International Environmental Modelling and Software Society (iEMSs)
2010 International Congress on Environmental Modelling and Software
Modelling for Environment's Sake, Fifth Biennial Meeting, Ottawa, Canada
David A. Swayne, Wanhong Yang, A. A. Voinov, A. Rizzoli, T. Filatova (Eds.)
http://www.iemss.org/iemss2010/index.php?n=Main.Proceedings

Fast Variable Selection for Extreme Values

A. Phatak[ab], **C. Chan**[ab] **and H. Kiiveri**[ac]

[a]*CSIRO Mathematics, Informatics and Statistics, Private Bag 5, Wembley, WA, Australia*
(Aloke.Phatak@csiro.au, Carmen.Chan@csiro.au, Harri.Kiiveri@csiro.au)

[b]*CSIRO Climate Adaptation Flagship*

[c]*CSIRO Transformational Biology Platform*

Abstract: In this paper, we outline and explore the use of RaVE, a sparse variable selection method that can be used for selecting variables when modelling extreme values. We illustrate its use by modelling the location parameter of a series of length 49 of annual rainfall maxima at two stations in North-West Western Australia. The ensemble of potential predictors consists of 1980 atmospheric variables. Preliminary results show that RaVE can produce parsimonious yet sensible models. Future work will focus on devising criteria to select the best choice of hyperparameters and on methods to choose groups of spatially contiguous variables so as to aid interpretation.

Keywords: generalized extreme value distribution; variable selection; extreme rainfall

1 INTRODUCTION

Though much of the discussion of the potential impacts of climate change has been cast in terms of changes in, for example, *average* temperatures or rainfalls, there is an increasing recognition that we need to understand the behaviour of climate *extremes*. In particular, the climate adaptation community has recognized that understanding the drivers, frequency, and impacts of climate extremes is central to devising adaptation strategies in a wide range of areas: human health; managing natural ecosystems; designing built environments; managing water resources; and many others. A recent example of the severe impact of an extreme event that may well be a harbinger of future climate is the European heatwave of 2003, in which "unusually large numbers of heat-related deaths were reported in France, Germany, and Italy." [Stott et al., 2004].

Central to the statistical modelling of climate extremes is extreme value analysis based on the generalized extreme value (GEV) distribution [Coles, 2001]. The objective here is to model the behaviour of a process at unusually large (or small) levels with a view to estimating what extreme values might occur in the future. In practical terms, this means postulating and fitting plausible models for the parameters—location, scale, and shape—of the GEV distribution. A summary of statistical modelling of extremes can be found in reviews by Katz et al. [2002] and Katz et al. [2005].

A common class of statistical models for modelling trends, or more generally, nonstationarity, in extremes includes regression models for the GEV parameters. In such models, covariates may include sine and cosine functions in time to model diurnal, seasonal, or annual variability, but they may also include variables that represent larger scale synoptic features or other environmental processes. For example, Furrer and Katz [2008] include an ENSO (El Niño Southern Oscillation) index as a covariate in modelling extreme rainfall, as do Mendez et al. [2007] in modelling monthly extreme sea levels. The *choice* of explanatory variables to use is often guided by domain-specific knowledge, but in many applications, such knowledge may not be available or there may well be little consensus on the the most appropriate choice. In such circumstances, it would be useful to have an automatic variable selection method.

Relatively little work has been done, however, on automatic variable selection methods for GEV-distributed random variables. In part, this reflects the fact that the GEV distribution is not a member of the the exponential family of distributions [Coles, 2001], for which variable selection methods exist and are widely used. Furthermore, there is often a paucity of extreme value data, especially, for example, if we consider annual extrema, and many more potential explanatory variables. This problem is sometimes known in the literature as the '$n \ll p$ problem', and has led to the development of new implicit variable selection methods.

In this paper, we introduce and the demonstrate the use of a fast, sparse variable selection method for selecting potential predictors for extremes. The method, known as RaVE (for Rapid Variable Elimination), is due to Kiiveri [2008]. It is related to other implicit variable selection methods such as well-known Lasso [Tibshirani, 1996], but it provides a much more flexible framework for model fitting and selection. We demonstrate proof-in-principle use of RaVE for modelling extremes by using it to select potential explanatory variables for modelling the location parameter of GEV distributions describing the annual maximum rainfall at two stations in North-West Western Australia (NWWA). The initial ensemble of predictors, from which a much smaller subset was selected, comprises a total of 1980 atmospheric variables from NCEP-NCAR reanalysis data [Kalnay et al., 1996]; there were 20 variables at each of 99 (11×9) grid points, for a total of 1980 potential predictors.

2 Variable Selection for the Generalized Extreme Value Distribution

2.1 Generalized Extreme Value Distribution

This section provides a brief introduction to the GEV distribution; more details may be found in the monographs of Coles [2001] and Davison [2003, Section 6.5.2], from which this summary is drawn.

Let Y represent the annual or seasonal maximum rainfall obtained from a series of daily rainfall measurements. Then, under certain conditions, the limiting distribution of Y is a member of the generalized extreme value distribution family, which has cumulative distribution function given by

$$F(y) = \exp\left\{-\left[1+\xi\left(\frac{y-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, \tag{1}$$

where $1+\xi(y-\mu)/\sigma > 0$, and where the parameters satisfy $-\infty < \mu < +\infty$, $\sigma > 0$, and $-\infty < \xi < +\infty$. The parameter μ is the *location* parameter; σ is the *scale* parameter; and ξ is the *shape* parameter. Depending on the value of ξ, the GEV distribution can take be a member of one of three families of distributions: Gumbel ($\xi = 0$), Frechét ($\xi > 0$), or Weibull ($\xi < 0$). In many, though not all, rainfall series, the shape parameter takes on a small positive value [Koutsoyiannis, 2004], which leads to a heavy-tailed Fréchet distribution.

If the characteristics of a series of n rainfall maxima $\{Y_t\} = Y_1, Y_2, \ldots, Y_n$ do not change throughout time, then we might postulate a very simple model in which a GEV with constant parameters gives rise to $\{Y_t\}$. It may well be, however, that because of seasonal effects or changing climate patterns, the parameters depend on time or other covariates. Consequently, it is necessary to postulate a model for how the parameters change with time. For example, if yearly maxima appear to be increasing linearly with time, a plausible model might be $Y_t \sim \text{GEV}(\mu(t), \sigma, \xi)$, where $\mu(t) = \beta_0 + \beta_1 t$. More generally, we can write $\theta = g(\boldsymbol{x}^T\boldsymbol{\beta})$, where θ is any of μ, σ, or ξ, $\boldsymbol{x}$ is a vector of covariates, $\boldsymbol{\beta}$ is a vector of coefficients to be estimated, and g is the inverse-link function. For example, in modelling maximum sea-levels, Mendez et al. [2007] postulate linear models for all three GEV parameters that use sines and cosines terms to model regular cycles, along with a covariate describingthe variability in ENSO.

In the example described in this paper, we model only the location parameter as a linear function of predictors, and assume that the scale and shape parameters remain constant. The challenge is in selecting a small number of covariates x from a much large ensemble of potential predictors.

2.2 Variable Selection Using RaVE

RaVE [Kiiveri, 2008] is a general statistical engine that provides a mechanism for eliminating redundant variables in a wide variety of existing statistical models such as generalized linear models, multiclass logistic regression, proportional hazards survival models, and many others. This mechanism is based on the notion of model *sparsity*, and hence allows for sensible estimation of parameters when the number of observations (n) is much less than the number of potential explanatory variables (p).

Let $L(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{\phi})$ be the log-likelihood function for a model that we would like to fit in order to relate a response vector $\boldsymbol{y}$ of length n to an $n \times p$ matrix of data $\boldsymbol{X}$. The p-vector $\boldsymbol{\beta}$ contains the parameters of primary interest, and the q-vector $\boldsymbol{\phi}$ contains the parameters of secondary interest. In the context of the problem described in this paper, $\boldsymbol{y}$ is a vector of n annual rainfall maxima, $\boldsymbol{X}$ is a matrix of values of p atmospheric variables. Since we are only interested in modelling the location parameter of the annual maxima, we postulate a linear model of the form $\mu_i = \boldsymbol{x}_i'\boldsymbol{\beta}$, where $\boldsymbol{x}_i$ is the ith row of $\boldsymbol{X}$. The vector $\boldsymbol{\phi}$ contains the scale and shape parameters, which we assume to be constant. The problem then becomes one of selecting a much smaller set of variables that may be influencing the nonstationary behaviour of the annual maxima.

In conventional statistical models such as linear or logistic regression, Lasso [Tibshirani, 1996] is a widely used for variable selection when $n \ll p$. In Lasso, the log-likelihood is penalized by the addition of an $L1$-norm penalty term of the form $\lambda \sum_{i=1}^{p} |\beta_i|$, where λ, known as the regularization parameter, controls the sparsity of the solution, that is, the number of variables that are selected. The Lasso solution can also be interpreted as the Bayes posterior mode under independent double-exponential priors for the coefficients [Tibshirani, 1996]. By contrast, RaVE is formulated explicitly as a Bayesian hierarchical model. The prior for the elements of $\boldsymbol{\beta}$ is specified in such a way as to capture the underlying assumption that only a few of the coefficients are likely to be non-zero. Kiiveri's formulation leads to the prior of the form

$$p(\beta_i) = \left[\frac{2^{(0.5-k)}}{\sqrt{\pi}\Gamma(k)}\right] \frac{\delta K_{(0.5-k)}(\delta|\beta_i|)}{(\delta|\beta_i|)^{(0.5-k)}},\ \delta = \sqrt{\frac{2}{b}} \tag{2}$$

where K denotes a modified Bessel function of the third kind (which is a rapidly decaying function), and Γ denotes the gamma function. An expectation-maximization algorithm is used to obtain maximum *a posteriori* estimates of the coefficients β_i. Algorithmic details may be found in Kiiveri [2008], and software to carry out the calculations is available as an R-language [R Development Core Team, 2008] package *GeneRaVE* [CSIRO Bioinformatics, 2008].

Using RaVE for modelling GEV-distributed variables requires the user to formulate an expression for the (log-)likelihood $L(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{\phi})$, and the first and second derivatives of $L(\cdot)$ with respect to the linear predictor $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$. These expressions, and other implementation details, may be found in Phatak [2009]. Note that RaVE has two hyperparameters k and b (or δ) that have to be selected. In linear regression or logistic regression, N-fold cross-validation can be used to determine optimal values of k and b, but in this initial exploration of the use of RaVE for variable selection for the GEV, we calculate a range of possible solutions and examine diagnostic plots to assess how well the models fit.

There are two ways in which RaVE can be used: first, as a means of selecting variables *and* estimating their associated coefficients; or second, as a 'pre-filter' to select which variables to retain, followed by, say, maximum likelihood estimation of the coefficients. We use the latter method.

3 RAINFALL MAXIMA IN NWWA

3.1 Rainfall Data

We obtained daily rainfall records from 19 stations in North-West Western Australia for the years 1958–2007, although we discuss results for only two stations, one each in the Kimberley (Station 1) and Pilbara (Station 41) regions of NWWA. Furthermore, we consider only the wet season, which runs from November to April. During this period, rainfall in these regions is driven by tropical systems that may include tropical cyclones, monsoonal rains, and tropical depressions [Land and Water Australia, 2008]. On any given day, the reported rainfall is the cumulative rainfall in the 24-hour period up to 9:00 a.m. on that day.

3.2 Atmospheric Data

For the years 1958–2007, atmospheric data was obtained from the NCEP-NCAR reanalysis data set Kalnay et al. [1996] at a $2.5^\circ \times 2.5^\circ$ latitude-longitude grid resolution across the area defined by 10° to 30° latitude S and 107.5° to 132.5° longitude E, for a total of $99 (= 9 \times 11)$ grid points (the grids are shown in Figs. 2 and 3). The group of potential predictors consisted of 20 atmospheric variables at each grid point: air temperature, dew-point temperature depression, geopotential height, specific humidity, and east-west and north-south components of wind speed at 500, 700, and 850 hPa, as well as mean sea-level pressure and total-totals. Thus, there was a total of $1980 (= 20$ predictors/grid point $\times$ 99 grid points$)$ potential explanatory variables, roughly forty times more than the number of annual maxima (49). There may well be additional variables that we *ought* to be considering in order to increase the information content of this ensemble, but it was chosen with the much longer term objective of studying the downscaling of rainfall extremes using atmospheric predictors from general circulation models.

3.3 Assessing Nonstationarity in Extremes

Before carrying out any regression modelling, we first assessed the degree of nonstationarity of the rainfall maxima series at the two stations using visual diagnostics.

Figure 1 shows the annual maxima at Station 1; superimposed on the plot is a smoothed estimate of the location parameter, obtained using the the local polynomial fitting method of Davison and Ramesh [2000] with a bandwith of 0.3. Though Davison and Ramesh [2000] describe a bootstrapping method for automatic choice of bandwidth, they also suggest that when exploratory data analysis is carried out with a view to suggesting hypotheses to be explored further, it is equally useful to gain insights into the data by trying different values of the bandwidth. For these data, different bandwidths lead to smooths that suggest the same conclusion: there appears to be a trend, or other nonstationary behaviour, in the location parameter.

We can also assess nonstationarity by examining whether a model in which the parameters are constant is plausible. Figure 1 also shows probability and quantile plots of the residuals after fitting the parameters by maximum likelihood. Although the probability plot appears to be linear, the quantile plot indicates some lack of fit, especially for very large values of rainfall. Hence, combining the results of the visual examination of the plots in Fig. 1, we tentatively conclude that annual maxima at Station 1 are nonstationary and proceed to search for covariates that may explain this behaviour. Plots such as those in Fig. 1 for Station 41 (not shown) yield a similar result.

3.4 Variable Selection

Before we describe the results of RaVE for Stations 1 and 41, it is worth asking the question, "how many variables should *any* variable selection method be expected to choose in this instance, given

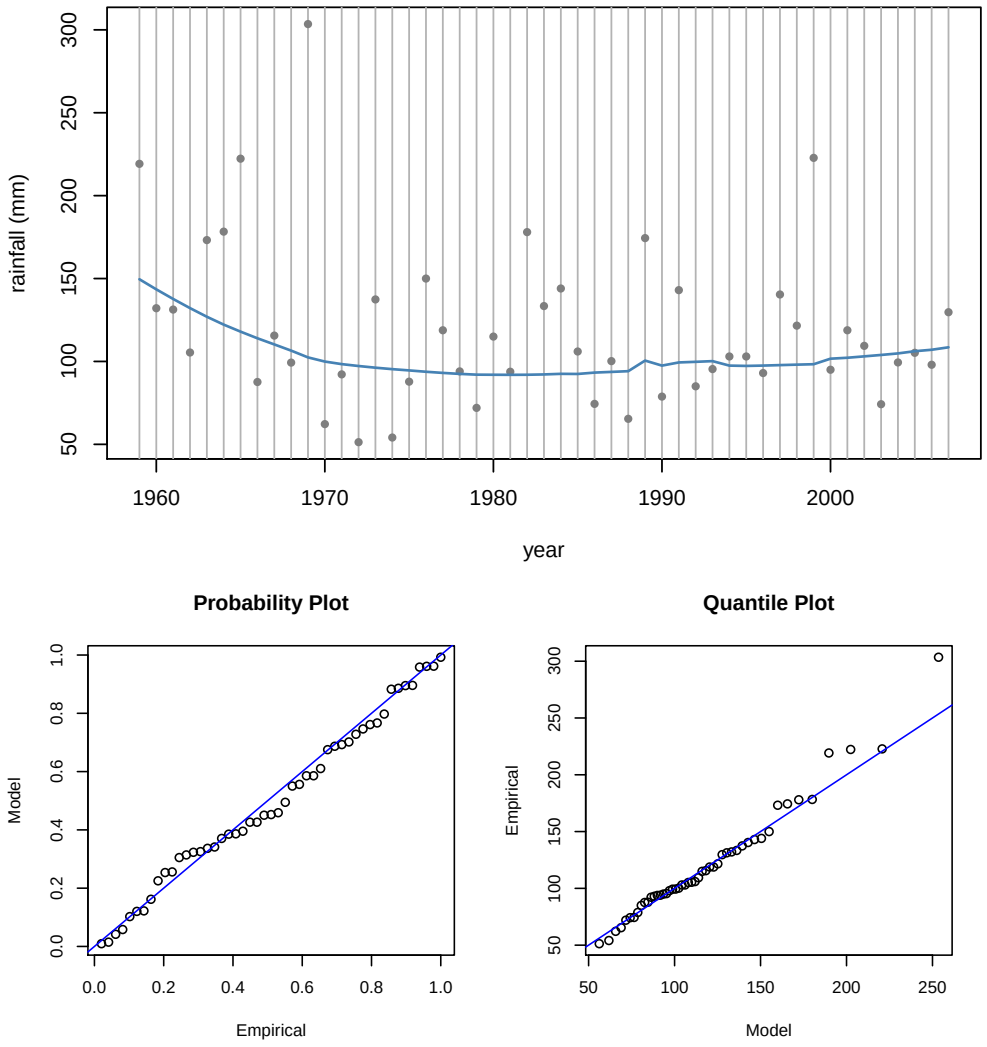

Figure 1: (Top) Annual rainfall maxima at Station 1 (Kimberley) during the wet-season. Solid line indicates the smoothed location parameter, calculated using the local polynomial fitting method of Davison and Ramesh [2000]. (Bottom) Probability plot (left) and quantile plot (right) of residuals after fitting a constant location, scale, and shape parameter to annual rainfall maxima at Station 1.

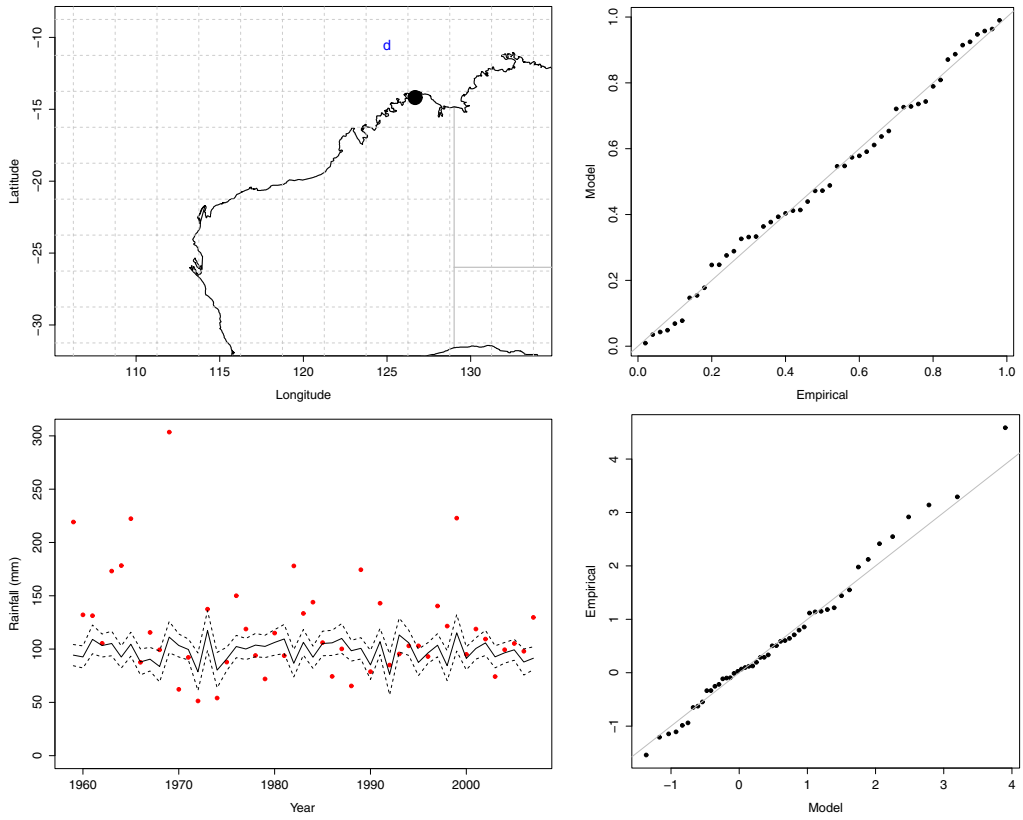

Figure 2: Station1: (Top left) Map of NWWA showing the grid points at which NCEP-NCAR data were obtained, the station location (black circle), and the variable that was selected for a b/k combination of $10^4/0.1$; (Top and bottom right) Probability and quantile plots after maximum likelihood estimation of the coefficient of the single variable chosen; (Bottom left) Estimated location and approximate 95% pointwise confidence intervals.

that there are only 49 observations?" Though a precise number cannot be specified, clearly, the answer is "very few"! A rough rule-of-thumb states that 5–10 observations are required for every parameter estimated; hence, taking into account the scale and shape parameter leaves sufficient information for between 5 and 10 additional parameters. Indeed, as we observed, models that contained more than about 10 variables severely overfitted the data.

We fitted RaVE models for values of $b \in \left\{1, 10, \ldots 10^6\right\}$ and $k \in \{0.1, 0, 2, \ldots, 0.9\}$. The number of variables selected is relatively insensitive to the value of the hyperparameter b, but very sensitive to the value of k. This is in contrast to sparse logistic regression for modelling rainfall occurence, where the number of variables selected is insensitive to k [Phatak et al., 2009]. For a fixed value of b, the number of variables selected quickly increases from one to greater than 10 as k is varied from 0.1 to 0.5.

Figure 2 shows results for Station 1 for $b = 10^4$ and $k = 0.1$. A single variable (dew point temperature depression at 850 hPa at 10°S, 125°E) was chosen, and is shown on a map of NWWA along with the station location. Figure 2 also shows the fitted location parameter along with quantile and probability plots obtained by maximum likelihood estimation of the coefficient of dew point temperature depression. The value of the coefficient is -4.33 with a standard error of ± 1.7. If we compare the diagnostic plots in Fig. 2 with those in Fig. 1, we can see that the addition of a covariate appears to have explained some of the variability in the annual maxima. In particular, the quantile plot lies closer to the 45° line. Although we do not show the corresponding plots here for lack of space, fitting a model with $k = 0.4$ leads to 9 variables being selected, overfitting has occurred and the performance is much degraded—although the estimated location tracks the

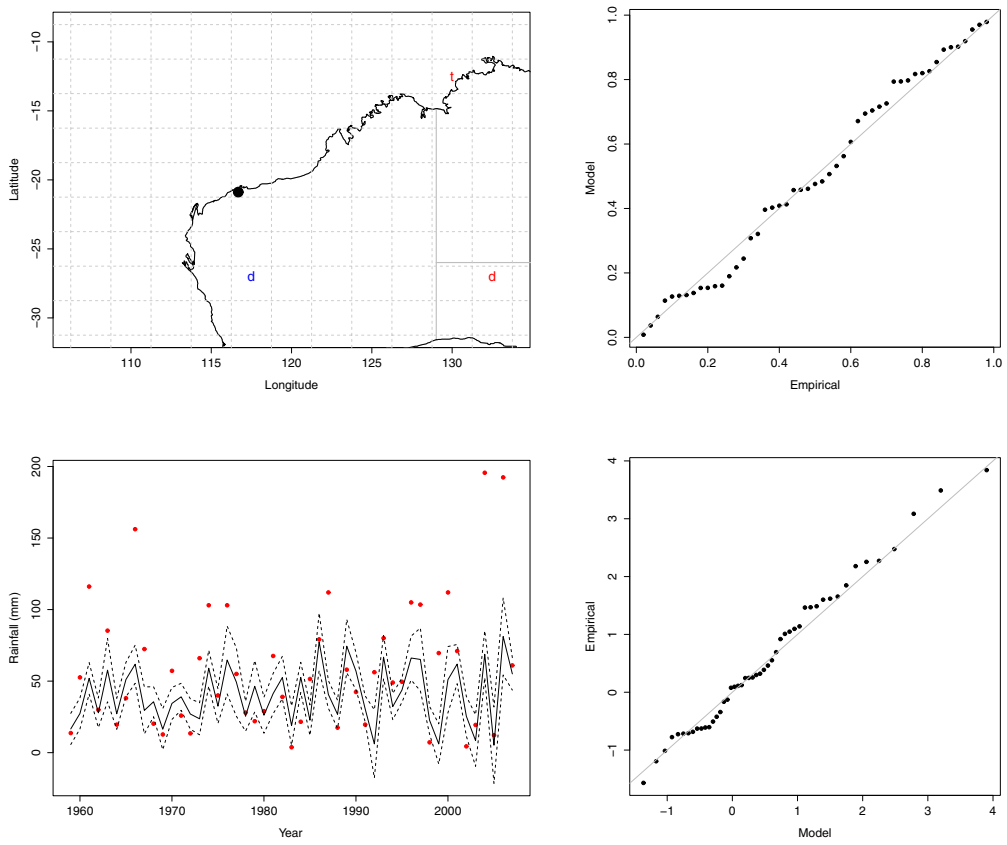

Figure 3: Station 41: (Top left) Map of NWWA showing station location (black circle), and the variables that were selected for a b/k combination of $10^4/0.3$; (Top and bottom right) Probability and quantile plots after maximum likelihood estimation of the coefficient of the three variable chosen; (Bottom left) Estimated location and approximate 95% pointwise confidence intervals.

annual maxima more closely, the quantile and probability plots show clear lack of fit, and the approximate 95% confidence intervals of many of the cofficients include zero.

For Station 41 in the Kimberley region, Fig. 3 shows the analogous series of plots to those in Fig. 2. Here, a b/k combination of $10^4/0.1$ selected three variables and gave the most sensible results, based on a visual examination of diagnostic plots. Again, more variables were selected for larger values of k. The three variables selected were dew point temperature depression at 500 hPa at 27.5°S, 117.5°E (coef: -1.4, se: 0.4) and at 27.5°S, 132.5°E (coef: 0.8, se: 0.3), and total-totals at 12.5°S, 130°E (coef:5.8, se: 1.8).

4 Discussion and Conclusions

Our intent in this short communication has been to introduce and demonstrate proof-in-principle use of RaVE for variable selection for extreme values, a topic that has, to the best of our knowledge, yet to be tackled in either the literature on variable selection or modelling of extreme values. The example we have discussed here is a particularly difficult one because the number of potential explanatory variables (1980) is much larger than the number of observations (49). Nevertheless, the results described here show that RaVE can produce parsimonious models that yield sensible results. Given the ensemble of predictors that we started with, RaVE selected a single dew point temperature depression near Station 1. Given what is known about the predominantly tropical rainfall mechanisms in the Kimberley, such a result is plausible. For Station 41 in the Pilbara,

the three variables selected may well reflect the multiplicity of mechanisms bringing rain to the Pilbara in the wet season, but further investigation is required.

Routine use of RaVE for modelling extreme values requires model selection and goodness-of-fit measures. Cross-validated likelihood may provide a partial solution. In addition, it would help the interpretation of the results if spatially contiguous *groups* of variables were chosen instead of variables at individual grid points. These and other topics will be the subject of future research.

ACKNOWLEDGMENTS

We would like to thank Steve Charles of CSIRO Land and Water for providing us with the rainfall and NCEP-NCAR reanalysis data used in this study.

REFERENCES

Coles, S. *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics. Springer, New York, 1st edition, 2001.

CSIRO Bioinformatics. *GeneRaVE*. `https://www.bioinformatics.csiro.au/GeneRave/index.shtml`, 2008.

Davison, A. C. *Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, August 2003.

Davison, A. C. and N. I. Ramesh. Local likelihood smoothing of sample extremes. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62(1):191–208, 2000.

Furrer, E. M. and R. W. Katz. Improving the simulation of extreme precipitation events by stochastic weather generators. *Water Resources Research*, 44(12), DEC 27 2008.

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, A. Leetmaa, B. Reynolds, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. Mo, C. Ropelewski, J. Wang, R. Jenne, and D. Joseph. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77(3):437–471, 1996.

Katz, R. W., G. S. Brush, and M. B. Parlange. Statistics of extremes: Modeling ecological disturbances. *Ecology*, 86(5):1124–1134, 2005.

Katz, R. W., M. B. Parlange, and P. Naveau. Statistics of extremes in hydrology. *Advances in Water Resources*, 25(8-12):1287–1304, 2002.

Kiiveri, H. A general approach to simultaneous model fitting and variable elimination in response models for biological data with many more variables than observations. *BMC Bioinformatics*, 9(1):195, 2008.

Koutsoyiannis, D. Statistics of extremes and estimation of extreme rainfall: II. Empirical investigation of long rainfall records. *Hydrological Sciences Journal–Journal des Sciences Hydrologiques*, 49(4):591–610, AUG 2004.

Land and Water Australia. Weather drivers in Western Australia. `http://www.managingclimate.gov.au/Publications_and_Tools/Communicating_Climate_Change/index.html`, August 2008. Downloaded on 1 February 2010.

Mendez, F. J., M. Menendez, A. Luceno, and I. J. Losada. Analyzing monthly extreme sea levels with a time-dependent GEV model. *Journal of Atmospheric and Oceanic Technology*, 24(5): 894–911, MAY 2007.

Phatak, A., B. Bates, and S. Charles. Statistical downscaling using sparse variable selection methods. Submitted to *Environmental Modelling & Software*, December 2009.

Phatak, A. Using GeneRaVE for Extremes: An Initial Attempt. Technical Report CMIS 09/96, CSIRO, June 2009.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

Stott, P. A., D. A. Stone, and M. R. Allen. Human contribution to the European heatwave of 2003. *Nature*, 432(7017):610–614, December 2004.

Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.