
Tertile 3: 2010, 2013, 2008, 2009, 2014, 2004.

References

1. Svoboda, M., D. LeComte, M. Hayes, R. Heim, K. Gleason, J. Angel, B. Rippey, R. Tinker, M. Palecki, D. Stooksbury, D. Miskus, and S. Stephens, 2002: THE DROUGHT MONITOR. Bull. Amer. Meteor. Soc., 83, 1181–1190, https://doi.org/10.1175/1520-0477-83.8.1181
2. Otkin, J.A., M. Svoboda, E.D. Hunt, T.W. Ford, M.C. Anderson, C. Hain, and J.B. Basara, 2018: Flash Droughts: A Review and Assessment of the Challenges Imposed by Rapid-Onset Droughts in the United States. Bull. Amer. Meteor. Soc., 99, 911–919, https://doi.org/10.1175/BAMS-D-17-0149.1
3. Otkin, J.A., M.C. Anderson, C. Hain, and M. Svoboda, 2014: Examining the Relationship between Drought Development and Rapid Changes in the Evaporative Stress Index. J. Hydrometeor., 15, 938–956, https://doi.org/10.1175/JHM-D-13-0110.1
4. Chen, L.G., Gottschalck, J.; Hartman, A.; Miskus, D.; Tinker, R.; Artusa, A. Flash Drought Characteristics and Prediction. NOAA 2018.
5. Chen, L.G., Gottschalck, J.; Hartman, A.; Miskus, D.; Tinker, R.; Artusa, A. Flash Drought Characteristics and Prediction. Atmosphere 2019, 10, 498.
6. Anderson, M. C., J. R. Norman, J. R. Mecikalski, J. A. Otkin, and W. P. Kustas, 2007: A climatological study of fluxes and moisture stress across the continental U. S. based on thermal remote sensing. 1: Model formulation. J. Geophys. Res., 112, D10117, doi:10.1029/2006JD007506.
7. Anderson, M. C., J. R. Norman, J. R. Mecikalski, J. A. Otkin, and W. P. Kustas, 2007: A climatological study of fluxes and moisture stress across the continental U. S. based on thermal remote sensing. 2: Surface Moisture Climatology. J. Geophys. Res., 112, D11112, doi:10.1029/2006JD007507.