Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 95 (2016) 229 – 236

Procedia Computer Science

Complex Adaptive Systems, Publication 6
Cihan H. Dagli, Editor in Chief
Conference Organized by Missouri University of Science and Technology
2016 - Los Angeles, CA

Classifying Drought in Ethiopia Using Machine Learning

Michael B. Richman[a]*, Lance M. Leslie[a] and Zewdu T. Segele[a]

[a]*School of Meteorology, University of Oklahoma, 120 David L. Boren Blvd, Suite 5900, Norman, OK 73072, USA*

Abstract

This study applies machine learning to the rapidly growing societal problem of drought. Severe drought exists in Ethiopia with crop failures affecting about 90 million people. The Ethiopian famine of 1983–85 caused a loss of ~400,000–1,000,000 lives. The present drought was triggered by low precipitation associated with the current El Niño and long-term warming, enhancing the potential for a catastrophe. In this study, the roles of temperature, precipitation and El Niño are examined to characterize both the current and previous droughts. Variable selection, using genetic algorithms with 10-fold cross-validation, was used to reduce a large number of potential predictors (27) to a manageable set (7). Variables present in ≥ 70% of the folds were retained to classify drought (no drought). Logistic regression and Primal Estimated sub-GrAdient SOlver for SVM (Pegasos) using both hinge and log cost functions, were used to classify drought. Logistic regression (Pegasos) produced correct classifications for 81.14% (83.44%) of the years tested. The variable weights suggest that El Niño plays an important role but, since the region has undergone a steady warming trend of ~1.6°C since the 1950s, the larger weights associated with positive temperature anomalies are critical for correct classification.

Peer-review under responsibility of scientific committee of Missouri University of Science and Technology

Keywords: Classification, Support Vector Machines, Pegasos, Logistic Regression, Drought, Global Climate Change

1. Introduction

East Africa frequently experiences catastrophic droughts. Currently, severe drought across the Horn of Africa (affecting countries including Ethiopia, Kenya, Somalia, Uganda and Djibouti) has placed over 12 million people in urgent need of assistance. In some areas, the current drought is the worst in over 60 years, as over the past year, the Horn of Africa has experienced two consecutive failed rainy seasons, resulting in one of the driest periods since

* Corresponding author. Tel.: +1-405-325-1853; fax: +1-405-325-7689.
E-mail address: mrichman@ou.edu

Peer-review under responsibility of scientific committee of Missouri University of Science and Technology
doi:10.1016/j.procs.2016.09.319

1950/51. Ethiopia, the focus of this study, has a population of about 90 million, of which 10 million are in need, and over 2 million are acutely malnourished because ~80% of the population is reliant on agriculture.

Ethiopia was in drought even before the El Niño phase of El Niño Southern Oscillation (ENSO) hit. Whereas El Niño typically brings more rain to California and the southern United States, it causes drought in other parts of the world, including eastern and southern Africa. Between August and October, 2015 the number of people in need of aid doubled, and numbers have continued to rise sharply, since the drought was exacerbated by El Niño.

The massive Ethiopian famine of 1983–1985, which resulted from a combination of drought and conflict with neighboring countries, 1983–85 left an estimated 400,000–1 million people dead [1], from a population which at that time was much lower, at 40 million, than the present 90 million. The back-to-back recent Ethiopian wet season rainfall failures, which largely are blamed on the current El Niño event, have created a drought, that in some areas of the country, is worse than that of 1983–1985 million.

The main goal of this study is to find a relatively small set of predictors that accurately classify Ethiopian drought years from non-drought years. The target years are the known set of drought years for the period 1953–2013, for which data is available. Although the approach in this study has been applied to both north and central Ethiopia, only the results for northern Ethiopia are described in detail in this study, as the method applied is the same for any climate sub-region of Ethiopia.

2. Data and Methods

2.1. Data

Monthly high resolution gridded rainfall and temperature data for 1953–1993 were obtained from the Climate Research Unit at the University of East Anglia [2]. Station anomalies (from the 1961–1990 means) were interpolated into 0.5° latitude/longitude grid cells covering global land surface. The monthly mean temperature and precipitation totals the cells for the regions 11–13.5°N and 39–41°W were averaged to provide predictor data for the agriculturally important region of northern Ethiopia (Fig. 1).

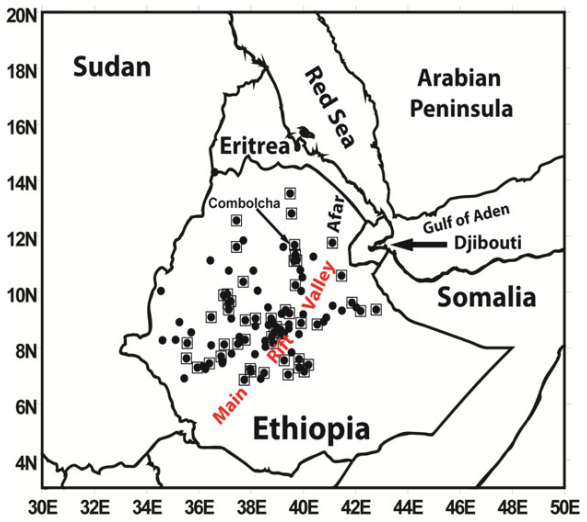

Fig. 1. Map of Ethiopia with location Combolcha centered in northern Ethiopia.

2.2. Methods

El Niño is known to contribute to drought conditions in Ethiopia [3]. This link was extended for various time scales finding numerous SSTA links to Ethiopian precipitations [4]. Hence the Niño3.4 sea surface temperature anomalies (SSTA) and the Trans-Niño Index (TNI) are included, as well as the Atlantic Meridional Oscillation

(AMO), the Pacific Decadal Oscillation (PDO), the Tropical Northern Index (TNI), North Pacific Index (NPI)and Western Hemisphere Warm Pool (WHWP) are included. Analysis of the precipitation and temperature data shows substantial trends; therefore, the global temperature time series and a linear trend (year number) are added to the predictor pool. The growing season in Ethiopia ends by October and the monthly precipitation and temperature data, limited to months for January–September, are included with the aforementioned climate drivers for a total of 27 potential predictors.

A wavelet analysis is used to spectrally decompose locally (in time) the north Ethiopian precipitation and temperature time series for the period 1953–2013. The technique described by [5] uses the Morlet wavelet [6, 7] and provides wavelet-filtered time series for the major temporal modes of Ethiopian precipitation and temperature data. The wavelet power spectrum describes the spectral characteristics of the time series in a time–frequency domain [8].

After communication with the Ethiopian Meteorological Service, years of drought were identified as 1965, 1969, 1972, 1976, 1978, 1982, 1984, 1987, 1990, 1992, 1993, 1997, 2002, 2004 and 2010, and binarized (1, 0) for classification. A genetic algorithm (GA) [9] with a cross-over probability of 0.6 and mutation probability of 0.033 was applied for a feature subset selection to the 27 predictors. Predictors highly correlated with the class, yet uncorrelated with each other are desirable. A 10-fold cross-validation GA process was applied to the 27 predictors to select the subset of predictors (7) that appear in ≥ 70% of the GA folds: Niño3.4 (70%), TNI (90%), precipitation in Jan. (70%), Aug. (100%), Sep. (80%) and temperature in Jun. (100%) and Aug. (90%). The classification process used Primal Estimated sub-GrAdient SOlver (Pegasos) for support vector machines (SVM). Pegasos is an effective algorithm for approximately minimizing the objective function of SVM, allows for a number of cost functions (e.g, hinge, log loss convex function, ϵ-insensitive, cost-sensitive loss) and control of regularization, a low computational complexity, a fast rate of convergence for linear and nonlinear kernels and was found to achieve excellent generalization [10]. Both the hinge and log loss functions are investigated.

3. Results

3.1. Wavelet Analysis of Precipitation and Temperature

Fig. 2a shows the precipitation time series, revealing a slight downward linear trend. The near-surface temperature time series indicates a large and steady increase throughout the entire 1953–2013 period (Fig. 2b). There is an increase over the period of ~1.6^{0}C. The global wavelet spectrum for precipitation (Fig. 3a) indicates statistically significant peaks at the 99th percentile for ENSO timescales (2–7 years). There is another low frequency

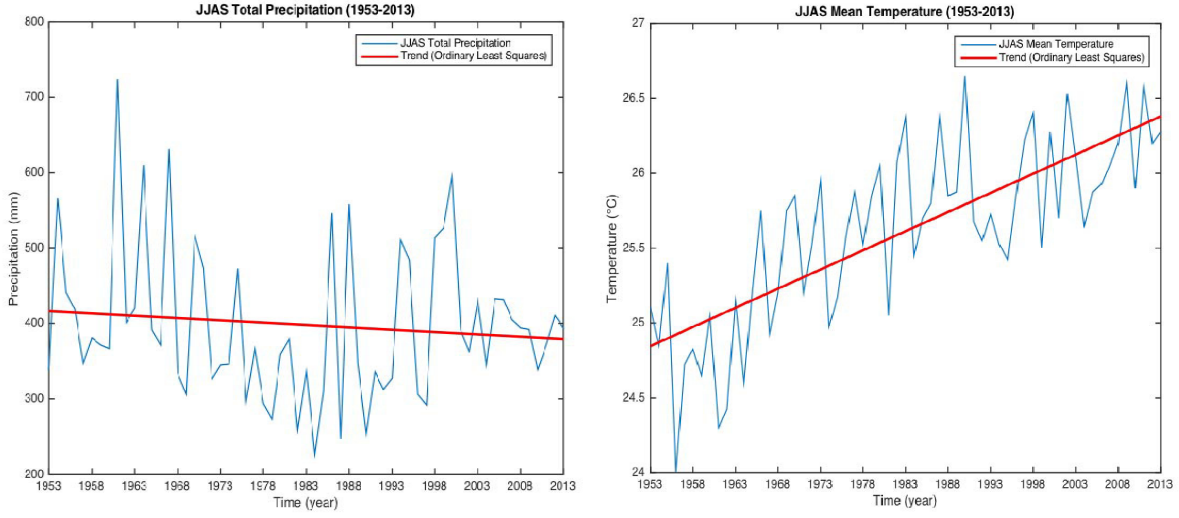

Fig. 2. (a) Northern Ethiopian precipitation time series and trend; (b) Same as (a), except for temperature.

peak suggested at about 30 years but, with a 61-year dataset, it is not statistically significant. The temperature global wavelet spectrum (Fig. 3b) shows a statistically significant ENSO global wave spectrum signal, and several other possible decadal and multi-decadal periods.

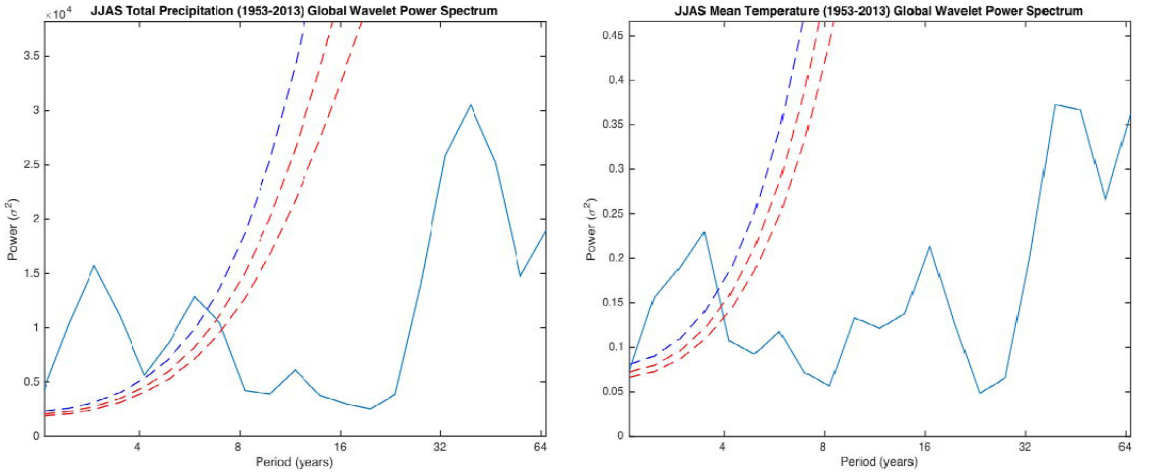

Fig. 3. (a) Northern Ethiopian precipitation local wavelet spectrum with dashed lines corresponding to the 90th, 95th and 99th percentile significance levels. The 99th percentile is the blue dashed line; (b) Same as (a), except for temperature.

The local wavelet power spectrum, with the region above the lower red dashed line (Fig. 4a) being significant at the 95th percentile, shows strong power at ~2- to 5-year periodicity from 1958 to 1963 and from ~2- to 8-years during 1988 to 2003 for precipitation. For temperature (Fig. 4b), the local wavelet spectrum underlines the strength of the ENSO signal in the 1960s, 1980s and the late 1990s to early 2000s at a 3-year periodicity.

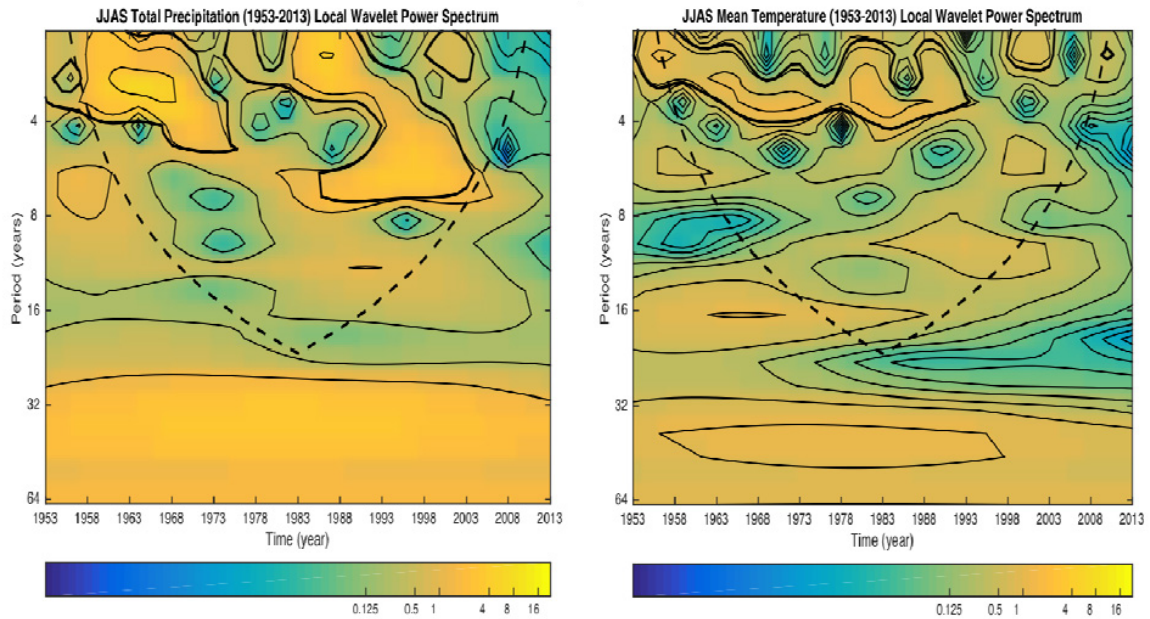

Fig. 4. (a) Northern Ethiopian precipitation local wavelet spectrum power with dashed line corresponding to the 95th percentile significance level; (b) Same as (a), except for temperature.

3.2. Classification of drought

Results are shown for both logistic regression and Pegasos with log and hinge cost functions. After a 10-fold cross-validation, the inputs found to optimally classify drought were Niño 3.4, August precipitation and August temperature. Evaluation of the drought classification are made by examining (1) how logistic regression compare to Pegasos methods, (2) the role of cost function selection and (3) the number of training epochs for the Pegasos method. Four evaluation indices are examined for the log cost function (Fig. 5): accuracy (the number of correct positive and negative forecasts divided by the total number of forecasts), the probability of detection (the number of correct positives divided by the number of observations of drought), the false alarm rate (the number of false positives divided by the number of yes forecasts) and the Heidke skill score or kappa that Pegasos, (a normalized function of the number of correct positive and negatives minus the number of false positives and negatives). Comparison of the mean accuracy of logistic regression (0.814) to Pegasos with log cost function (Fig. 5a) shows

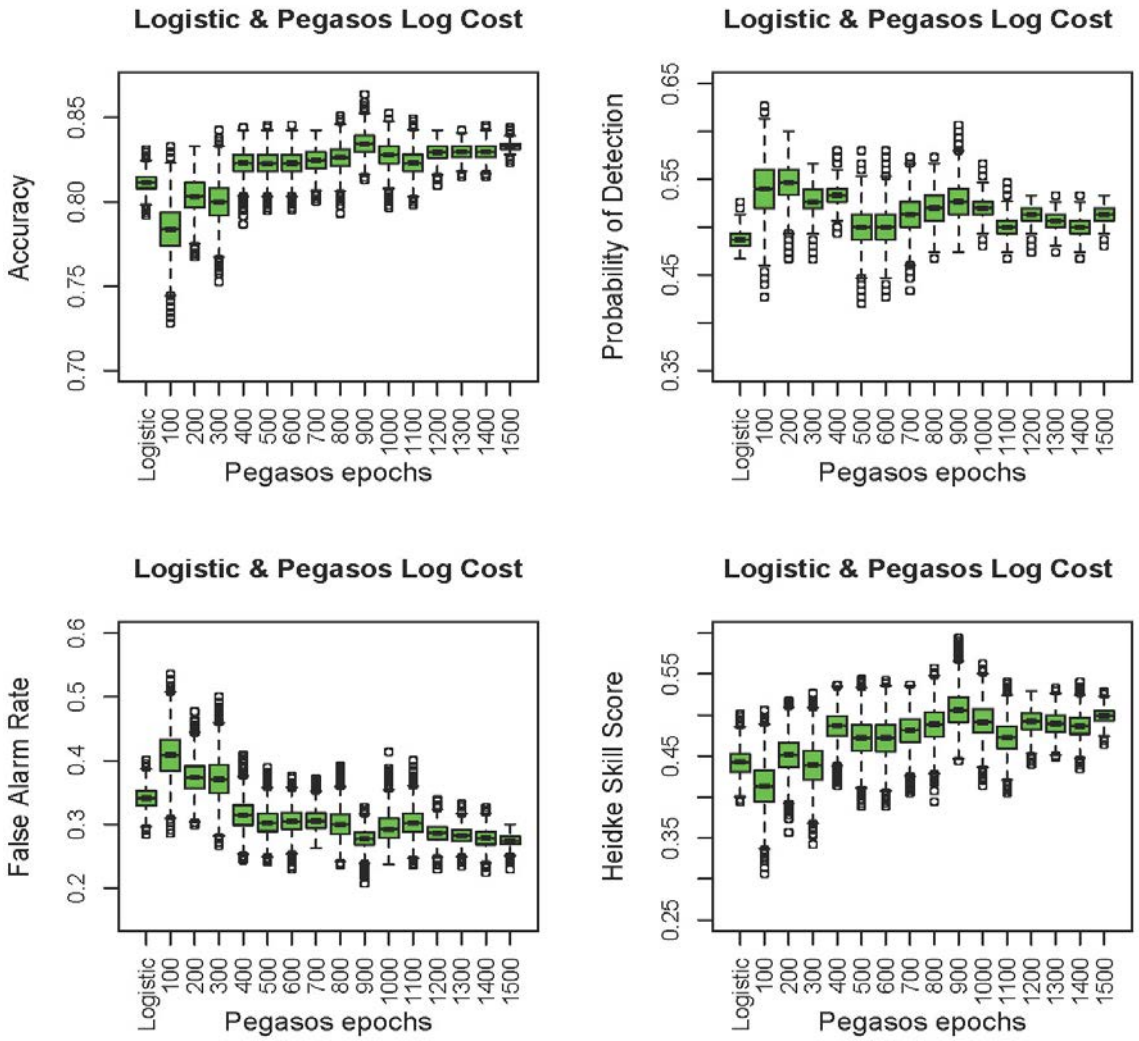

Fig. 5. (a) Boxplots of 5000 replications of 100 random starts for accuracy for logistic regression and as a function of various training epochs for Pegasos with log loss function 10-fold cross-validations with Ethiopian drought targets; (b) same as (a), except for probability of detection; (c) same as (a) except for false alarm rate; (d) same as (a) except for Heidke skill score (kappa).

that Pegasos requires ~400 epochs to achieve a value in excess of logistic regression, suggesting Pegasos is undertrained for the 100–300 epochs. The accuracy slowly increases beyond 300 epochs, reaching an optimum at 900 epochs (mean of 0.834) and then decreases slightly. For the probability of detection, the undertrained results have desirable large values (Fig. 5b) but secondary maxima occur at 900 and 1200 epochs (means of 0.527 and 0.513) that exceed the value of logistic regression (mean of 0.487). When viewed along with the false alarm rate, where smaller values are desirable, all epochs ≥800 have values lower than logistic regression (e.g., at epoch 900, the mean is 0.277 vs. 0.341; Fig. 5c). These goodness of the forecast, relative to a reference random guess, is evaluated via the Heidke skill score (kappa). Skill based on training Pegasos for ≥400 epochs exceed that of logistic regression, with a clear maximum skill at 900 epochs (mean of 0.507 vs. 0.443; Fig. 5d). Fig. 5 indicates that Pegasos, with training to 900 epochs, provides an effective classification system for Ethiopian drought. In some cases, the improvement is probably in the sampling error range. For the probability of detection and kappa, the 25th percentile of the 900 epoch solution does not overlap the median at 1500 epochs, which is an indication of improvement over other log loss solutions. Moreover, the interquartile ranges of the kappa boxplot for 900 epochs and that for logistic regression are separated widely (Fig. 5d), a strong indication of a superior Pegasos solution.

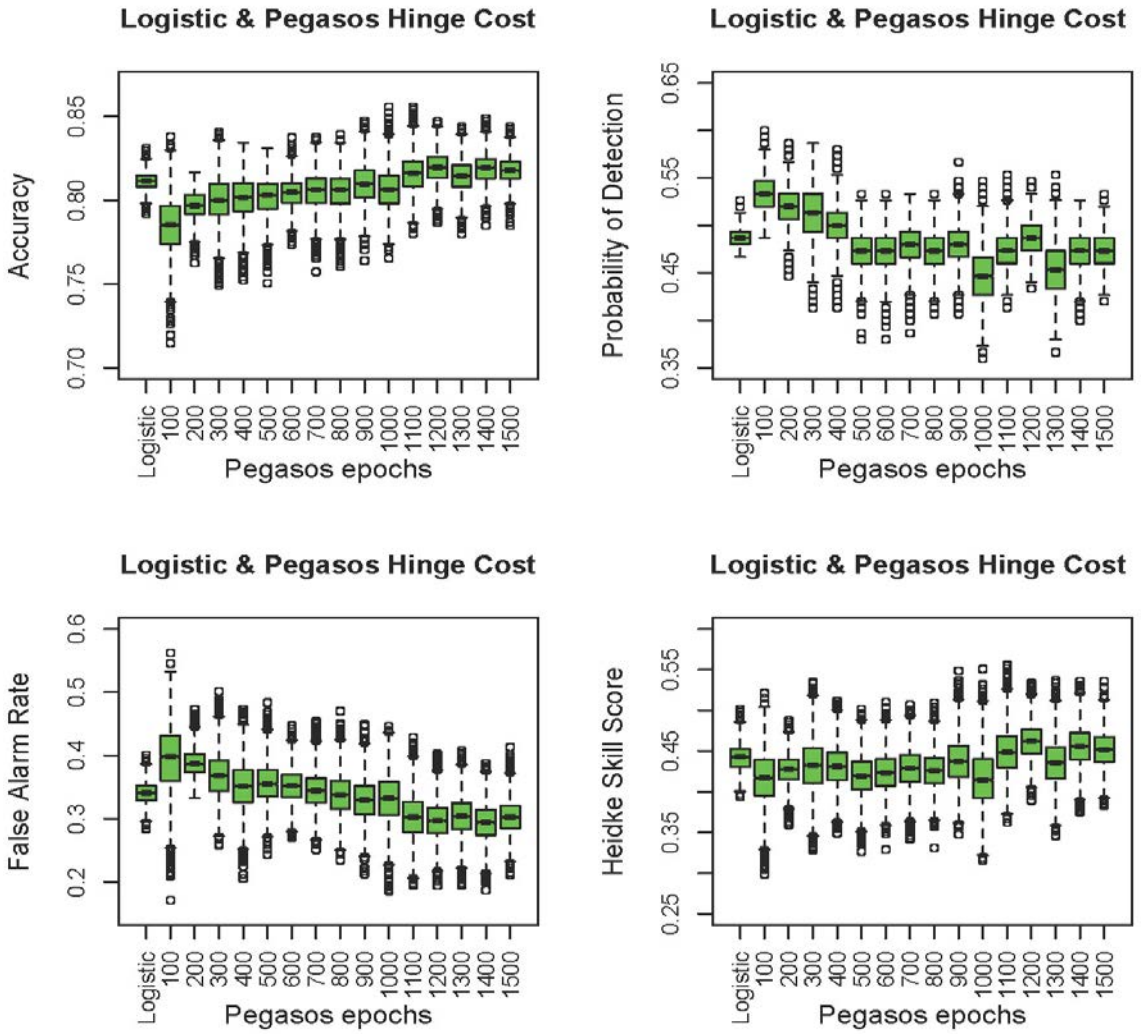

Fig. 6. Same as Figure 5, except for Pegasos with hinge loss function.

Comparing the log loss to the hinge loss (Fig. 6) elucidates the sensitivity of the classification to the loss function. The accuracy of the hinge loss (Fig. 6a) shows that ≥1100 epochs are required to exceed the value for logistic regression. At epoch 1200, the Pegasos accuracy has a mean of 0.820 vs. 0.811 for logistic regression. However, this improvement is unlikely to be statistically significant, given the uncertainty in the solutions shown in the boxplots. For the probability of detection (Fig. 6b), the undertrained models for epochs ≤400 have large values but, as the number of epochs exceeds 400, only epoch 1200 had a value equal to that of the logistic value (both have a mean of 0.487). However, the false alarm rate has values less than that associated with logistic regression (mean of 0.341) for epochs ≥900, with a minimum at 1200 epochs (0.297) (Fig. 6c). The improvement of the forecast over a random guess is seen in the Heidke skill score (Fig. 6d), where epochs ≥1100 have values exceeding logistic regression (e.g., epoch 1200 has a mean value of 0.462 vs. 0.443 for logistic regression). Examining the interquartile range in the statistics, the 1200 epoch Pegasos false alarm rate has a 75th percentile that is smaller than the 25th percentile of the logistic regression solution, an indication of improvement for Pegasos (Fig. 6c). However, for the accuracy, probability or detection and kappa, the interquartile ranges of the two techniques overlap. Taken collectively, the analyses summarized in Fig. 6 suggest that the hinge loss Pegasos model requires ~1100 epochs for a forecast system to achieve the level of accuracy of logistic regression but may offer an advantage of fewer false positives. All three models have more misclassifications with false negatives compared to false positives.

The weights associated with each model were examined to apportion the climate driver signal (Niño 3.4 SSTA) from the precipitation and temperature signal (Table 1). The correlations between these three inputs were small (all ≤ 0.182, Table 2), implying that they are nearly independent. Concentrating on the most accurate log loss Pegasos model, the Niño3.4 SSTA is the least important (11.3% of the total weight), with August temperature nearly twice as large (18.2%) and August precipitation was, by far, the largest (70.5%).

Table 1. Weights (percentage of total weights) assigned to logistic regression and Pegasos with log and hinge loss functions.

Model	August Precipitation	August Temperature	Niño3.4 SSTA
Logistic	-0.060 (3.4%)	1.054 (59.9%)	0.646 (36.7%)
Pegasos (log loss)	-14.678 (70.5%)	3.783 (18.2%)	2.347 (11.3%)
Pegasos (hinge loss)	-12.984 (56.3%)	4.534 (19.7%)	5.527 (24.0%)

Table 2. Correlations between predictors

Predictor	August Precipitation	August Temperature	Niño3.4 SSTA
August Precipitation	1.000	-0.182	0.012
August Temperature		1.000	0.119
Niño3.4 SSTA			1.000

4. Conclusions

Considering that the famine of 1983–1985 killed between 400,000–1,000,000 of 40 million people (or approximately 1–2.5% of the total population), the current multi-year drought in Ethiopia could prove to be even more catastrophic, given the population growth to 90 million since the mid-1980s. This study shows that machine learning techniques can successfully be applied to drought (no drought) classification, in the case of the present and past droughts in Ethiopia. By applying variable selection techniques, 27 variables were reduced to 7. Of these 7 variables, only 3 were required to maximize the skill of the classification using Pegasos and logistic regression. The

variables were August precipitation, August temperature and Niño 3.4 sea surface temperature anomalies relative to the 1981–2010 mean values. These variables must be predicted prior to the growing season to classify correctly drought (no drought) in northern Ethiopia with ~83% accuracy.

Depending on the technique applied, the weights suggest that August temperature plays an important role in the drought (18.2–59.9%). The traditional logistic method gives weights that differ considerably from the newer Pegasos methods. The difference between the two loss functions within Pegasos led to relatively small differences in the weights. The log loss model was most accurate with training of 900 epochs. It had a kappa that improved upon that generated from logistic regression by 14.4%. The weights from the Pegasos methods suggest close to 20% of the drought classification is tied to August temperature. This is consistent with the findings for the role of temperature in the 2011–2015 California drought [11]. Moreover, Pegasos methods identify a significant role for August precipitation and El Niño (56.3–70.5% and 11.3–24.0%, respectively) in drought classification for Ethiopia. As current drought prediction for Ethiopia is based solely on El Niño, this work suggests that the line of advance is improved prediction of August precipitation and temperature with sufficient lead-time to enable action that pre-empts starvation.

5. Acknowledgement

This paper is dedicated to the memory of the late Peter J. Lamb. His enthusiastic mentorship and interest in promoting international collaboration motivated numerous scientists on five continents to undertake research to help Africans help themselves.

6. References

1. de Waal, A. Evil Days: Thirty Years of War and Famine in Ethiopia. New York & London: Human Rights Watch 1991.
2. Harris, I., Jones, P.D., Osborn, T.J. and Lister, D.H. Updated high-resolution grids of monthly climatic observations – the CRI TS3.10 dataset. *International J. Climatology* 2014; **34**, 623-642.
3. Haile, T. Causes and characteristics of drought in Ethiopia. *Ethiopian Journal of Agricultural Sciences* 1988; **10**, 85-97.
4. Segele, Z.T., Lamb P.J. and Leslie, L.M. Seasonal-to-interannual variability of Ethiopia/Horn of Africa monsoon. Part I: associations of wavelet-filtered large-scale atmospheric circulation and global sea surface temperature. *J. Climate* 2009; **22**, 3396-3421.
5. Torrence, C., and G. P. Compo, G.P. A practical guide to wavelet analysis. *Bull. Amer. Meteor. Soc.* 1998; **79**, 61–78.
6. Chapa, S. R., V. B. Rao, V.B. and. Prasad, G.S.S.D. Application of wavelet transform to Meteosat-derived cold cloud index data over South America. *Mon. Wea. Rev.* 1998; **126**, 2466–2481.
7. Huang, N. E., and Coauthors. The empirical mode decomposition and the Hilbert Spectrum for nonlinear and non-stationary time series analysis. *Proc. Roy. Soc. London* 1998; **A454**, 903–995.
8. Yang, S., Ding, X., Zheng, D. and Li, Q. Depiction of the variations of Great Plains precipitation and its relationship with tropical central-eastern Pacific SST. *J. Appl. Meteor. Climatol*. 2007; **46**, 136–153.
9. Goldberg, D.E. Genetic algorithms in search, optimization and machine learning. Addison-Wesley. 1989
10. Shalev-Shwartz, S., Singer, Y., Srebro, N. and Cotter, A. Pegasos: primal estimated sub-gradient solver for SVM. *Mathematical Programming*. 2011; **12**, 3-30.
11. Richman, M.B. and Leslie, L.M. Uniqueness and causes of the California drought. *Procedia Computer Science* 2015; **61**, 428-435.