The problem of determining the threshold for statistical analysis by the POT method

Application to wave data on the Moroccan Atlantic coast

Hosny BAKALI[1*], *Ismail AOUICHE*[1], and Najat SERHIR[1]

[1] Er-SHEMSS, Laboratory of Civil Engineering, Hydraulics, Environment and Climate, Hassania School of Public Works, Morocco

Abstract. In a study of extreme waves by the Peak Over Threshold (POT) method, the determination of the threshold of data censoring is an essential step. A wrong choice of the threshold can lead to erroneous results of the wave height design and consequently a bad design of maritime structures such as breakwaters for deep sea ports. In this study, we analyzed the influence of the threshold variation on the results of the hundred-year return period waves, generally considered for the design of maritime structures. The sensitivity study allowed us to confirm that the exponential model is the best probability distribution to describe wave data in two points on the Moroccan Atlantic coast for the wave data period from 1958 to 2019. This study also confirmed that a wrong choice of the statistical distribution and a wrong choice of the threshold lead to significant errors in the estimation of design wave height.

* Corresponding author: albakalihosny@gmail.com

1 Introduction

The determination of the design wave height is a crucial step in the design of maritime structures. The significant wave height generally considered is the 100 years return period wave, which is determined by statistical analysis of the wave data available over periods of a few decades. The first recordings of waves are made in the sixties [1]; therefore the available periods are insufficient for an accurate calculation of waves with low occurrence probability. The extreme waves heights are determined by extrapolation from statistical distributions fitted to the wave data.

To select the data for statistical analysis, two methods are used:

• The first approach is called BLOCK MAXIMA method [2]. The data is the maximum values over regular time intervals, generally taken equal to one year. In this particular case, the method is referred to by the "ANNUAL MAXIMA METHOD (MA)".

• The second approach is the Peak Over Threshold method (POT) [3].

The two methods consider independent and identically distributed (iid) of random variables. POT method required iid values above a chosen threshold.

The POT method has the advantage of considering all the significant waves [2]; nevertheless, the determination of the censoring threshold is the key step for obtaining the best relevant data of extreme waves and consequently the best estimation of the wave height design.

The threshold is set to obtain an average number of events per year over the entire data period equal to or slightly less than N_a, which is the number per year of extreme storms in the study area [4]. Authors of [5] propose the determination of a first threshold u_1 to obtain a number of events between 5 and 10 similar to the number of significant storms in extratropical zones. The second threshold u_2, is the abscissa where the mean residual life plot function becomes linear [8]. The u_2 threshold should not be very high to maintain a significant number of data for statistical analysis. The N_a number should remain approximately from 2 to 5 [5].

The existing methods do not allow a single choice of threshold regardless of the size of the data and the study area. A local sensitivity study of extreme waves for the case of the Moroccan Atlantic coast is therefore required to propose the safest approach.

In this paper, we will present the impact of the threshold variation on the predicted 100 years return period significant wave height. We will carry out a sensitivity study at two points on the northern Atlantic Moroccan coast. The wave data are taken from the SIMAR-44 database for the period from 1958 to 2019 on two points: SIMAR network N° 1050036 and N° 1042030 located on the coast of the cities of Mohammedia and Safi.

2 Methodology

2.1 Method for selection independent waves

Authors of [5] and [6] recommend the verification of the following criteria for the choice of the extreme values of storms; these criteria are fixed for each site according to the local meteorological and maritime conditions:

• The minimum duration of the storm above the chosen threshold U is greater or equal to N_H with a possibility of a storm subside below the threshold for a duration of n_H (n_H is around 3 hours)

• A time lag of 1 to 3 days should be between two extreme values.

According to Caires and Sterl [7], the N_H time interval between two independent storms is 48 hours. To ensure the independence of the storms, we will adopt the criterion of minimum interval between two successive extreme values of 48 hours.

2.2 Determination of the threshold

The mean excess function (MEF) describes the prediction of the exceeding threshold u when an excess occurs [8], is defined by (1):

$$\hat{e}_n(u) = \frac{\sum_{i=1}^{n}(X_i - u)_{\{X_i > u\}}}{\sum_{i=1}^{n} 1_{\{X_i > u\}}} \quad (1)$$

A linear trend of the Mean Excess Function in the threshold u_T indicates a stabilization of the parameters σ_u and ξ of the Generalized Pareto Distribution fitted to the data, u_T is therefore a possible value of the threshold [8]

2.3 Software used

We used the HYFRAN-PLUS software, designed for the analysis of hydraulic data. This software allows the adjustments by several statistical models as well as the comparison between these adjustments by graphical method and the goodness of fit tests [9].

2.4 Calculation of empirical non- exceeding probabilities

The general formula for determining the empirical probabilities of non-exceedance is (2):

$$P_k = \frac{k-\alpha}{n+1-2\times\alpha} \quad (2)$$

We will consider the compromise distribution proposed by Cunnane [10], with $\alpha = 0.4$

2.5 Method for selecting the most suitable distribution model

The model selection is made by a multi-distribution analysis. We studied the theoretical distributions: Gumbel, GEV, Weibull, Gamma, Inverse Gamma, Lognormal, and Exponential.

A first choice of the most suitable distributions is made by comparing the graphical fittings to the data [11]. The final selection of the best theoretical model

is made by referring to the goodness-of-fit tests: $\chi 2$ [12], AIC [13], and BIC [14].

3 Results

3.1 Synthesis of the results on the coast of Mohammedia city - Point SIMAR network N° 1050036

3.1.1 Determination of censorship thresholds for wave data

Figure 1 presents the mean residual life plot function of the extreme wave's data. In addition to the graphically determined thresholds, and for comparison purposes, we added an additional value u'= 5.00.

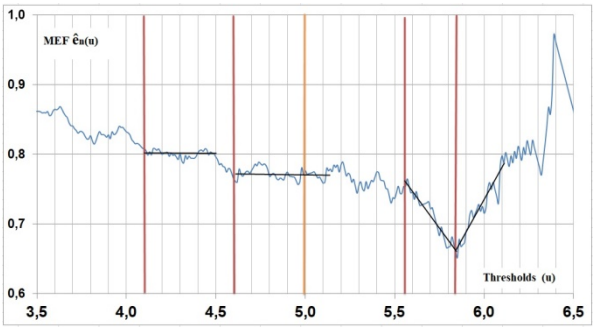


Fig. 1. Graphical determination of censoring thresholds

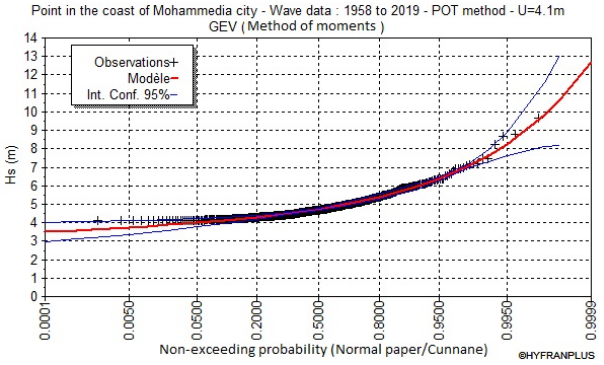
The total and the average per year wave's number above thresholds are presented in Table 1:

Table 1. Overall and average per year number of extreme events for each threshold

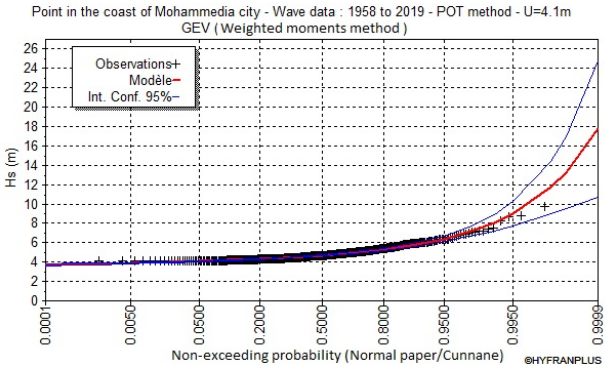
Threshold	Value	Number of events considered	The average number of waves per year N_a
U_1	4.10 m	447	7.21
U_2	4.60 m	226	3.64
U'	5.00 m	138	2.22
U_3	5.55 m	72	1.16
U_4	5.85 m	55	0.89

3.1.2 Graphical adjustments of extreme waves data on the coast of Mohammedia city

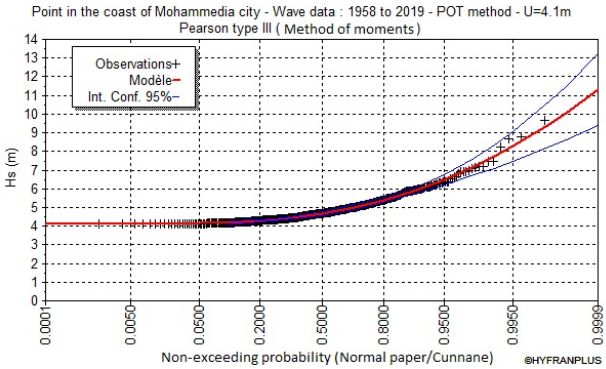
Figure 2 presents some examples of graphical adjustments for the threshold U_1 = 4.10 m:



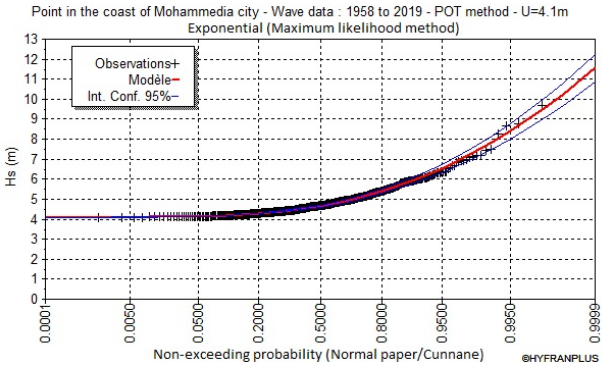
(a)



(b)



(c)



(d)

Fig. 2. (a) Graphical adjustment of GEV distribution & Method of moment. (b) Graphical adjustment of GEV distribution & weighted moments method (c) Graphical adjustment of Pearson III distribution & Method of moment (d) Graphical adjustment of the Exponential distribution $ Maximum likelihood method.

We present in Table 2 the results of the goodness-of-fit tests of the most graphically appropriate distributions.

Table 2. Results of fit tests for the models with the best graphical adjustments

Statistical distribution	GEV		Pearson type 3	**EXP**
Method for estimating model parameters	Method of moments	Method of weighted moments	Method of moments	**Maximum likelihood**
Value of $\chi 2$	96.01	57.82	15	**15**
AIC	831.92	795.21	-	**709.49**
BIC	844.23	807.52	-	**717.69**

Based on the graphical comparison and the goodness-of-fit tests; we conclude that the adequate model is the exponential distribution with maximum likelihood method for the estimation of the distribution parameters.

Figure 3 presents the results of the hundred-year return period wave's height as a function of several threshold values. The wave's height values retained are illustrated in full black patterns.

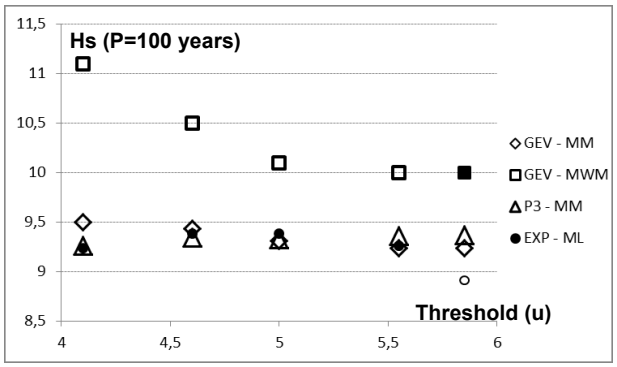


Fig. 3. Variation of the hundred-year return period waves for different forecasting models depending on thresholds in the point in Mohammedia's coast

3.2 Synthesis of the results on the coast of Safi city - Point SIMAR network N°1042030

3.2.1 Determination of censorship thresholds for wave data.

We present in figure 4 the mean excess function of the extreme wave's data.

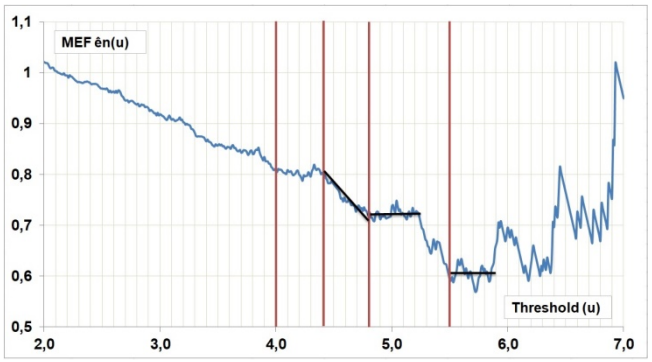


Figure 4: Graphical determination of censoring thresholds

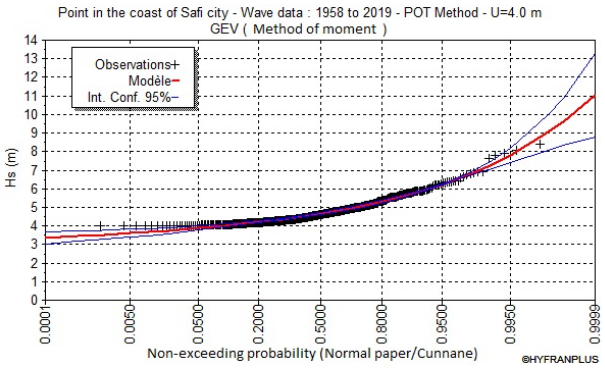
The total and the average per year storms number above thresholds are presented in table 3:

Table 3. Overall and average per year number of extreme events for each threshold

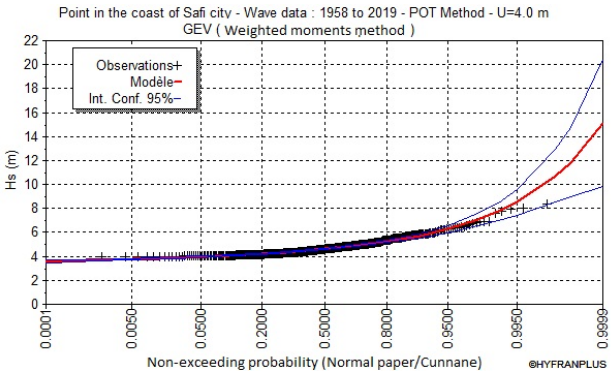
Threshold	Threshold Value	The number of events considered	The average number of storms per year: Na
U_1	4.00	405	6.53
U_2	4.40	248	4.00
U_3	4.80	164	2.64
U_4	5.50	73	1.17

3.2.2 Graphical adjustments of extreme wave data on the coast of Safi city.

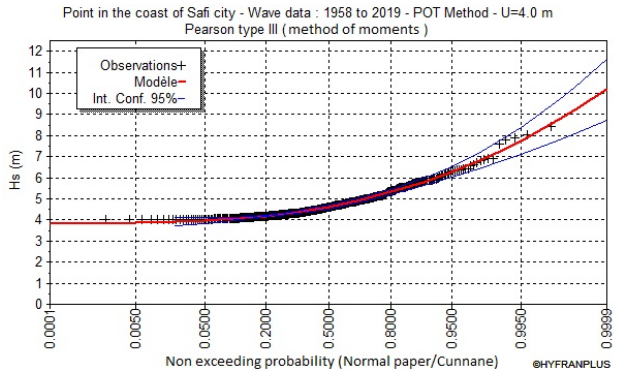
Figure 5 presents the graphical adjustments for the threshold $U_1 = 4.00$ m



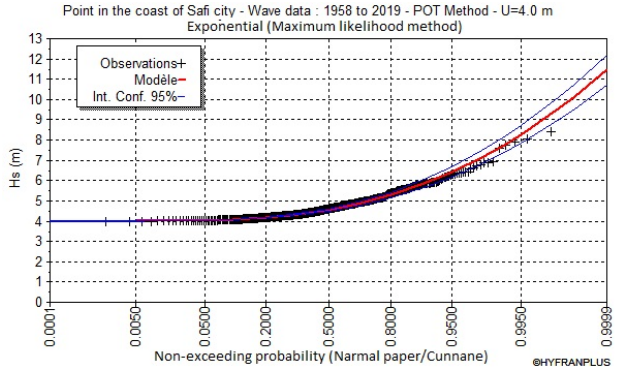
(a)



(b)



(c)



(d)

Figure 5: (a) Graphical adjustment of GEV distribution & Method of moment. (b) Graphical adjustment of GEV distribution & weighted moments method (c) Graphical adjustment of Pearson III distribution & Method of moment (d) Graphical adjustment of the Exponential distribution & Maximum likelihood method.

The results of the Goodness-of-fit tests of the best graphical fitting models are given in table 4:

Table 4. Results of fit tests for the models with the best graphical adjustments

Statistical distribution	GEV		**Pearson type 3**	EXP
Method for estimating model parameters	Method of moments	Method of weighted moments	**Method of moments**	Maximum likelihood
Value of $\chi 2$	107.84	74.06	**42.99**	20.28
AIC	762.96	733.38	**714.43**	642.00
BIC	774.97	745.39	**726.44**	650.01

Based on the graphical comparison; we conclude that the adequate model is the Pearson III distribution with the method of moment for the estimation of the model parameters.

The results of the hundred years return period wave's height for the statistical distributions with the best graphical fitting are presented in Figure 6:

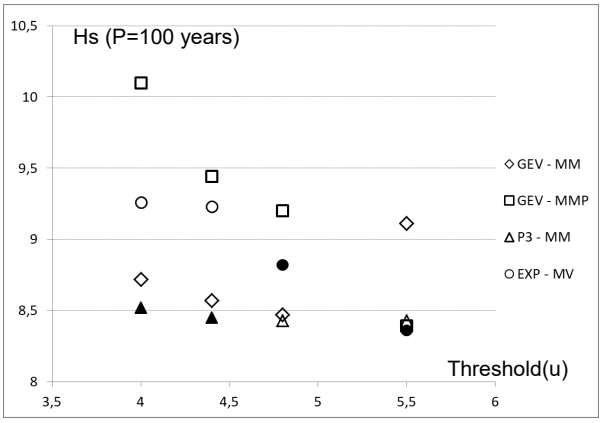


Figure 6: Variation of the hundred-year return period waves for different forecasting models depending on thresholds in the point in Safi's coast

3.3 Results analysis

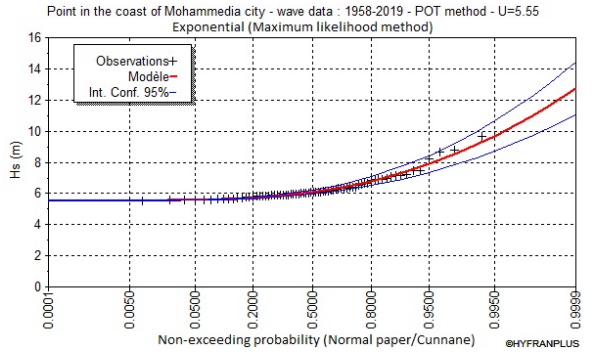
The sensitivity study results highlights the following points:

- The GEV model with the estimation of the distribution parameters by the weighted moment method presents the biggest sensitivity to the variation of the threshold values.
- The asymptotic trend of the forecasting model remains insensitive to the change of the threshold. Indeed, the Exponential model remains an excellent graphical adjustment of the selected data for the different thresholds values.
- The extreme values of thresholds (u < 4.5m and u > 5.5m) generate significant disparity in the forecasting extreme waves.
- The values of thresholds in the interval [4.5; 5.5m] generate insignificant disparity in the forecasting extreme waves. Hence, the stability interval of forecasting the hundred-years return period wave height indicates that the threshold must be within the range between 4.5m and 5.5m approximatively.
- The threshold stability interval corresponds to the interval number (Na) between 1.2 and 3.6 in the case of Mohammedia's point and from 1.2 to 4 for Safi's point. these results confirm the importance to keep the N_a number between 2 to 5 as recommended by Mazas and Hamm [5]
- The too-high threshold implies an erroneous forecasting model for extreme waves; this observation is validated as mentioned by Coles [8].

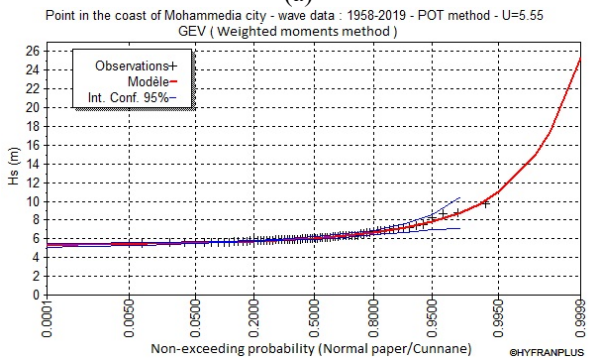
The parent distribution is a local characteristic that varies from one zone to another [15]. In the study area between Mohammedia and Safi the best fitting model to the data is the exponential model. This threshold sensitivity study showed that an overestimation or

underestimation of the censoring threshold led to an erroneous forecasting model.

It should be noted that for the data in the point in Mohammedia's coast. The threshold U_3 = 5.55 seems to be a transition between the two models: exponential and GEV, Figure 7 presents the two adjustments. Indeed, the fitting tests favor the GEV model. However, the graphical comparison indicates that this theoretical distribution presents an overestimation of the extreme values.



(a)



(b)

Figure 7: (a) Results of graphical adjustment for exponential model for threshold U_3 = 5.55. (b) Results of graphical adjustment for GEV model for threshold U_3 = 5.55

4 Conclusion & Discussion

Statistical study of wave data over a 62-years data period assumes that the data is stationary and there is no long-term variation due to climate change. This hypothesis is not valid according to studies of wave climate evolution during the 20th century [16]. Therefore, this study could be extended to examine new models of the trend of extreme values depending on climate change.

Examination of the available data for the 62 years does not reveal a worsening of extreme events in the study area. Hence, the hypothesis made concerning the identically distributed data is a priori a security hypothesis. Extending this analysis to the entire Moroccan Atlantic coast will allow the decomposition of this coast into homogenous zones. Each zone can be described with an adequate theoretical model for forecasting extreme values.

The regional analysis for the determination of the censorship threshold allowed the definition of the interval of the most suitable values. This analysis could be extended to other points for the confirmation of the results and the determination of the intervals of the thresholds of censorship on the entire Moroccan Atlantic coast.

REFERENCES

1. Walden Nasmyth P 1970. Oeanique turbulence. The University of British Columbia.
2. Ferreira A and De Haan L 2015 On the block maxima method in extreme value theory Annals of Statistics. Vol. 43. pp. 276-298.
3. Pickands J 1975 Statistical inference using extreme order statistics The annals of statistics.
4. Mathiesen M. Goda Y. Hawkes P J. Mansard E. Jesús Martín M. Peltier E. Thompson E F and Van Vledder G 1993 Case studies of extreme wave analysis: a comparative analysis Journal of Hydraulic Research. pp. 803-814.
5. Mazas F and Hamm L 2011 A multi-distribution approach to POT methods for determining extreme wave heights Coastal Engineering. Vol. 58. pp. 385-394.
6. Mazas F and Hamm L 2010 Théorie statistique du renouvellement pour la détermination des houles extrêmes - Partie 1 : le point sur les méthodes disponibles La houille blanche.
7. Caires S and Sterl A 2005 100-Year Return Value Estimates for Ocean Wind Speed and Significant Wave Height from the ERA-40 Data journal of climate. pp. 1032-1048.
8. Coles S 2001 An Introduction to Statistical Modeling of Extreme Values. Springer.
9. El adlouni S et Bobée B 2014 Analyse Fréquentielle avec le logiciel HYFRAN-PLUS.
10. Cunnane C. 1978 Unbiased plotting positions - A review Journal of hydrology. 37. pp. 205-222.
11. El Adlouni S. Bobé B and Ouarda T 2008 On the tails of extreme event distributions in hydrology Journal of Hydrology. Vol. 355. pp. 16-33.
12. PEARSON. K. 1900. On the criterion that a given system of deviations given system of deviations of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine Series. Volume 5. pp. 157-175.
13. Hirotugu. A. 1974. A New Look at the Statistical Model Identification. IEE Transactions on automatic control. Vol AC-19. NO. 6. pp. 716-723.
14. Schwarz. G.1978. Estimating the dimension of a model. The Annals of Statistics. Vol 6. no. 2. pp. 461-464.
15. Goda Y 2000 Random seas and design of maritime structure. Singapore: World scientific.
16. Alvaro S . Ralf W . Arno B. Andreas S . Lennart B and Heinz G 2013 Projection of Global Wave Climate Change toward the End of the Twenty-First Century JOURNAL OF CLIMATE. Vol. 26. pp. 8269-8288.