Original Research Article

Simulation of Meteorological Drought of Bankura District, West Bengal: Comparative Study between Exponential Smoothing and Machine Learning Procedures

Abstract

Drought is one of the slow-onset natural disasters affect the quality of environment and environmental factors altogeather. Simulation of drought is needed for proper planning and management of water resources. This study has been developed using the following five key points: a) primarily from rainfall SPI, PN, DI, RAI, CZI and Z-score are estimated on yearly basis, those indices are added and a new index standardized total drought (S_d) has been established. b) Considering S_d as the input parameter a comparative assessment has been made between 4 individual models (3 models from exponential smoothing, 1 model from machine learning) in simulation and prediction of drought status in next 18 time steps (years) in Bankura District and Winexpo model outperforms the other models as it obtains minimized SE, RMSE, MAE, and MAPE and highest R^2 value. c) The cumulative drought proneness of the region is also assessed and it is found that the whole district will be drought prone within the year 2100. This study establishes a comparative study between Exponential smoothing and machine-learning procedures and also introduces a new combined index standardized total drought.

Keywords: Simulation; Meteorological drought; Winexpo.

1. Introduction:

Drought is one of the natural disasters that human being has been suffering since the ancient era (Wu et. al 2011, Zarch and Amin 2015) and it is the costliest (Wilhite 2000), long-lasting most severe natural hazard (Mishra and Singh, 2010). It is recurrent natural phenomena associated with the lack of water resources for a prolonged period of dryness can occur in arid, semi-arid and rain-forested region (Mishra et. al 2007, Mishra and Desai 2011, Wilhite and Glantz 1985, Abdourahamane 2018) however confusion and debates among scholars prove that there are no universal accepted definitions of drought. Drought forecasting is a critical element in drought risk management (Ozger et al 2012). Meteorological drought that transforms in a hydrological, agricultural and socio-economic events, onsets with a marked reduction in rainfall sufficient to trigger hydro-meteorological imbalance for a prolonged period (Wilhite and Hayes, 1998; Mishra and Singh, 2010).Thus drought monitoring and assessment are hot topics among hydrologists and meteorologists and attract world-wide attention (Yu et. al 2014, Jain et. al 2010, Todisco et. al 2013); its' preparedness and mitigation depends upon the large scale drought monitoring and forecasting over a large geographical area (Ozger et al 2012, Wu et. al 2011). Many drought forecasting models already developed in the field of civil engineering. Mishra and Desai (2006) developed ARIMA and multiplicative seasonal ARIMA models to forecast drought using SPI series. These models are able to simulate drought up to 2 months lead time. Morid et.al 2007 simulated Effective Drought Index (EDI) and SPI using Artificial Neural Network (ANN). Mishra and Desai (2007) compared linear stochastic models with recursive multistep neural network model to the 6 months lead time. Barros and Bowden (2008) employed self-organizing maps (SOM) and multivariate linear regression analysis to forecast SPI of Murray Darling basin of Australia in 12 months of forthcoming scenarios. Many scholars worldwide tested SVM in climatological and hydrological applications (Dibike et. al 2001, Asefa et. al 2004, Tripathi et. al 2006, Wang et. al 2008). Hastie et. al 2008 used Support Vector machine for prediction of drought in eastern Australia. Belayneh and Adamowski in 2012 forecasted meteorological drought using neural network, wavelet neural network and SVM. Exponential smoothing is quite new in this field originally developed in the field of business mathematics in 1960. Exponential smoothing is able to simulate drought in a long term time frame. This study attempts to simulate drought using exponential smoothing in a long-term time frame.

2. Study Area and Background Information

The District Bankura is bounded by latitude $22^0 38'$ N to 23°38´ N and longitude $86^0 36'$ E to $87^0 47'$E covering an area of 6,882 square Kilometers (2,657sq. mile). River Damodar creates the north and north-east boundary of the district. The neighboring districts are Bardhaman in the north, Paschim Medinapore in the south, Hoogly in the east and Purulia in the west (Figure 1).

Bankura is located in the south western central part of the State of West Bengal belonging transition zone between the plains of Bengal on the east and Chhota Nagpur plateau on the West (District Statistical Handbook, 2014). It is a part of Midnapur Division of the State and a part of "Rarh" region thus can be stated as "rarh in Bengal' (Nag and Ghosh 2013a, bankura.gov.in). The areas to the east and north-east are rather flat belonging to the low lying alluvial plains, known as rice bowl of Bengal (bankura.gov.in, Disaster Management Plan of Bankura District 2017, Bankura.gov.in, Nag and Ghosh 2013b).

3. Data sets and Methodology

Figure 2 constructively describes the methodological overview of this paper. 1901 to 2017 monthly rainfall data has been used for overall analysis and 1901 to 1978 data obtained from Govt. of India water portal website mentioned in table 1. From 1979 to 2014 daily station wise rainfall data obtained from National Centres for Environmental Protection (NCEP) official website (https://globalweather.tamu.edu/). 2015, 2016 and 2017 yearly rainfall data were collected from Disaster Management Plan of Bankura District 2017 published by District Disaster Management Cell (Table 1). We got 6 individual rainfall stations available for Bankura District and monthly and daily rainfall data have been added to get yearly rainfall trend. Thus 117 years are taken into consideration.

3.1 Formation of Standardized Total Drought (S_d)

There are several indices developed to assess meteorological drought but the most common are SPI, DI, PN, Z-Score, RAI and CZI (Chen et. al 2009). First of all, from the rainfall data, the above mentioned 6 well-known indices i.e. SPI, DI, CZI, PN, Z-score, and RAI have been estimated on yearly basis and later those are combined and formed a new Index Standardized Total Drought (S_d). So, those six indices are utilized to estimate the true nature of meteorological drought and standardized total drought (yearly basis) becomes the sole input variable for every models of our study.

It can be computed as follows:

$$\text{Total Drought}(T_d) = (\text{SPI} + \text{DI} + \text{PN} + \text{ZScore} + \text{RAI} + \text{CZI}) \quad (1)$$

$$\text{Standardized Total Drought}(S_d) = \frac{T_d - \overline{T_d}}{\delta} \quad (2)$$

Where, T_d is the total drought.

$\overline{T_d}$ is the mean of T_d

δ is the standard deviation of the total drought.

Based on estimated S_d values the individual drought categories are subdivided into 9 sub-groups (table 3). The whole subgroups are ranging between <-10 to >10 category and <-10 denotes the most extreme category whereas >10 denotes wet category. Every 9 sub categories are coded as 1 to 9 (table 3).

3.2 Exponential and Holt-Winter Forecast and Winexpo Method:

Exponential smoothing is the technique to smoothing the time series in exponential window function. Exponential smoothing assigns decreasing weights over time. Holt in 1957 and Winter in 1960 developed smoothing technique and later their method was combined and making Holt-Winter smoothing technique to forecast the recursive trend from the historically observed data series (https://otexts.org/fpp2/holt-winters.html,). Here we use the single exponential smoothing technique as Kaleker in 2004 used in his thesis:

$$S_{t+1} = \alpha * y_t + (1-\alpha) * S_t \qquad 0<\alpha<1,\ t>0 \tag{3}$$

Eq. (11) can be written as

$$S_{t+1} - S_t = \alpha * €_t \tag{4}$$

The Holt-Winter method time series can be represented using the following model:

$$y_t = (b_1 + b_2 t) * S_t + €_t \tag{5}$$

Where b_1 is the permanent component, b_2 is the linear trend component, S_t is the multiplicative seasonal factor, $€_t$ is the random error component, t is the time and t+1 is the lead time from t.

From the Eq. (13)

$$S_t = \frac{y_t}{b_1+b_2 t} + €_t \tag{6}$$

Sum of all the seasons can be written as

$$\sum_{t=12} S_t = M \tag{7}$$

Where L is the length of the year.

So, the Eq. (7) can be written as,

$$\sum_{t=1}^{12} y_t = (b_1 + b_2 \sum_{t=1}^{12} t) * \sum_{t=1}^{12} S_t + €_t \tag{8}$$

Assuming, $\sum_{t=1}^{12} y_t = Y$, $\sum_{t=1}^{12} t = T$ and $\sum_{t=1}^{12} S_t = M$ we get from Eq. (16)

$$Y_t = (b_1 + b_2 T) * M + €_t \tag{9}$$

And Eq. (14) can be written after the sum of all the seasons

$$M = \frac{Y_t - €_t}{b_1 + b_2 T} \tag{10}$$

Winexpo method has been developed by us to combine the traditional exponential and Holt-Winter method. Combining Eq. (12) and Eq. (18) we get,

$$\frac{S_{t+1} - S_t}{M} = \frac{\alpha * €_t}{\frac{Y_t - €_t}{b_1 + b_2 T}} \tag{11}$$

Or, $$\frac{S_{t+1} - S_t}{M} = \frac{\alpha * (b_1 + b_2 T)}{(Y_t - €_t)} + €_t \tag{12}$$

Winexpo is one of the integrative models as it holds the combination of Holt-Winter exponential smoothing and traditional exponential smoothing.

3.4 Support Vector Machine model (SVM)

Support Vector Machine (SVM) is the supervised learning models that analyse data used for classification and regression analysis (Cortes et. al 1995, Vapnik and Vapnik 1998, Vapnik and Cortes 1995). The x related all points can be mapped in the hyperplane can be defined by the relation $\sum_i \alpha_i k(x_{i,} x) = \text{constant}$ where k (x_i, x) is the kernel function used to suit the problem. Kernel function becomes small where y grows further away from x so it becomes the matter of closeness of each point of y to x. With the kernel function SVM actually use the relative closeness between the each point in the feature space. The detailed method of analysis can be expressed as following:

Suppose our training data is consist of N pairs $(X_{1,} Y_1)$, (X_2, Y_2)………….. (Xn, Yn); where Xi € R^p and Y_i € {-1, 1}. Define a hyperplane by, {x: $f(x) = x^T\beta + \beta_0 = 0$}, where β is a unit vector. A classification rule induced by f(x) is G(x) = sign $\{x^T\beta + \beta_0\}$. Now the signed distance from the point x to the hyperplane is 0. Here we are able to find the hyperplane that creates biggest margin between training points for class 1 and -1. So, the optimization problem just reverses and forms the following dimension:

$$\max_{\beta,\beta_{0,\|\beta\|}=1} = M \tag{13}$$

Subject to,

subject to, $$y_i\{x^T\beta + \beta_0\} \geq M \;\; ; \; i = 1,2, \dots \dots \dots .., N \tag{14}$$

We have used here Least Square Support Vector Machine is based on structural risk minimisation (Vapnik and Vapnik, 1998,) in the model weight. It counters convex quadratic programming associated with Support Vector Machine (SVM) (Suykens et. al 1999, Suykens and Vandawalle, 1999). The least square version of the SVM classifier is obtained by reformulating the minimization problem as

$$\min J_2\,(w, b, e) = \frac{\mu}{2} x^T \beta + \frac{\infty}{2} \sum_{i=1}^{N} e_i{}^2$$

Subject to equality constraints,

$$y_i\left[x^T\beta\ +\ \beta_0\right] = 1 - e_i,\ i=1,2,\ldots\ldots\ldots n \tag{15}$$

Eq. 36 can be written as

$$e_i = 1 - y_i\left[x^T\beta\ +\ \beta_0\right] \tag{16}$$

The eq. 37 hold the case of regression. To solve the eq. 37 we use Lagrangian multiplier by which it can be solved.

$$L_2(w, \beta, e, \alpha) = J_2(w, e) - \sum_{i=1}^{n} \alpha_i \left\{[\beta\ +\ \beta_0] + e_i - y_i\right\} \tag{17}$$

Where, α_i€ R , the Lagrangian multipliers. For evaluation performance test of SVM we use the error estimation and Kappa Coefficient statistic as well as the accuracy. The definition of Cohen's Kappa is as follows (Galton 1892, Smeeton 1985):

$$k = \frac{p_o - p_e}{1 - p_e} \tag{18}$$

Where, P_0 is the relative observed agreement among variables; Pe is the hypothetical probability of chance agreement. If the rates are in the complete agreement then k =1 and if there is no agreement then k = 0.

3.7 Estimation of Cumulative Hazard Proneness:

To estimate the cumulative drought-proneness of the study region over the years we took help of the hazard function and survival analysis. Let T be a non-negative random variable representing the waiting time until the occurrence of an event. For simplicity we can adopt the term 'survival analysis' referring to the event of interest as 'hazard proneness' and to the waiting time we state as 'survival time'. We can assume T is a continuous random variable with probability density function (p.d.f.) f(t) and cumulative distribution function (c.d.f.)

Pr $\{k < t\}$ given that probability that the event has occurred by duration t. Complement of c.d.f. the survival function becomes

$$S(t) = \Pr\{T \geq t\} = 1 - F(t) = \int_t^{\infty} f(x)dx \tag{19}$$

Which gives probability of being 'less drought prone' just before duration t more generally the probability that the event of interest has not occurred by duration t. Here we use the following distribution of T is given by hazard function or instantaneous route of occurrence of the event defined as

$$\Omega(t) = \lim_{dt \to 0} \frac{\Pr\{t \leq T < t+dt, T \geq t\}}{dt} = \frac{f(t)}{S(t)} \tag{20}$$

Where f (t) is the derivative of S (t)

$$S_t = \exp\{-\int_0^t \Omega(x)\}\, dx \tag{21}$$

3.9 Error Estimation

3.9.1 Standard Error estimation (SE):

The standard error can be stated as (Hyndman et. al 2006, Makiridakis & Spyros 1993)

$$SE = \frac{\partial}{\sqrt{n}} \tag{22}$$

Where $\partial$the standard deviation of the distribution and n is is the number of samples.

3.9.2 Root of Mean Squared Error (RMSE):

Root of mean squared deviation can be stated as (Hyndman et. al 2006, Anderson & Woessner 1992)

$$RMSE = \frac{\sqrt{\sum_{t=1}^{T}(\overline{y_t} - y_t)^2}}{\sqrt{T}} \tag{23}$$

Where, The RMSD of predicted values for $\overline{y_t}$ times t of a regression's dependent variable y_t with variables observed over T times.

3.9.3. Mean Absolute Error (MAE):

MAE measures average magnitude errors in the set of predictions without considering their direction. It is the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight (Strook and Daniel 2011):

$$\mathrm{MAE} = 1/\mathrm{n}\sum_{\mathrm{j}=1}^{\mathrm{n}}\left|\mathrm{y_j} - \overline{\mathrm{y_j}}\right| \quad (24)$$

Where y_j is the observed value and $\overline{y_j}$ is the predicted value.

3.9.4. Mean Absolute Percentage Error (MAPE)

Mean Absolute Percentage Error (MAPE) is a measure of prediction accuracy of a forecasting method of accuracy. MAPE can be stated as (Hyndman and Koehler 2006)

$$\mathrm{MAPE} = \frac{100\%}{\mathrm{n}}\sum_{\mathrm{t}=1}^{\mathrm{n}}\left|\frac{\mathrm{y_t}-\mathrm{F_t}}{\mathrm{y_t}}\right| \quad (25)$$

Where, y_t is the actual value and F_t is the forecasted value.

3.10 Significance test

3.10.1 Anderson-Darling Test:

The Anderson-Darling test is the hypothesized distribution is F, and cumulative distribution is F_n and the formula can be written as

$$\mathrm{A}^2 = \mathrm{n}\int_{-\infty}^{\infty}\frac{(\mathrm{F_n(x)}-\mathrm{F(x)})^2}{\mathrm{F(x)}\left(1-\mathrm{F(x)}\right)}\mathrm{dF(x)} \quad (26)$$

3.10.2 Kolmogorov-Smirnov Test:

Kolmogorov Smirnov test is a nonparametric test of the equality of continuous one dimensional probability distribution with compare of a sample with reference probability distribution (Kolmogorov 1933, Smirnov 1948). Kolmogorov Smirnov test statistic can be expressed as

$$\mathrm{F_n(x)} = 1/\mathrm{n}\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{I}_{[-\infty,\mathrm{x}]}(\mathrm{X_i}) \quad (27)$$

Where $I_{[-\infty,x]}(X_i)$ is the indicator function, equal 1 if $(X_i) \leq x$ and equal to 0 otherwise.

The Kolmogorov-Smirnov statistic of a given cumulative function F(x) is

$$\mathrm{D_n} = \sup_{\mathrm{x}}(\mathrm{F_n x} - \mathrm{F_x})$$

(69)

Where sup is the supremum of the set of distance between the $F_n x$ and F_x. In our case this model has been run at 95% significance level.

3.10.3 Shapiro -Wilk Test

Shapiro and Wilk test of the normality formula can be written as,

$$W = \frac{(\sum_{i=1}^{n} a_i x_i)^{\wedge}2}{\sum_{i=1}^{n}(x_i - \bar{x})^{\wedge}2} \quad (70)$$

a_i is the $(a_1 \ldots\ldots\ldots a_n)$, $\bar{x}$ is the mean.

The constants a_i can be written as $(a_1 \ldots\ldots\ldots a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}$ here $m = (m_1 \ldots\ldots\ldots m_n)^{\wedge}T$

and $m_1 \ldots\ldots\ldots m_n$ are the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution, and V is the covariance matrix of those order statistics.

4. Application and Discussion

Fluctuation of rainfall and a negative exponential trend are specified in Figure 3 (Y_t = 1418.88 × (0.999642^t). Rainfalls are more or less normally distributed at 95% confidence interval (Figure 4a). Residuals versus fit plot (Figure 4b) displays that the points are randomly distributed on both sides of zero with no recognisable patterns thus our rainfall data are having a constant variance. Residuals of rainfall are having a mean close to zero and the histogram is symmetric close to around zero (Figure 4c). Residuals versus order fit (Figure 4d) shows that the residuals fall randomly around the centre line. Before proceed with rainfall and estimated 6 indices the reliability of those 6 indices are judged using Cronbach's Alpha. The overall value of Cronbach's alpha is 0.9694. Average SPI and Z-score between the time frame 1901-1939 are -0.06 and 0.299 , in between 1940 -1980, 0.037 and 0.382 respectively and from 1980-2035 the average SPI and Z-score becomes -2.345. Average PN value from 1901-1939 is 100.792 %, 1940-1980 PN becomes 100.641%; 1980-2035 it is diminished and become 98.967%. In the same way average DI is estimated by us and from 1901-1939 DI 5.76%, 1940 to 1980 5.73% and DI from 1980 to 2035 4.64% value of DI is obtained. CZI and RAI are also decreased from 0.32 (1901-1939) and 0.38 to 0.26 (1940-1980) ,0.28 and later 1980-2035 it reaches to 0.14 and 0.19. Overall all the indices attain negative trend. SPI, PN, DI, RAI, CZI and Z-score are added and a new index Standardized Total Drought (S_d) has been formed to estimate overall trend of meteorological drought of Bankura District. Estimation and prediction of the trend of S_d using the traditional exponential smoothing has been done and a slightly negative trend is obtained (Values reach to -0.143 in 2035) (Figure

5a). The residuals of traditional exponential smoothing trend values are ranging between -15 to +5 (Figure 5b). In case of traditional exponential smoothing the average value between 1901-1939 experiences -0.170, 1940 to 1980 the value reaches to -0.034 whereas between the 1980 to 2035 the average value attains -0.134 thus overall trend is seemed to be more drought prone in recent upcoming periods. Similarly using Holt-Winter exponential smoothing analysis and prediction of drought has been done (Figure 5c) and residuals are fitted randomly as histogram plot based on the centre line (ranging between -2 to +5 range) (Figure 5d). In case of Holt-Winter exponential smoothing the average value between 1901-1939 achieve -0.163, between the time frame 1940-1980 and 1980 to 1935 it attain 0.061 and -0.261 values respectively. The combined model Winexpo attains 0.423 for 1901-1939, 0.51 for 1940-1980 and -1.423 for 1980-2035.

From the true classes determined from the categories of S_d (Table 3) SVM is capable to predict the nature of drought category. A user friendly SVM tool LSSVM is used to implement the classification of drought status of Bankura District. At data pre-processing stage raw values of S_d are linearly rescaled into [-1, 1] using the ranges of their minimums and maximums for binary distribution of classifiers. Applying the SVM each category against all is estimated in every case. In case of Extreme vs. others the model is obtained 43 support vectors, for extreme normal the model is obtained 33 support vectors, for mild drought vs. Others the model obtains 34 support vectors, most extreme vs. Others the model obtains 28 support vectors, normal vs. others obtains 51 support vectors, severe vs. others obtains 8 support vectors and wet vs. others obtains 20 support vectors. From the observed true classes of 135 observations (used simulated value using Winexpo) drought probability classes are predicted by SVM. SVM performs with a medium accuracy level. According to SVM identified drought categories over years over 80% years are concentrated within severe moderate, severe, extreme and most extreme categories and about 20% years are concentrated within Moderate, Normal, and Extreme Normal, wet categories (Figure 6a) whereas according to Winexpo identified drought categories 36% years are mingled with severe moderate, severe, extreme, most extreme and moderate categories and over 64% are mingled with normal, mild, extreme normal and wet categories (Figure 6b). The extreme normal versus others, wet versus others, mild versus others, normal versus others training sample sets achieve over 90% accuracy whereas extreme and most extreme versus others and severe moderate versus others category training samples achieve less than 30% accuracy (Table 4).

Overall average SVM achieve 0.724 as Cohen's Kappa and overall 60% accuracy has been achieved. So, SVM has performed moderately well in prediction of drought of our study area.

The significance test using three individual tests has been run at 95% and 99% confidence interval (Table 5). The traditional exponential smoothing experiences probability value 0.004 for Anderson-Darling test, 0.005 for Shapiro-Wilk test and 0.004 by Kolmogorov-Smirnov test. The Holt-Winter exponential smoothing attains 0.003 probabilities for Anderson-Darling test, 0.004 for Shapiro-Wilk test and 0.001 for Kolmogorov-Smirnov test. Winexpo model also attains probability value 0.002 for Anderson-Darling test, 0.004 for Shapiro-Wilk test and 0.003 for Kolmogorov-Smirnov test.

The Bayesian model of LSSVM extreme category versus others experiences 10.275 as Anderson-Darling test statistic value, 0.527 as Shapiro-Wilk test statistic value and 0.435 as KS test statistic value. LSSVM Bayesian most extreme vs others is mingled with 5.543 as Anderson-Darling test statistic, 0.727 as Shapiro-Wilk test statistic and 0.316 as KS test statistic. SVM extreme normal vs others achieves 2.165 as Anderson-Darling test statistic, 0.904 as Shapiro-Wilk test statistic and 0.482 as KS test statistic value. Similarly, Mild versus others, severe versus others, severe moderate versus others and wet versus others are also calculated (Table 5). All the Anderson –Darling test is successful and valid at 95% confidence interval as the significance level P-value achieves <0.005 value in all the nine combinations. Shapiro-Wilk and KS test for all the SVM nine possible combinations the probability value is <0.010 that means those values are significant at 99% confidence interval. Overall SVM model is significant at 95% confidence interval (in case of Anderson-Darling test) and 99% significance level (in case of Shapiro-Wilk test and KS test). As P values are <0.005 and <0.010 for all the cases the distribution is not normal here and null hypothesis that there were no difference between the observed class and predicted class can be rejected and the alternative hypothesis is accepted. The error estimation and goodness of fit statistics (Table 5) of the individual models indicate that Winexpo attains the lowest error and highest R-square value in comparison with the other models altogether.

Based on the whole aspects of meteorological drought the year wise hazard and cumulative failure functions are developed. The most extreme, extreme, severe, severe moderate, moderate and mild categories are included in the category of "hazard prone or failure "whereas normal, extreme normal and wet categories are included in "censored" category. Winexpo attains the best result so this model has been used here. According to simulation of drought category using winexpo over years (Table 5), almost 84 observations are fallen into "hazard-prone" category and 51 observations have fallen into the "censored" group. We had compared the distribution of yearly censored and failure categories based on Weibull and logistic probability fit but logistic probability fit gave us the better association (Correlation value 0.984 for logistic and 0.678 for Weibull). So, finally the logistic probability fit have been taken for year-wise estimation of cumulative hazard-proneness. The whole logistic model seemed to be more or less normal (Figure 8a and 8b) and it had achieved the 3.223 value as the Anderson-Darling test. From the survival function (Figure 8c) fitted based on logistic probability plot encounters the fact that as the time (year) will progress the drought proneness will increase and at the year 2100 the vulnerability will be almost intolerable that will lead to massive disruption over the local community. Reversely, the progression of hazard based on cumulative curve plotting (Figure 9, figure 8d) exhibits the fact that the whole district will be severely affected by drought within 2100. The significance test for hazard function is done in 95% significance level .So, it can be concluded that the district will face extreme to severe drought hazard in the recent future.

5. Conclusion

The evolution and quantification of drought are necessary for the proper planning and management of water resources to mitigate the hazard of future occurrences (Duan et. al 2014). By far the main challenge in this field is that a) to identify the correct method to analyze the meteorological drought b) to identify the spatial dimension over which the drought can be affected c) to simulate and predict the drought correctly as it is inherently needed for proper planning and management of water resources (Hastie et. al 2008). Continuous year wise monitoring and simulation is also an important issue even seriously neglected in the drought monitoring and assessment (Elhag & Zhang, 2018). In most of the cases of drought monitoring and assessment historical rainfall data is one of the input factors (e.g. Dogan 2012). Our study is also not an exception with the above scenarios. Taking rainfall as the sole input factor we estimated 6 essential meteorological indices and from those indices we form a new index Standardized Total Drought (S_d) and simulate it upto 2035 and make a comparative assessment of exponential smoothing and machine learning procedures. Cumulative drought-proneness of the region using hazard function has been analysed and we found that the whole region will be severely drought affected within 2100. Chatterjee 2018, Das et. al 2013, Khan et. al 2011, Rogaly 2010, Rogaly et. al 2001 also support the fact that Bankura is a historically a drought prone district and if no supportive action taken quickly in this regard the condition will get much severe in the upcoming periods. Lohar and Pal (1995) showed that mean monthly pre-monsoonal rainfall has decreased and temperature has increased significantly in the last decades of twentieth century. The extremities of rainfall and temperature drive a potential threat to agriculture, food security and socio-economic vulnerability. Thus a more detailed structural study is required to explore the synergetic effects of trends and patterns of other climatic variables. However the conclusion reached in this study can be an elementary step to improve the risk management strategy, review of agricultural practices and water use in this counterpart.

Conflict of Interest

There is no conflict of interest regarding the publication of this article.

References

1. Abdourahamane, Z. S., & Acar, R. (2018). Fuzzy Rule-Based Forecast of Meteorological Drought in Western Niger. *Theor Appl Climatol*. doi:10.1007/s00704-017-2365-5

2. Akhtari, R., Morid, S., Mahdian, M.H., Smakhtin, V.(2009). Assessment of Areal Interpolation Methods for Spatial Analysis of SPI and EDI Drought Indices. *Int. J. Climatol*. 29, 135–145.

3. Almedeij, J.(2015). Long-Term Periodic Drought Modeling. *Stochastic Environmental Research and Risk Assessment*, 1-10.

4. Anderson, M.P.; Woessner, W.W. (1992). *Applied Groundwater Modeling: Simulation of Flow and Advective Transport* (2nd ed.). Academic Press.
5. Andrews, J. L. and P. D. McNicholas (2011). Mixtures of Modified T-Factor Analyzers for Model-Based Clustering, Classification, and Discriminant Analysis. *Journal of Statistical Planning and Inference 141*(4), 1479–1486..
6. Asefa, T., Kemblowski, M.W., Urroz, G., McKee, M., Khalil, A. (2004). Support Vectors-Based Groundwater Head Observation Networks Design. Water Resources Research. 40 (11), W11509.doi:10.1029/2004WR003304
7. Banfield, J. D. and A. E. Raftery (1993). Model-based Gaussian and non-Gaussian Clustering. *Biometrics 49*(3), 803–821.
8. bankura.gov.in-official website of Bankura District.

9. Barros, A.P. and G.J. Bowden (2008), Toward Long-Lead Operational Forecasts of Drought: An Experimental Study in the Murray-Darling River Basin. *J Hydrol.*, 357(3-4): 349-367.
10. Barua, S., Ng, A.W.M., Perera, B.J.C., (2011). Comparative Evaluation of Drought Indexes: Case Study on the Yarra River Catchment in Australia. *J. Water Resour. Plann. Manage. – ASCE* ,37, 215–226.

11. Belayneh A., Adamowski J. (2013). Drought Forecasting Using New Machine Learning Methods. *Journal of Water and Land Development.*3-12.
12. Brockwell P.J., Davis R.A. (2002). *Introduction to Time Series Forecasting*. Springer (2nd ed.).. 1-434.
13. Chatterjee U (2018). Water Scarcity in Semi-Arid Regions of Bankura District, West Bengal, India – Problems and Prospects. *Khoj*. 87-96.
14. Chen, S.T., Kuo, C.C., Yu, P.S., (2009). Historical Trends and Variability of Meteorological Droughts in Taiwan. *Hydrol. Sci. J.* 54 (3), 430–441.
15. Cortes, Corinna; Vapnik, Vladimir N. (1995). Support-Vector Networks. *Machine Learning*. 20 (3): 273–297. CiteSeerX 10.1.1.15.9362. DOI:10.1007/BF00994018

16. Dibike, Y.B., Velickov, S., Solomatine, D., Abbott, M.B., (2001). Model Induction with Support Vector Machines: Introduction and Applications. *J. Comput. Civil Eng.* 15(3), 208–216.
17. Disaster Management Plan of Bankura District, (2017). Disaster Management Cell. 1-147.Available at http://www.wbdmd.gov.in/writereaddata/uploaded/DP/ Disaster%20Management%20Plan%20of%20Bankura.pdf.
18. District Statistical Handbook, 2014. Collected from Panchayet Bhawan, Salt-Lake city, Kolkata.
19. Deo, Ravinesh C, Kisi, Ozgur, Singh, Vijay P (2017). Drought Forecasting In Eastern Australia Using Multivariate Adaptive Regression Spline, Least Square Support Vector Machine And M5Tree Model, *Atmospheric Research,* doi:10.1016/j.atmosres.2016.10.004 (accepted manuscript).
20. Dogan, S., Berktay, A., & Singh, V. P. (2012). Comparison of Multi-Monthly Rainfall-Based Drought Severity Indices, With Application to Semi-Arid Konya Closed Basin, Turkey. *J. Hydrol.*, 470-471, 255–268. doi:10.1016/j.jhydrol.2012.09.003.

21. Duan, K., Xiao, W., Mei, Y., & Liu, D. (2014). Multi-Scale Analysis Of Meteorological Drought Risks Based on A Bayesian Interpolation Approach In Huai River Basin, China. *Stoch. Environ. Res. Risk A.,* 28(8*), 1985–1998.*doi:10.1007/s00477-014-0877-4
22. Durdu, O.F (2010). Application of Linear Stochastic Models for Drought Forecasting in the Buyuk Menderes River Basin, Western Turkey. *Stoch. Environ. Res. Risk A*. 24, 1145–1162.
23. Edwards, D.C., McKee, T.B., 1997. Characteristics of 20th Century Drought in the United States at Multiple Time Scales. *Atmos. Sci. Paper* .634, 1–30.
24. Elhag K,M., Zhang W. (2018). Monitoring and Assessment of Drought Focused on Its Impact on Sorghum Yield over Sudan by Using Meteorological Drought Indices for the Period 2001–2011. *Remote Sens.* 1-21. doi:10.3390/rs10081231.
25. Freitas Mas (2005). Um Sistema De Suporte À Decisão Para O Monitoramento De Secas Meteorológicas Em Regiões Semiáridas. *Rev. Tecnol.* ; 19: 84-95.
26. Galton, F. (1892). *Finger Prints* Macmillan, London
27. Gibbs, W.J., Maher, J.V., (1967). Rainfall Deciles as Drought Indicators. *Bureau of Meteorology, Bulletin No. 48, Melbourne, Australia*.
28. Hayes, M.J., (2000). Drought Indices. National Drought Mitigation Center, University of Nebraska, Lincoln, Nebraska, USA.
29. Hayes, M.J., Svoboda, M.D., Wilhite, D.A., Vanyarkho, O.V.(1999).Monitoring The 1996 Drought Using The Standardized Precipitation Index *Bulletin of the American Meteorological Society* 80, 429-438.
30. https://otexts.org/fpp2/holt-winters.html.
31. Hyndman, Rob J., & Koehler Anne. B. (2006). Another Look at Measures of Forecast Accuracy. *Int. J. Forecast*. 22 (4). 679-688.
32. Human Development Report of Bankura District. 2007-08.
33. Jain S.K., Keshri R., Goswami A., Sarkar A.(2010) Application Of Meteorological And Vegetation Indices For Evaluation Of Drought Impact: A Case Study For Rajasthan, India. *Natural Hazards*, 54 (3), 643.
34. Kalekar PS (2004). Time-Series Forecasting Using Holt-Winters Exponential Smoothing. *Kanwal Rekhi School of Information Technology*. 1-13. Available at: https://labs.omniti.com/people/jesus/papers/holtwinters.pdf.
35. Keyantash, J., Dracup, J.A., (2002). The Quantification of Drought: An Evaluation of Drought Indices. Bull. *Am. Meteorol. Soc*. 83, 1167–1180.

36. Khan J.H., Hassan T. and Shamsad (2011). Socio Economic causes of Rural Urban Migration in India. *Asia-Pacific Journal of Social Sciences*. 138-158.

37. Kolmogorov A. (1933). Sulla Determinazione Empirica Di Una Legge Di Distribuzione. G. *Ist. Ital. Attuari*. 4: 83–91.

38. Lana, X., Burgueno, A.(2000). Statistical Distribution and Spectral Analysis of Rainfall Anomalies for Barcelona (NE Spain). *Theor. Appl. Climatol*. 66, 211–227..

39. Makridakis, Spyros (1993). Accuracy Measures: Theoretical and Practical Concerns. *Int. J. Forecast*. 9 (4): 527-529.

40. McKee, T.B., Doesken, N.J., Kleist, J (1993). The Relationship of Drought Frequency and Duration to Time Scales. *Proceedings of the 8th Conference on Applied Climatology*, American Meteorological Society Boston, MA , 179-183.

41. Mishra A.K., Desai V.R (2006). Drought Forecasting using Feed-forward Recursive Neural Network. *Ecological Modeling*. pp. 127-138.

42. Mishra A.K., V. R. Desai, and V. P. Singh, (2007): Drought Forecasting Using a Hybrid Stochastic and Neural Network Model. *J. Hydrol*. Eng., 12, 626–638.

43. Mishra, A.K., Singh, V.P. (2010).A Review of Drought Concepts. *J. Hydrol.* 391, 202-216.

44. Moghaddasi R., Eghbali A., Rizi P.L. (2014). Analysis and Forecasting of Drought by Developing a Fuzzy-Based Hybrid Index in Iran. *MPRA*. 1-15.

45. Morid, S., V. Smakhtin, and K. Bagherzadeh, (2007): Drought Forecasting Using Artificial Neural Networks and Time Series of Drought Indices. *Int. J. Climatol*., 27, 2103–2111.

46. Mpelasoka, F., Hennesy, K., Jones, R., Bates, B. (2008). Comparison of Suitable Drought Indices For Climate Change Impacts Assessment Over Australia Towards Resource Management. *Int. J. Climatol*. 28, 1283–1292.

47. Nag S.K, Ghosh P. (2013a). Delineation of Groundwater Potential Zone in Chhatna Block, Bankura District, West Bengal, India Using Remote Sensing And GIS Techniques. *Environmental Earth Sciences*. 70(5). 2115-2127.

48. Nag S.K, Ghosh P. (2013b). Variation in Groundwater Levels and Water Quality in Chhatna Block, Bankura District, West Bengal — A GIS Approach. *Journal of Geological Society of India.*81 (2). 261-280.

49. Özger, Mehmet, Ashok K. Mishra, and Vijay P. Singh. (2012) "Long Lead Time Drought Forecasting Using a Wavelet and Fuzzy Logic Combination Model: A Case Study in Texas." *J Hydro. Meteorol.* 13(1): 284–97. https://doi.org/10.1175/JHM-D-*10-05007.1.*

50. Pandey, R.P., Dash, B.B., Mishra, S.K., Singh, R., (2008). Study Of Indices For Drought Characterization in KBK Districts In Orissa (India). *Hydrol. Process.* 22, 1895–1907.

51. Rogaly B, Biswas J, Coppard D, Rafique A, Rana K and Sengupta (2001) Seasonal Migration, Social Change and Migrants' Rights: Lessons from West Bengal . *Economic and Political Weekly.* 36(49) , 8-14, 4547-4559.

52. Rogaly, B. (2010). Workers on the move: Seasonal Migration and Changing Social Relations in Rural India. *Gender & Development*, 6(1), 21–29. doi:10.1080/741922628.

53. Rooy M.P , Van (1965). A Rainfall Anomaly Index Independent of Time and Space. *Notos.*; 14, 43p.

54. Smeeton, N.C. (1985). Early History of the Kappa Statistic. *Biometrics*. 41: 795. JSTOR 2531300

55. Smirnov N (1948). Table For Estimating The Goodness Of Fit Of Empirical Distributions. *Annals of Mathematical Statistics*. 19: 279– 281. doi:10.1214/aoms/1177730256

56. Suykens, J.A.K., Lukas, L., Van Dooren, P., De Moor, B., Vandewalle (1999), J. Least Squares Support Vector Machine Classifiers: A Large Scale Algorithm. *European Conference on Circuit Theory and Design, ECCTD*, Citeseer , 839-842.

57. Suykens, J.A.K., Vandewalle, J.(1999).Least squares support vector machine classifiers *Neural processing letters* 9, 293-300.

58. Todisco F., Mannocchi F., Vergni L (2013). Severity duration-Frequency Curves In the Mitigation of Drought Impact: An Agricultural Case Study. *Natural Hazards*, 65 (3), 1863.

59. Tripathi, Sh., V. V. Srinivas, R. S. Nanjundiah. (2006). Downscaling Of Precipitation For Climate Change Scenarios: A Support Vector Machine Approach. *J Hydrol*. 330, 621– 640.

60. Turkes, M., Tatli, H., (2009). Use of the Standardized Precipitation Index (SPI) and A Modified SPI For Shaping The Drought Probabilities Over Turkey. *Int. J. Climatol.* 29, 2270–2282.
61. Vapnik, V.N., Vapnik, V. (1998). *Statistical Learning Theory*. Wiley New York.
62. Vapnik, V. N., Cortes, C., 1995. Support vector networks. *Machine Learning* 20, 273–297.
63. Vrbik, I. and P. D. McNicholas (2014). Parsimonious Skew Mixture Models for Model-Based Clustering and Classification. *Computational Statistics and Data Analysis 71*, 196–210.
64. Vicente-Serrano, S.M., Gonzalez-Hidalgo, J.C., De Luis, M., Raventos, J., (2004). Drought Patterns in The Mediterranean Area: The Valencia Region (Eastern Spain). *Climate Res.* 26, 5–15.
65. Wang, W. C. Men, W. Lu.(2008). Online Prediction Model Based On Support Vector Machine. *Neurocomputing*, 71: 550-558.
66. Wang F.Q., Zheng Z., Kang P.P., Wang L. (2016). Applicability Evaluation On The Indexes Of Typical Drought In Henan Province, China. *Applied Ecology and Environmental Research*. 253-262.
67. Wilhite, D. (Ed.), (2000). *Drought: A Global Assessment, vols. I &II. Routledge Hazards and Disasters Series*, Routledge, London.
68. Wilhite, D.A., Hayes, M.J. (1998). Drought planning in the United States: Status and future directions. *The Arid Frontier*, Springer, 33-54.
69. Wilson, E.B., Hilferty, M.M., (1931). The Distribution of Chi-Square. *Proc. Nat. Acad. Sci*. USA 17, 684–688.
70. Wu, H., Hayes, M.J., Weiss, A., Hu, Q., (2001). An Evaluation of the Standardized Precipitation Index, the China-Z Index and the Statistical Z-Score. *Int. J. Climatol*. 21, 745–758.
71. Wu Z.Y., Lu G.H., Wen L., Lin C.A. (2011). Reconstructing and Analyzing China's fifty-nine Year (1951–2009) Drought History using Hydrological Model Simulation. *Hydrol. Earth. Syst. Sci*. 15. 2881-2894. doi: 10.5194/hess-15-2881-2894.
72. Support Vector Machine (SVM). *International Conference on Drought Management Strategies in Arid and Semi Arid Regions*. 1-16.

73. Zarch A., Amin M.(2015), Droughts In A Warming Climate: A Global Assessment Of Standardized Precipitation Index (SPI) and Reconnaissance Drought Index (RDI). *J. Hydrol* . 526. 183-195.
74. Zhai, J.Q., Su, B.D., Krysanova, V., Vetter, T., Gao, C., Jiang, T., (2010). Spatial Variation and Trends in PDSI and SPI Indices and Their Relation to Stream-flow in 10 Large Regions of China. *J. Climate* 23 (3), 649–663.
75. Zhang, G. P. (2003). Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model. *Neurocomputing*, 50, 159–175. doi:10.1016/s0925-2312(01)00702-0.

Table 1 Source of Rainfall Data

<table>
<tr><th colspan="4">Monthly Rainfall Data Station-wise 1979-2014 downloaded from NCEP data set (https://globalweather.tamu.edu/)</th><th></th></tr>
<tr><th>Id of Stations associated Bankura</th><th>Longitude</th><th>Latitude</th><th>Elevation(m)</th><td rowspan="7">Monthly total rainfall data downloaded from 1901-1978 from Indian Water Portal (www.Indianwaterportal.org.) and 2015,2016 and 2017 rainfall data obtained from Disaster Management Plan 2017 of Bankura district</td></tr>
<tr><td>229869</td><td>86.875</td><td>22.9488</td><td>133</td></tr>
<tr><td>229872</td><td>87.1875</td><td>22.9488</td><td>61</td></tr>
<tr><td>229875</td><td>87.5</td><td>22.9488</td><td>34</td></tr>
<tr><td>233869</td><td>86.875</td><td>23.261</td><td>127</td></tr>
<tr><td>233872</td><td>87.1875</td><td>23.261</td><td>95</td></tr>
<tr><td>233875</td><td>87.5</td><td>23.261</td><td>46</td></tr>
</table>

Table 2 Classes of Drought Indices

Drought Indicators	Extreme Drought	Severe Drought	Moderate Drought	Normal	Moderately Wet	Very Wet	Extreme Wet
SPI (Mckee et. al 1993)	<-2.00	-1.99 to -1.50	-1.49 to -1.00	-0.99 to 0.99	1.0 to 1.49	1.5 to 1.99	>2
PN (Dogan et. al 2012)	<40	40-55	55-80	80-100	>100		
DI (Mpelasoka et. al 2008)	<1	1-2	2-3	>3			
RAI (Rooy 1965, Freitas 2005)	<-3.0	-2.1 to -3.0	-1.2 to -2.1	-0.3 to -1.2		0.3 to -0.3	>0.3
CZI (Wilson and Hilferty 1931)	<-2	-1.5 to -1.99	-1.0 to -1.49	-0.99 to 0.99	1.0 to 1.49	1.5 to 1.99	>2
Z Score (Dogan et. al 2012)	<-1.25		-0.84 to -1.25	-0.52 to -0.84	-0.25 to -0.52	0.25 to -0.25	>0.25

Training set	Accuracy***	Cohen's kappa**
Extreme versus Others	0.847	0.978
Extreme Normal versus Others	0.187	0.086
Moderate versus Others	0.987	0.987
Most Extreme versus Others	0.847	0.978
Normal versus Others	0.253	0.222
Severe versus Others	0.987	0.998
Severe Moderate	0.876	0.965

versus Others		
Wet versus Others	0.153	0.042
Mild versus Others	0.165	0.078

Table 3 Probable classes of Standardized Total Drought (Sd)

Table 4 Performance matrix of Support Vector Machine (SVM)

Categories of Drought	**Code**	**Ranges of Drought**
Most Extreme	1	<-10.00
Extreme	2	-3.00 to -10.00
Severe	3	-2.99 to -2.50
Severe Moderate	4	-2.49 to -2.35
Moderate	5	-2.35 to -1.15
mild drought	6	-1.15 to 1
Normal	7	1-5
Extreme Normal	8	5-10
Wet	9	>10

Table 5 Error Estimation and Goodness of fit statistics (for error estimation 0.001 used as a multiplicative factor)

Model Name	**SE**	**Adjusted RMSE**	**Adjusted MAE**	**Adjusted MAPE**	**R^2(using Linear kernel)**
Traditional exponential smoothing**	0.024	0.996	0.790	25.65	0.39
Holt-Winter Smoothing**	0.026	1.006	0.654	95.43	0.04
Winexpo Model**	0.111	1.64	0.445	49.53	0.35
SVM-Most Extreme versus others	3.080	0.049	0.045	4.559	0.99
SVM-Extreme	1.303	0.038	0.019	2.048	0.94

versus others					
SVM-Severe versus others	11.180	0.026	0.026	1.915	0.95
SVM-Severe moderate versus others	11.345	0.023	0.045	1.934	0.99
SVM-Moderate versus others	5.533	0.015	0.008	0.833	0.99
SVM-Mild versus others	5.333	0.020	0.013	1.413	0.97
SVM-Normal versus others	1.668	0.033	0.019	2.048	0.52
SVM-Extreme Normal versus others	7.580	0.018	0.014	1.487	0.35
SVM-Wet versus others	83.724	0.001	0.008	0.900	0.34
Overall SVM versus other	0.130	0.02175	0.022	1.904	0.78

Table 6 Significance test of the models

Standardized Total Drought	**Anderson-Darling Test**		**Shapiro-Wilk Test**		**Kolmogorov-Smirnov Test**		**Type of Model**
	Test Statistic	**Significance Level**	**Test Statistic**	**Significance Level**	**Test Statistic**	**Significance Level**	
Traditional Exponential Smoothing	8.827	0.004 (<0.005)	0.916	0.005 (<0.05)	0.169	0.004 (<0.005)	Exponential Smoothing
Holt-Winter Exponential Smoothing	7.192	0.003 (<0.005)	0.917	0.004 (<0.005)	0.163	0.001 (<0.005)	
Winexpo Model	28.790	0.002 (<0.005)	0.529	0.004 (<0.005)	0.363	0.002 (<0.005)	Combined model
SVM-Extreme versus others**	10.275	<0.005	0.527	<0.010	0.435	<0.010	

SVM-Extreme normal versus others**	2.165	<0.005	0.904	<0.010	0.482	<0.010	
SVM-Mild vs others**	11.598	<0.005	0.482	<0.010	0.419	<0.010	Machine Learning
SVM-Moderate vs others**	10.550	<0.005	0.455	<0.010	0.427	<0.010	
SVM-Most Extreme vs others**	5.543	<0.005	0.727	<0.010	0.316	<0.010	
SVM-Normal vs. others**	5.274	<0.005	0.827	<0.010	0.261	<0.010	
SVM-Severe vs. others**	5.544	<0.005	0.597	<0.010	0.466	<0.010	
SVM-Severe moderate_vs_others**	2.131	<0.005	0.662	<0.010	0.462	<0.010	
SVM-Wet vs. Others**	1.108	<0.005	0.935	<0.05	0.236	<0.010	

**** Calculation done using rescaled vector, others indicate other categories combined, eq. Denotes equation.**

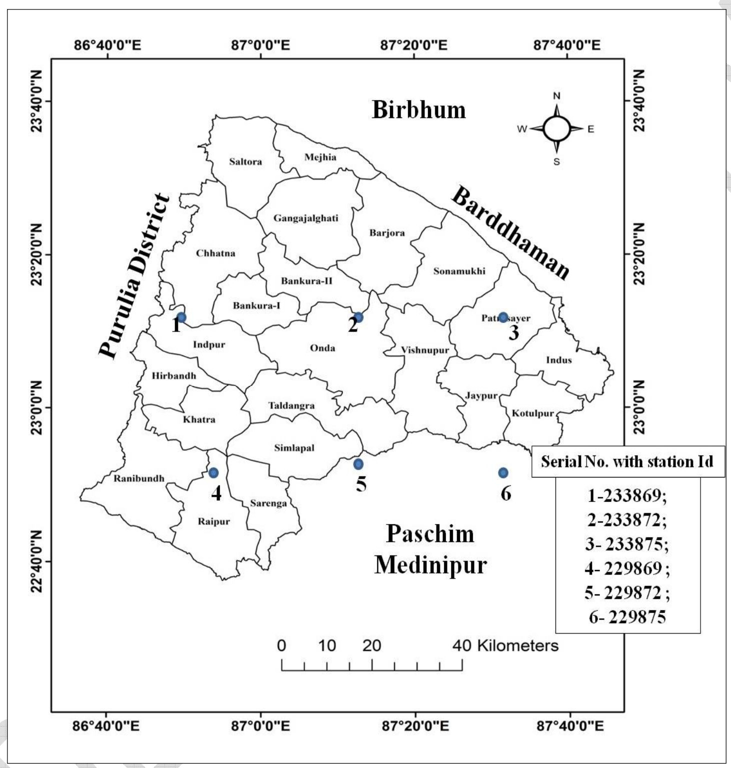

Figure 1 Bankura Location Map and location of Meteorological Stations

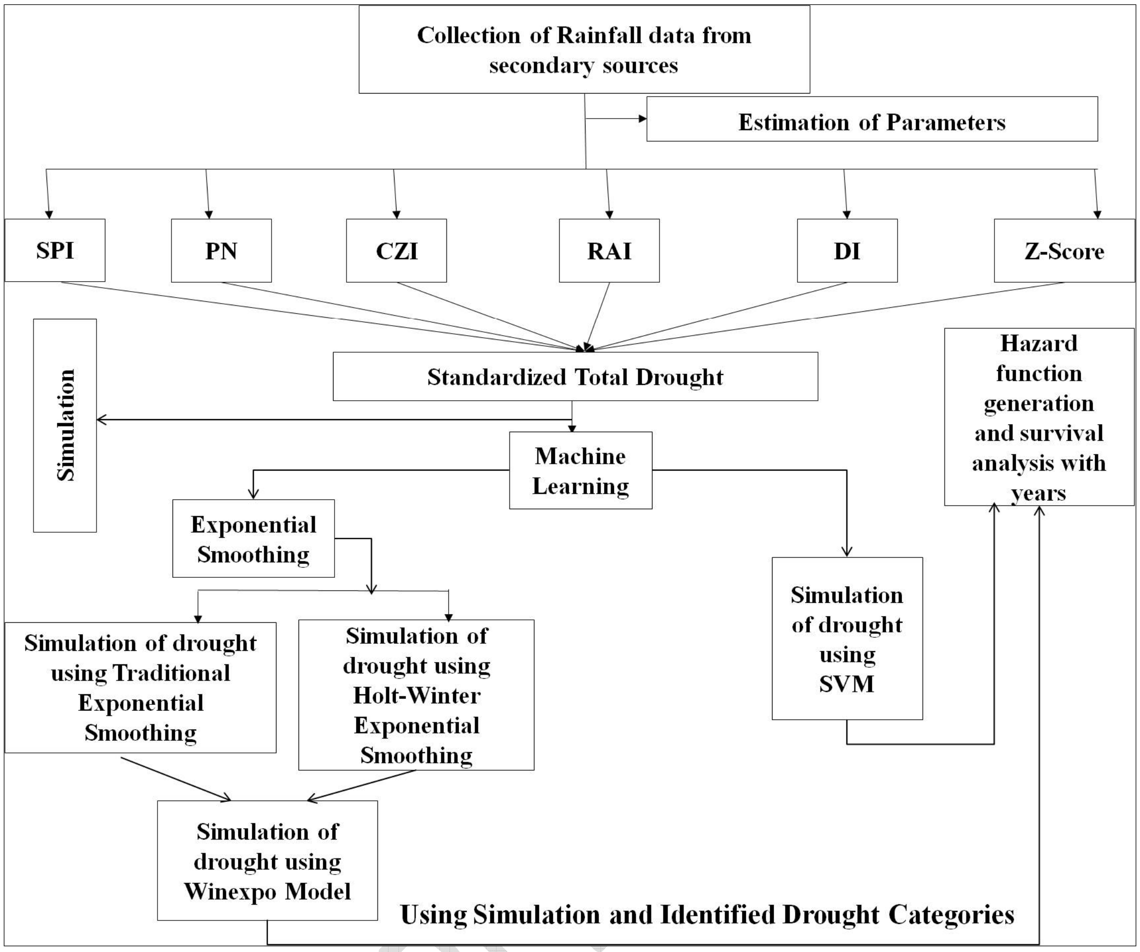

Figure 2 Methodological Overview

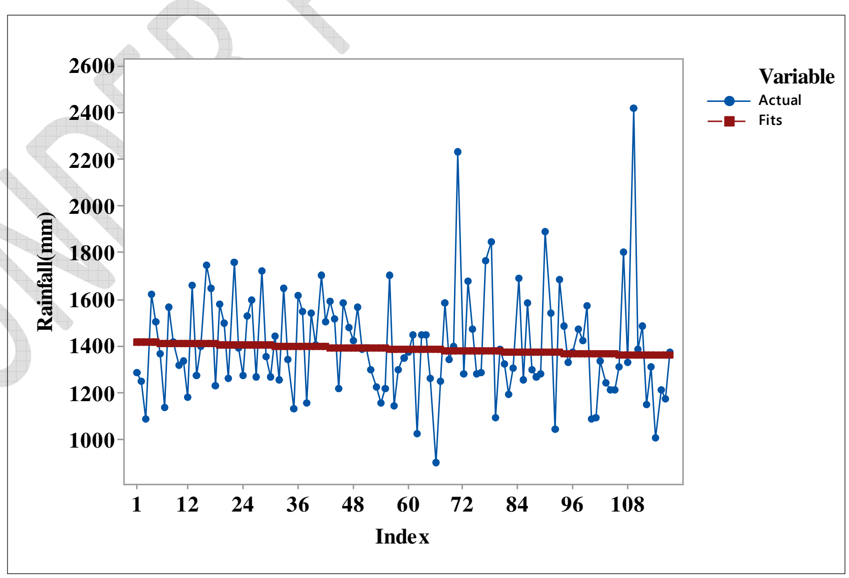

Figure 3 Fluctuation of rainfall according to yearly time steps (1901-2017)

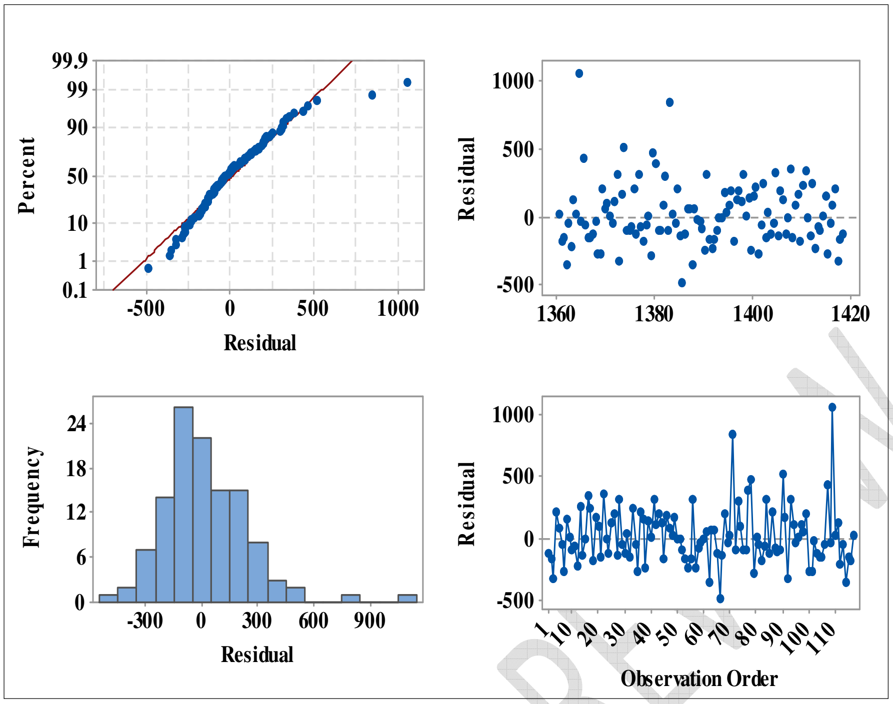

Figure 4a Normal probability Plot of Rainfall **Figure 4b** Fitted value of rainfall vs. Residual value **Figure 4c** Residual value versus Frequency value **Figure 4d** Observation order vs. Residual value

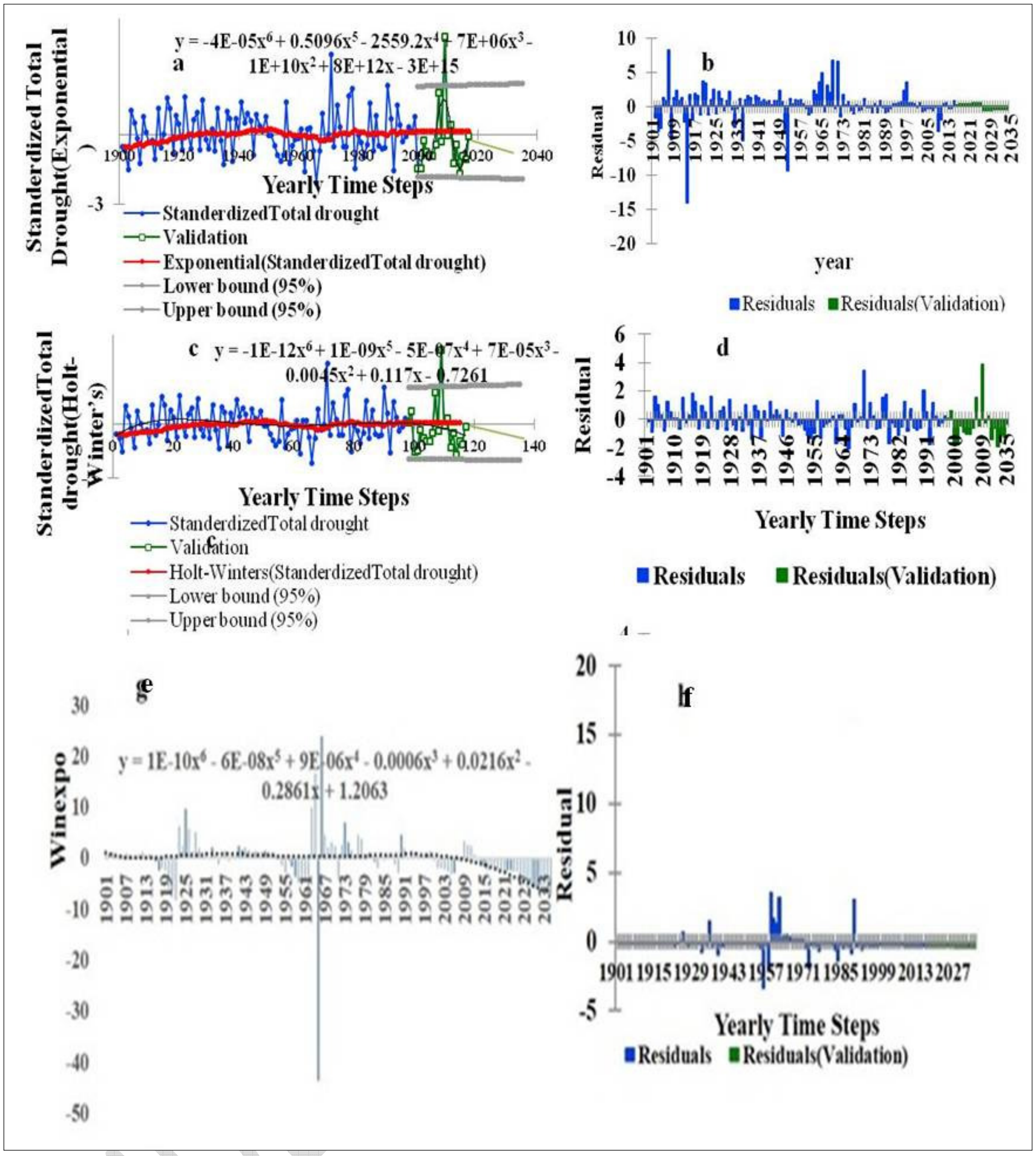

Figure 5a Exponential Smoothing of S_d ,Figure 5b Residual plots of exponential smoothing simulation Figure 5c Holt-Winter's exponential smoothing of S_d Figure 5d Residual plots of Holt-Winter smoothing Simulation Figure 5e Winnexpo exponential smoothing of S_d Figure 5f Residual plots of Winexpo simulation

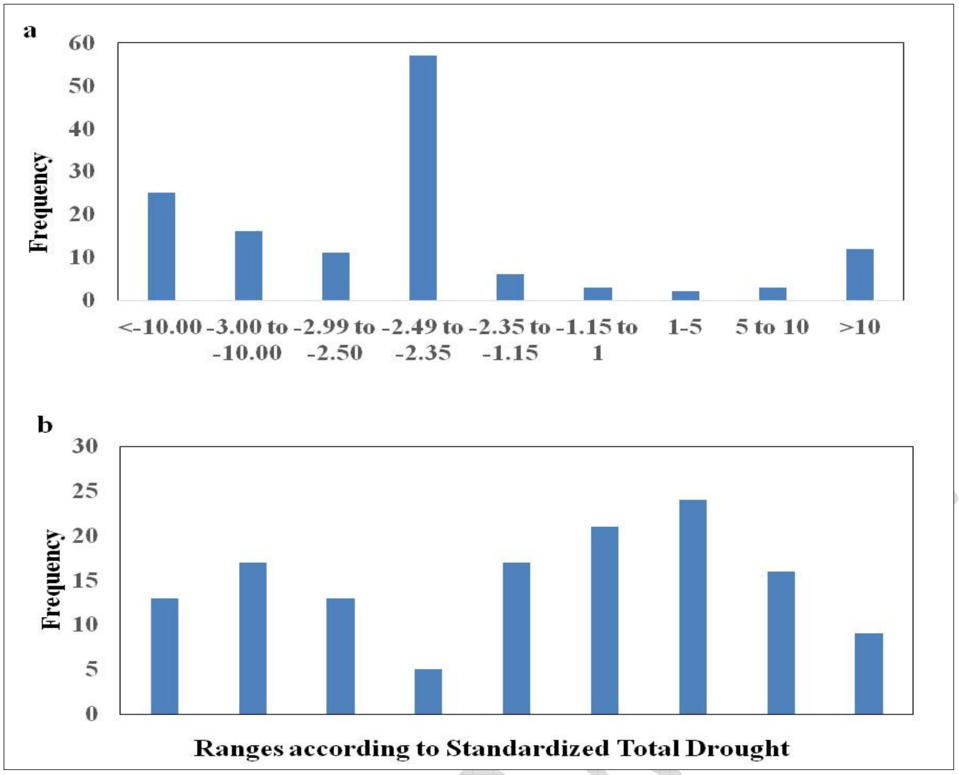

Figure 6a Frequency of drought under each drought categories based on standardized total drought, 1901-2035 (based on simulation method SVM) **Figure 6b** Frequency of Drought under each drought categories based on standardized total drought, 1901-2035 (based on simulation method Winexpo)

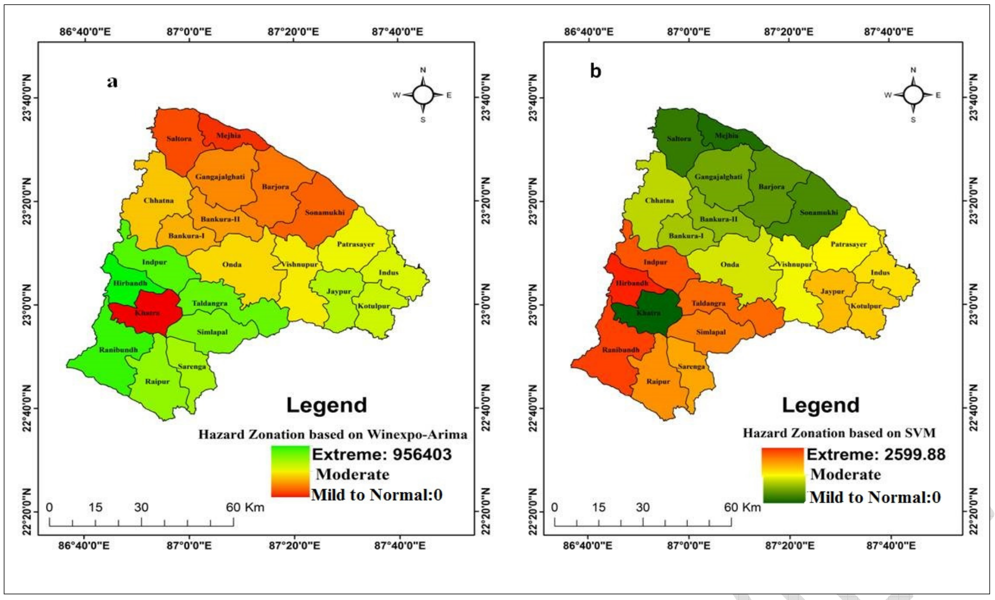

Figure 7a Drought-prone zone identification based on Winexpo (12 month time steps) (1901-2035), **Figure 7b** Drought-prone zone identification based on SVM (1901-2035)

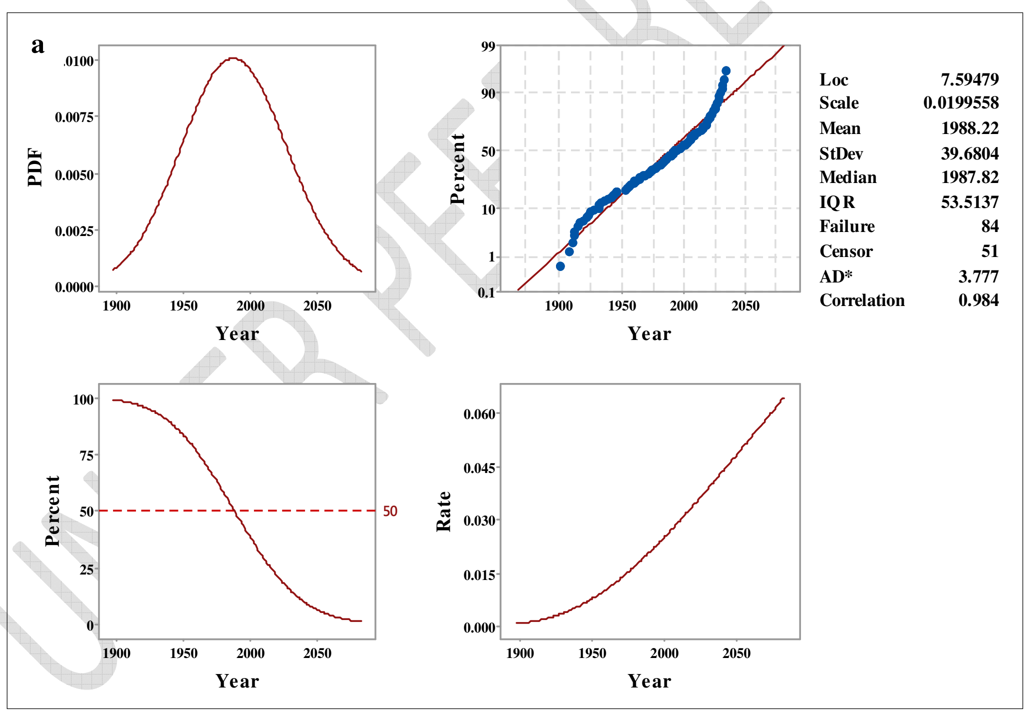

Figure 8a Probability density function of years **Figure 8b** Logistic probability fit for yearly variation of failure and censored categories **Figure 8c** Survival function based on logistic probability fit **Figure 8d** Progression of hazard rate with years

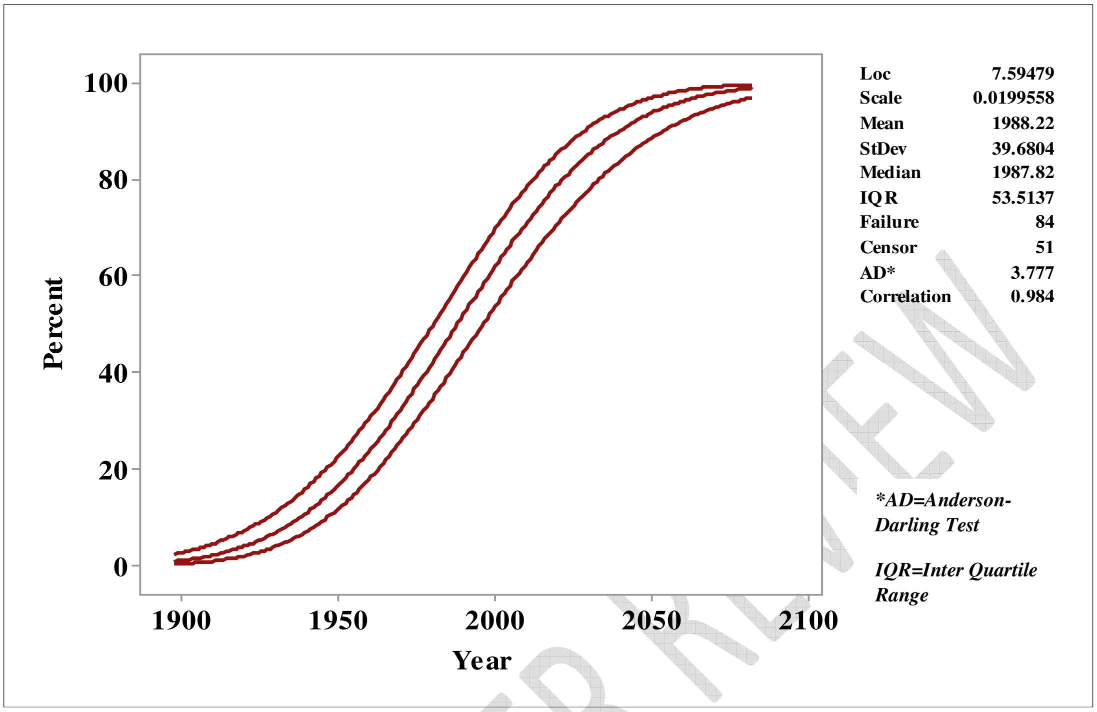

***Figure 9** Year-wise cumulative hazard curve plot over the years