STATISTICS IN TRANSITION-new series, July 2010
Vol. 11, No. 1, pp. 91—103

EXTREME VALUE MODELING OF THE MAXIMUM TEMPERATURE: A CASE STUDY OF HUMID SUBTROPICAL MONSOON REGION IN INDIA

Ripunjai K. Shukla, Manish Trivedi*, Manoj Kumar[1]

ABSTRACT

This experiment is sought to identify and fit the generalized extreme value distribution for extreme maximum temperature data of Ranchi by using the method of maximum likelihood. Ranchi District is located in Humid Sub-Tropical Monsoon Region in Jharkhand state in India. To examine the uncertainty of estimated parameters Q-Q plot and goodness of fit (K-S test) criteria were applied. The study revealed that three parameter Generalized Extreme Value Distribution fitted very well to the data. The estimates of 10, 50, 100 and 200 years return level for yearly extreme maximum temperature are described in that how they vary in future. Further, Exponential Smoothing technique is also applied to capture the trend of extreme maximum temperature in which the residuals fulfilled their assumptions, i.e. randomness and normality.

Key words: Extreme maximum temperature, Maximum likelihood, Generalized Extreme Value etc.

Introduction

Modelling the behaviour of extreme events is important in many fields. These include hydrology, climatology, engineering as well as insurance and finance. In these fields, one is often concerned about the maximum or minimum level of some values rather than the expected or average case.

Extreme events manifest themselves in a variety of ways: large insurance claims, flood levels of rivers, extreme temperature, value-at-risk analysis of a portfolio, wind-speeds, and wave heights during a storm, to name a few. With the ever-growing securitization of risk, there is a commensurately growing need to properly model events that could shock the financial system [Parisi, (2000)].

---

1 Department of Applied Mathematics, Birla Institute of Technology Mesra, Ranchi (JH)-835215, Email: manish_trivedi1976@yahoo.com (Dr. Manish Trivedi - Corresponding Author), msinha_09@rediffmail.com (Dr. Manoj Kumar), ripu121@yahoo.co.in (Mr. Ripunjai K. Shukla).

Among several popular extreme value models, the Generalized Extreme Value (GEV) model play leading role for extreme data. Emil Julius Gumbel, a German mathematician, developed new distributions in the 1950s for extreme events, known as Gumbel distribution. The GEV distribution model combines three types of extreme value model into one expression, with Type I for Gumbel model, Type II for Frechet model and Type III for Weibull model [Huang et al, (2008)]. The GEV Distribution model has been used by [Onofrio et al., (1999)] and [Phein and Fang, (1989)] in Hydrology, Meteorology and Flood analysis. [Michele and Salvadori, (2005)] used GEV distribution as well as Pareto distribution also to estimate extreme of the parameters. [Morrison and Smith, (2002)] developed the model for flood peaks using the generalized extreme value distribution. The Gumbel-distribution, GEV distribution and the Generalized Pareto Distribution is new and quickly growing branch of statistics [Demoulin and Roehrel (2004)].

There is a growing dissatisfaction with the use of standard statistical tools for the prediction of extremes and rare events. Examples abound in several scientists disciplines of gross under estimation, based on historical data, of the probabilities of extreme events that subsequently occur and cause catastrophic damage. This is not restricted to hydrological events: examples appear in ecology (estimating the probability of species extinction), [Ludwig (1996)]; fish management (estimating the exhaustion of fisheries), [Hilborn and Mangel (1997)], [Malakov, (1999)]; insurance (calculating the probabilities of enormous claims, [Smith and Goodman, (2000)]; and geophysical sciences [Smith, (1989)]. Recently [Pinter et al. (2001)] have also emphasized the need for a reassessment of food hazards, highlighting the potential impact of the dynamical structure of rivers which have been modified due to different land usage patterns. Standard methodology for modelling extremes consists of adopting an asymptotic model to describe stochastic variation at extreme levels of a process, inference, and forecasting on the basis of the inferred model. Asymptotically motivated models remain the centrepiece of our modelling strategy, since without such an asymptotic basis, models have no rationale for extrapolation beyond the level of observed data. However, in this article we argue that there are two principal reasons why naive adoption of this paradigm leads to systematic underestimation of the probability of disastrous events. First, model and prediction uncertainty are often overlooked or ignored. Since such uncertainties can be substantial, designs made without allowance for these effects can be disastrously anti-conservative. Moreover, it is commonplace to reduce the dimensionality of extreme value models and to proceed on the basis of a simplified model. This procedure may lead to similar estimated models, but is likely to lead to overly confident measures of precision as the principal source of model uncertainty is artificially discarded. The second aspect is that of model homogeneity. We argue that a false assumption of a stationary model may also lead to considerable underestimation of the probability of a disastrously extreme event. While recent techniques, such as the local likelihood method of [Ramesh and Davison (2002)], identify local temporal variation by parameter estimation at each individual time point via appropriate data weightings, we take a different approach. With no empirical evidence of temporal trends, we model non-stationary effects through the assumption of within-season homogeneity for model parameters within a number of seasons, whilst allowing for variations across seasons whose starting point and duration are treated as unknown. Our main departure from standard methodology is a preference for Bayesian inference. Though we also include classical likelihood analyses for comparison, we argue that the Bayesian approach provides a more coherent framework for keeping track of, and incorporating, all of the uncertainties involved in the prediction process. Computations for such models are intractable using conventional techniques, but are now almost routine using stochastic algorithms such as Markov chain Monte Carlo (MCMC).

Extremes are unusual or rare events and in classical data analysis often labelled as outliers and even ignored. For the events which do not happen very often for that one should apply extreme value theory. Earthquakes, hurricanes, and stock market crashes are surprising phenomena which do not follow any rule, but can be modelled by the distributions of extreme value models. For example, very high temperature can be fatal and is therefore more important than average temperature when assessing the consequences of climate changes. Policy makers are concerned with changes in the probability of extremes of climate parameters, i.e. extremely hot summer in the next coming decades. Therefore, the present investigation was carried out to assess and predict the changes in extreme maximum temperature of Ranchi.

Methodology

Distributional Study

The annual maximum of temperature was modelled using GEV distribution. Denoting daily observations by X_1, $X_{2,\ldots\ldots}$, the classical model for extremes is obtained by studying the behaviour of M_n = max [X_1, ……, X_n] for large values of n, where M_n corresponds naturally to the annual maximum. Asymptotic considerations suggest that the distribution of M_n should be approximately that of a member of the generalized extreme value distribution family [NERC,(1975)]; [Leadbetter et al., (1983)], for example, having distribution function

$$f(x;\mu,\sigma,k)=\frac{1}{\sigma}\left[1+k\left(\frac{x-\mu}{\sigma}\right)\right]^{-1-1/k}\exp\left\{-\left[1+k\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/k}\right\},\ k\neq 0$$

$$\text{and}\qquad f(x;\mu,\sigma,k)=\frac{1}{\sigma}\left[-\left(\frac{x-\mu}{\sigma}\right)\right]-\exp\left\{-\left[\left(\frac{x-\mu}{\sigma}\right)\right]\right\},\ k=0 \tag{1}$$

Where $\sigma>0$, $-\infty<\mu$, $k<\infty$ and the range of x is such that $\left[1+k(x-\mu)/\sigma\right]>0$.

The location, scale and shape of the distribution are controlled by the parameters μ, σ and k respectively.

For a random sample of annual maxima $x_1, \ldots, x_n$, the log likelihood is written as

$$l(\mu,\sigma,k) = -n.\log\sigma\left[-\left(\frac{1}{k}+1\right)\right]\sum_{i=1}^{n}\log\left[1+k\left(\frac{x_i-\mu}{\sigma}\right)\right]-\sum_{i=1}^{n}\left[1+k\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-1/k} \quad k\neq 0$$

(2)

and,

$$l(\mu,\sigma,0) = -n.\log\sigma - \sum_{i=1}^{n}\log\left[\left(\frac{x_i-\mu}{\sigma}\right)\right]-\sum_{i=1}^{n}\left[\left(\frac{x_i-\mu}{\sigma}\right)\right], \quad k=0$$

This expression allows us to calculate the maximum likelihood estimates of the parameters. In the most applications the quartiles of the fitted distribution are to interest as they allows us to make predictions about the level exceeded once every 1/(1-T) years on the average. This is called the return level and is given by [Nadarajah and Choi, (2007)].

$$X_T = \mu - \frac{\sigma}{k}\left[1-\left\{-\log(1-\frac{1}{T})\right\}^{-k}\right], \quad k \neq 0$$

$$X_T = \mu - \sigma.\log\left\{-\log\left(1-\frac{1}{T}\right)\right\}, \quad k = 0$$

(3)

If L_1 is the maximum likelihood of three parameter distribution and L_2 is the maximum likelihood of the two parameter distribution, then the preferable model from the both fitted models can be judged by test statistic $\lambda = -2.\log(L_2/L_1)$, would be assumed to follow the χ^2 distribution with 1 degree of freedom (Wald, 1943). Since the number of parameters differs by 1 but it would be asymptotically true as the number of observations tend to infinity.

Model Incorporating Trend

The time series forecasting assumes that a time series is a combination of a pattern and some random error. The goal is to separate the pattern from the error by understanding the pattern trend, its long term increase and decrease and its seasonality. Several methods of time series forecasting are available such as the moving averages methods, linear regression with time, exponential smoothing, etc. This study concentrates on Exponential Smoothing technique as applied to extreme temperature time series data.

Single Exponential Smoothing

This is also known as simple exponential smoothing. Simple smoothing is used for short-range forecasting, usually just one step ahead into the future. The model assumes that the data fluctuates around a reasonably stable mean (no trend or consistent pattern of growth). The following specific formula for simple exponential smoothing is given in [Makridakis et al, (2003)]:

$$F_t = \alpha * X_t + (1- \alpha) * F_{t-1} \quad (4)$$

Where X_t is the observation at time t, F_t is the forecasted value at time t and α is the damping factor having the range $0 < \alpha < 1$.

When applied recursively to each successive observation in the series, each new smoothed value (forecast) is computed as the weighted average of the current observation and the previous smoothed observation; the previous smoothed observation was computed in turn from the previous observed value and the smoothed value before the previous observation, etc.

Initial Value

The initial value of F_t, plays an important role in computing all the subsequent values. Setting it to x_1 is one method of initialization. Another possibility would be to average the first four or five observations. The smaller the value of α, the more important is the selection of the initial value of F_t.

Testing the homogeneity of Mean Square of Errors

Bartlett test carried out to check the homogeneity of error Mean Square of Error that whether they are homogeneous or not at various values of damping factor α. The test statistic of Bartlett test is given by the following formula [Gupta and Kapoor, (1970)]

$$\chi^2 = \sum_{i=1}^{d} \left(v_i \log \frac{S^2}{S_i} \right) \Big/ \left[1 + \frac{1}{3(d-1)} \left\{ \sum_{i=1}^{d} \left(\frac{1}{v_i} \right) - \frac{1}{v} \right\} \right] \quad (5)$$

Where $S^2 = \dfrac{\sum_{i=1}^{d} v_i S_i^2}{\sum_{i=1}^{d} v_i} = \dfrac{\sum_{i=1}^{d} v_i S_i^2}{v}$, $\sum_{i=1}^{d} v_i = v$

In the above expression S_i^2 is the i^{th} Mean Square of Error and v_i is the corresponding degree of freedom. The test statistic follows χ^2 distribution with (d-1) degree of freedom.

Data source

The data of daily basis maximum temperature for the years 1960 to 2006 were collected form Birsa Agricultural University, Ranchi (Jharkhand) and daily wise extreme maximum temperature has been extracted from the whole data set then the analysis of that extreme maximum temperature has been carried out to further study.

Fitting of generalized extreme value distribution

The data are presented graphically in Fig.1, which does not give much insight to any long term change. The summary statistics given in Table-1 revealed that data showed slight longer tail on left side with more flatness, i.e. negatively skewed and platykurtic with slow variability. The standard deviation (S.D.) and standard error (S.E.) are 1.182 and 0.174 respectively.

Figure 1. Yearly extreme maximum temperature

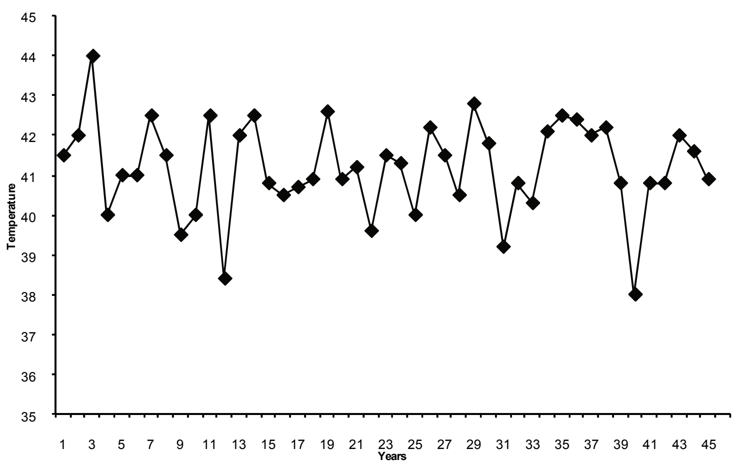


Table 1. Summary Statistics of extreme maximum temperature data

Statistics	Value
Sample Size	46
Range	6
Mean	41.213
Standard Deviation	1.182
Coefficient of Variation	2.870
Standard Error	0.174
Skew ness	-0.505
Kurtosis	0.661

Maximum likelihood estimates of the various parameters of GEV distribution with and without shape parameter are given in Table-2. The probability density function (PDF) and cumulative density function (CDF) of GEV distribution (only for k≠0) indicated in Fig.-2 and Fig.-3 respectively. From the both types of fitted model the GEV, when k≠0, is better due to likelihood ratio test failed, i.e. λ (Wald's statistics)= 0.6721 < χ^2 (3.84) at 1 degree of freedom. The bracketed values in table-2 are the S.E. of the respective parameters.

Table 2. The maximum likelihood estimates of GEV distribution

Estimated parameters		
Location ($\hat{\mu}$)	Scale($\hat{\sigma}$)	Shape ($\hat{k}$)
40.907 (0.2108)	1.2578 (0.1987)	-0.4707 (0.1404)
40.681 0.1497	0.9215 0.0570	–

Figure 2. PDF of the GEV distribution for Extreme Maximum Temperature

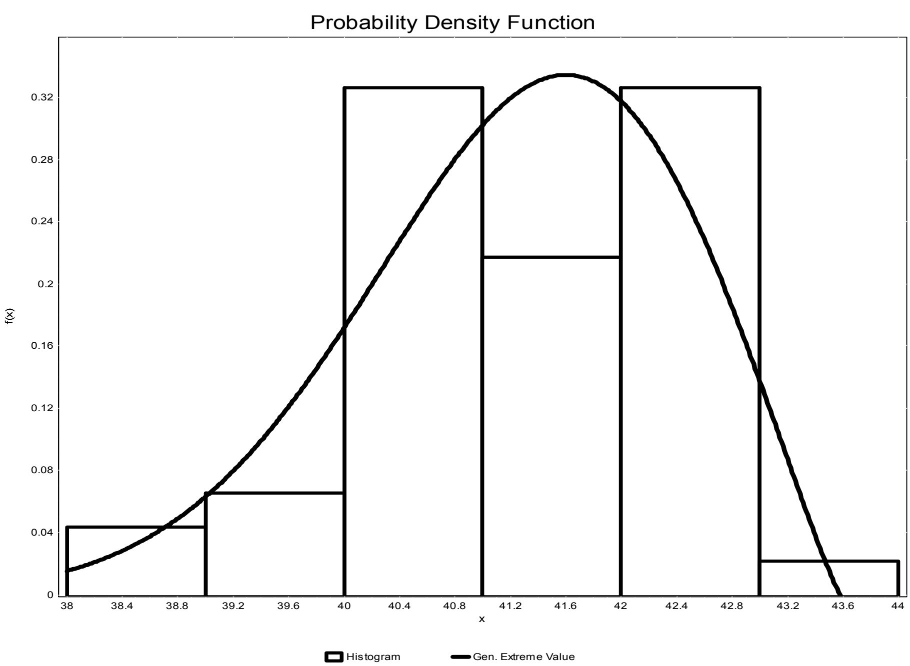


Figure 3. CDF of the GEV distribution for Extreme Maximum Temperature

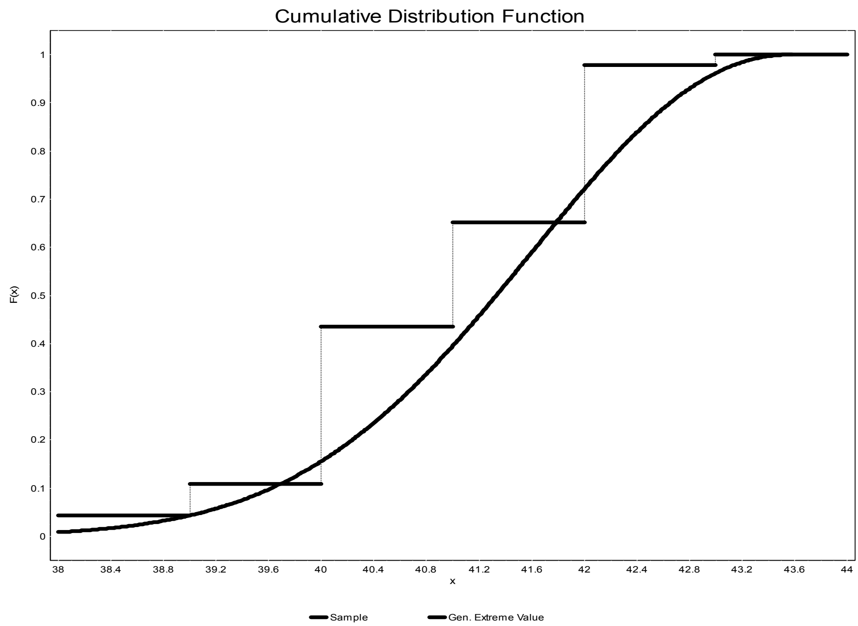
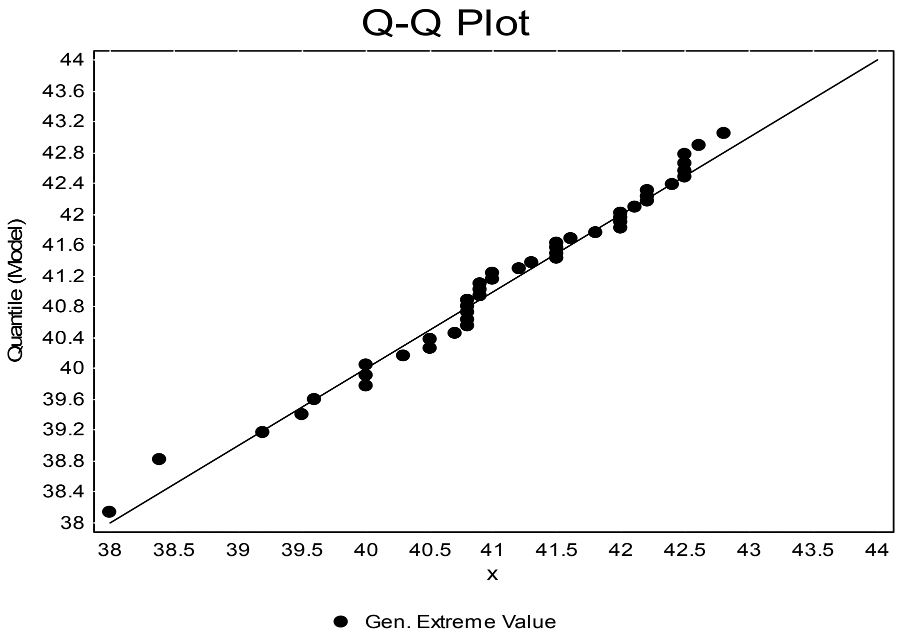


Table 3. Goodness of fit of the GEV Distribution model

Kolmogorov-Smironov Test				
Sample Size Statistic P-Value	46 0.08285 0.88448			
p	0.2	0.1	0.05	0.01
Critical Value	0.15457	0.17665	0.19625	0.23544
Null Hypothesis	Accept	Accept	Accept	Accept

The diagnosis of the model by the Q-Q plot indicated that quantiles are much closer to the diagonal reference line (Fig.4). A Q-Q plot is where the observed quantile is plotted against the quantile predicted by the fitted model. For example, to check the goodness of fit of model 1, we would plot the sorted values (in the ascending order) of the observed annual maximum daily rainfall versus expected quantiles Y_i determined by $F(Y_i) = (i - 0.375)/(n + 0.25)$ (Nadarajah and Choi, 2007), where F is the CDF of GEV distribution and n is the number of observations in the data series. Confirmation of good fitting is more emphasized by non-significant Kolmogorov-Smironov (K-S) test (Table-3), in which the null hypothesis of goodness of fit is accepted.

Figure 4. Q-Q plot of GEV Distribution model



Once the best model for the data is determined the next interest to derive the return levels of the Extreme Maximum Temperature in the summer seasons. The p year return level, say x_p, is the level exceeded on average only once in T years already indicated in equation (3). The return level for 10, 50, 100 and 200 years are consistently increasing towards long run in future indicated in the Table -4.

Table 4. Return level of maximum extreme temperature

Return Level				
Year	T= 10	T= 50	T=100	T= 200
Temperature (°C)	42.90	43.30	43.37	43.43

Modeling by exponential smoothing of extreme maximum temperature

The asymptotic arguments which lead to the generalized Pareto model as a distribution of threshold excesses by an independent sequence can be generalized to consider sequences of dependent, though stationary, series. Under a weak condition that limits the effect of any long-range dependence, it can be shown that the generalized Pareto family has the same status as a limiting family for threshold excesses in this more general setting (see [Leadbetter et al, (1987)] for example). This provides some justification for use of the generalized Pareto distribution as a model for extreme daily rainfalls, even if there is some temporal dependence in the series. The exponential smoothing was carried out at various value of damping factor (α). The Mean Square of Errors (MSE) and errors assumption (i.e. randomness and normality) and Bartlett test to check the homogeneity of Mean Square of Errors are illustrated in Table-5.

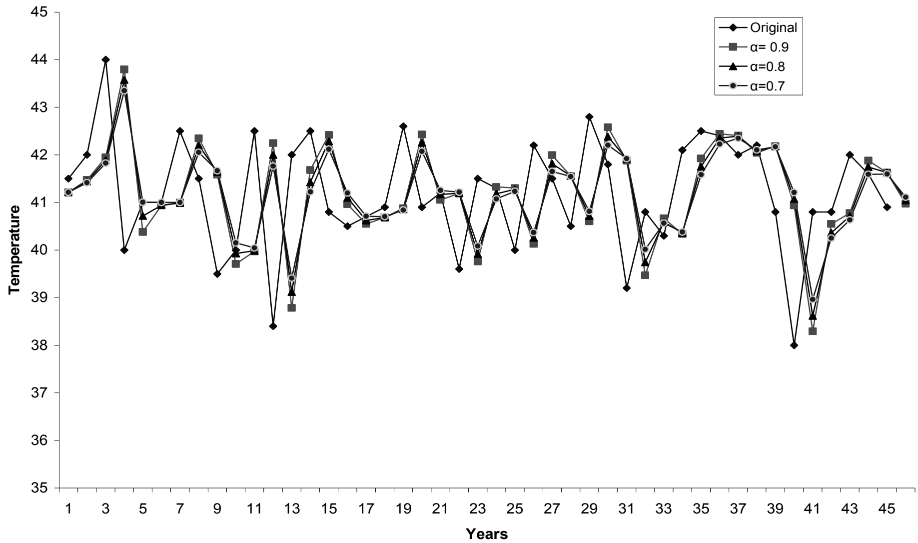
Table 5. Validation of the model at various values of α

α	MSE	Run Test (p = 5%)	K-S test (p = 5%)	Bartlett Test (χ^2) (p= 5%)
0.9	2.639	PASS	PASS	PASS
0.8	2.442	PASS	PASS	
0.7	2.272	PASS	PASS	

From Table-5, it is seen that hypothesis is successfully passed for all the values of α varying from 0.7 to 0.9 with an increment of 0.1 for both Run Test and K-S test. Bartlett test is also passed for Mean Square of Errors, i.e. there is no significant difference among the Mean Square of Errors and the predicted values by this technique are much closer to observed values when α = 0.9. Moreover, it is observed that the Mean Square of Errors is slightly higher for larger values of α but these are statistically same as indicated by Bartlett test (see Table-5).

The observed and predicted extreme of maximum temperature for different values of α is given in Fig.5. Form Fig. 5, it is clearly visible that when α = 0.9, the predicted values are much closer with observed values. Generally speaking, the more the value of α is the more the fitness of the predicated value with the observed value.

Figure 5. Observed and predicted values by exponential smoothing at various values of α



Conclusion

From the present study it is concluded that (i) the model is best fitted when the damping factor is closer to 1. (ii) The extreme maximum temperature gradually and slightly increases from time to time for the next 200 years. An overall look of the study shows that by using these techniques one can be risk-free of extreme temperature arising in the environmental disaster in the sectors like agriculture, forest fire, drought, etc. For further study, researcher can make a forecast at one step ahead whereas in distributional study one can predict the return level of extreme temperature for long run of time in the future.

Acknowledgement

The authors are thankful to the Birsa Agricultural University, Ranchi (Jharkhand) for providing necessary data by which the present investigation has been possible.

REFERENCES

D' ONOFRIO, E. E., FIORE, M. M. and ROMERO, S. I. (1999); Return periods of extreme water levels estimated for some vulnerable areas of Buenos Aires. *Continental Shelf Research* 19, 1681–1693.

DAVISON, A. C. and SMITH, R. L. (1990) Model for exceedances over high thresholds. Journal the Royal Statistical Society, Series B, 52, No. 3, 393–442.

DE MICHELE, C. and SALVADORI, G. (2005); Some hydrological applications of small sample estimators of Pareto and Extreme Value distributions, *Journal of Hydrology* 301, 37–53.

DEMOULIN, V. C. and ROEHREL, A. (2004); Extreme Value Theory can save your neck. *White Paper Jan-08, 2004* (Supported by Swiss National Science Foundation).

GUPTA, S. C., KAPOOR, V. K. (1970); Fundamental of Mathematical Statistics. *Chapter- 13*, pp-13.68.

HILBORN, R. and MANGEL, M. (1997) The ecological detective: Confronting models with data. Monographs on Population Biology. Princeton University Press.

HUANG, W., SUDONG XU and NNAJI, S. (2008); Evaluation of GEV model for frequency analysis of annual maximum water levels in the coast of United States, *Ocean Engineering*, doi: 10.1016/j.oceaneng.2008.04.010.

LEADBETTER, M.R., LINDGREN, G. and ROOTZEN, H. (1983); Extremes and related properties of random sequences and processes. Springer-Verlag, New York.

LUDWIG, D. (1996) Uncertainty and the assessment of extinction probabilities. Ecological Applications, 6 (4), pp. 1067–1076.

MALAKOV, D. (1999) Bayes offers a New Way to make sense of numbers. Science, Vol. 286, pp. 1460–1464.

MAKRIDAKIS, S., WHEELWRIGHT, S. C. and HYNDMAN, R. J. (2003); FORECASTING: Methods and Applications (Third Edition), Chapter-4, pp 136–180.

MORRISON, J. E. and SMITH, J. A. (2002); Stochastic modeling of flood peaks using the generalized extreme value distribution. *Water Resource Research* 38, 411–412.

NADARAJAH, S. and CHOI., D. (2007); Maximum daily rainfall in South Korea. *J. Earth Syst. Sci.* 116, No. 4, pp. 311–320.

PARISI, F. (2000); Special Report: Extreme Value Theory and Standard and Poor's Rating. *ABC research,* New York :(1) 212–438–2570.

PHEIN, H. N., FANG, T. S., (1989); Maximum likelihood estimation of the parameters and quartiles of the general extreme-value distribution for censored samples. *Journal of Hydrology* 105, 139–155.

PINTER N., THOMAS R. and WLOSINSKI H. (2001) Assessing ood hazards on dynamic rivers. EOS, Transactions, American Geophysical Union, Vol. 82, Number 31, 333 and 339–340.

RAMESH, N. I. AND DAVISON, A. C. (2002) Local models for exploratory analysis of hydrological extremes, Journal of Hydrology, 256, 106–119.

SMITH, R. L. (1989) Extreme value analysis of environmental times series: an example based on ozone data (with discussion), Statistical Science, 4, 367–393.

SMITH, R. L. and D. J. GOODMAN (2000) Bayesian risk analysis. Technical Report, Dept. Statistics, UNC-Chapel Hill.

WALD, A. (1943). Test of statistical hypothesis concerning several parameters, when the number of observation is large. Transactions of American Mathematical Society, 54:426–483.