ABSTRACT
Resource adequacy is the ability of the electric system to supply the sufficient resources to satisfy demand reliably at all times. One key challenge for optimally estimating resource adequacy is that demand is difficult to predict with certainty, particularly during summer in California where the main factor affecting it is the temperature and the demand sensitivity to it. Cooling demand responses to extreme high summer temperatures generate sudden changes and large variability in demand testing the ability of the system to meet its highest demand and so predicting it with some degree of certainty is the ultimate goal in resource adequacy planning. This paper discusses how to implement the Markov chain Monte Carlo (MCMC) procedure with the SAS® Enterprise Miner™ platform to effectively fit Bayesian predictive modeling to the extreme cooling demand distribution to accurately assess risk in electricity summer demand. Specifically, it discusses how extreme value theory (EVT) analysis of extreme demand can be used to estimate its probability distribution and calculate extreme demand quantiles for low probabilities. In addition, the paper discusses how to extend SAS® Enterprise Miner™ capabilities by deploying an extension node under the model menu to perform EVT analysis.

INTRODUCTION
Clearly California is summer-peaking – a response dominated by summer space cooling needs – and extreme high summer temperatures cause significant impacts on cooling demand. Demand is difficult to predict with certainty and frequent and sudden excessive demand puts the ability of the system to meet its highest demand at risk. Predicting demand with some degree of certainty is the ultimate goal in resource adequacy planning; the prime concern is to model and estimate the probability of excessive demand caused by extreme and/or rare temperatures that cannot be entirely explained or captured reliably by existing modeling techniques. It is essential to know that unexplained portion of demand known as residuals, the difference between actual demand and the one suggested by the regression model, the positive residuals correspond to under prediction of demand and assessing its risk is absolutely crucial for resource adequacy planning.

Extreme value theory and the estimation of the parameters of the EVT model based on Bayesian method using the MCMC procedure provide a useful risk measure because EVT provides a more appropriate distribution to fit extreme and rare events and Bayesian estimation accounts for uncertainties on the parameters of an assessment of the likelihood of under prediction of demand. The analysis is implemented in SAS Enterprise Miner 6.1 by creating an EVT node. The node extends the functionality of Enterprise Miner and offers an easy way to modify it to one’s analytical needs and share it with others.

DATA DESCRIPTION AND EARLIER MODELING EFFORTS
The data set consists of daily electricity system peak demand in megawatts (MW) between 1 pm and 6 pm from June 15 to September 15, 2010, 93 days of summer, from the publicly available database of the California Independent System Operator (CAISO). For brevity here, the data set only includes San Diego Gas and Electric (SDGE) utility. Matching daily maximum and minimum temperatures in degrees Fahrenheit (deg F) come from representative weather stations across SDGE planning area taken from the National Climatic Data Center and weighted by air-conditioning saturation to represent the temperatures for the planning area.

![Graphs showing daily system demand by temperature, by month, and cooling degree days by month for summer 2010. Weekdays and LOESS smoothing curve shown as blue and solid line and weekends as red and dashed line.](image_url)
The scatter plots in Figure 1A, 1B, and 1C with LOESS smoothing curve show daily system demand by maximum daily temperature and by month and cooling degree days by month for summer 2010, respectively. Daily system peak clearly varies nonlinear with maximum daily temperature, and thus with cooling degree days, and depends significantly on the day of the week, with higher loads on weekdays (blue and solid line) than on weekends (red and dashed line). To capture day of the week effect, a variable DOW takes value one on weekends and holidays and zero otherwise. Cooling degree days, CDD65, defined as the positive difference between daily mean temperature and the reference temperature of 65 (deg F) and zero otherwise, exhibits intra-seasonal variation not only with the mean but also with its variance. The slightly decline in system demand at five cooling degree days corresponds to a daily mean temperature of 69.5 to 70 (deg F).

Despite the influence of temperature values and other temperature-derived variables, of a set of Fourier series of sine and cosine terms selected based on Akaike’s information criterion, and of autoregressive error terms selected on significant t-tests from a stepwise autoregression with the backstep option, a variable selection using the Variable node with the R-square and AOV16 variables (binned variables) options reveals CDD65 and DOW showing the highest relation in describing the variation in logarithmic daily peak demand. Logarithm transformation fits better the data. For this reason, CDD65 and DOW are the only two significant explanatory variables considered here.

A stepwise autoregressive model of first order using the SAS Code node establish the relationship between daily peak demand and CDD65 and DOW explanatory variables. Based on the results, the variables capture the variations in peak demand remarkably well. The autoregressive model has an r-square value of 90.87%, and includes .0316 for CDD65 and -.1101 for DOW as the estimated coefficients. The coefficient of DOW indicates 10.43% lower peak loads on weekends and holidays than on weekdays and the coefficient of CDD65 indicates daily demand increases by 9.94% for every additional 3 cooling degree days.

The residual time series for the autoregressive regression represent the difference between actual and predicted peak loads and so the portion of the actual peak demands that the model fails to capture or explicitly explain. The positive residuals offer the most challenging and interesting discontinuous time series to analyze for resource adequacy planning because in summer the interest lies in assessing the risk of under prediction of peak demand during periods of extreme temperature conditions. To do this, it is necessary to model the extreme tails of the positive residual distribution and extreme value theory seems better suited for that purpose.

**EXTREME VALUE THEORY**

Extreme value theory models the extreme tails of a distribution (Coles, 2001) and quantifies the behavior of a process at both extremes of a distribution. EVT estimates the probability of an extreme and/or rarely observed event, i.e., an extreme positive residual due to an extreme and/or rare temperature. The interest here is to estimate the probability of an extreme positive residual from the stepwise autoregressive model and to add it to the estimated daily load to yield peak demand under a specific level of risk.

Identifying extreme positive residuals using the statistics of extremes (Coles, 2001) involves fitting either the generalized extreme value (GEV) distribution that covers three limiting distributions (Gumbel, Fréchet, and Weibull) to maximum residuals of a sequence of positive residuals or the generalized Pareto (GPD) distribution to positive residuals exceeding a high threshold.

The GEV distribution is given by the likelihood function

\[ G(x; \mu, \phi, \xi) = \exp \left\{ - \left( 1 + \frac{x - \mu}{\phi} \right)^{-\frac{1}{\xi}} \right\} \]

where \( \xi \) and \( \phi > 0 \) denote the location and scale parameters, and \( \xi \) represents the shape parameter that determines the tail of the distribution, light-tailed \( \xi \to 0 \) for Gumbel type of distribution, like the normal distribution, bounded \( \xi < 0 \) for Weibull, like the beta distribution, and heavy-tailed \( \xi > 0 \) for Fréchet, like the Student t distribution. The GPD distribution is given by the likelihood function

\[ G(x; \tau, \sigma, \epsilon) = \left\{ 1 - \left( 1 + \frac{x - \tau}{\sigma} \right)^{-\frac{1}{\epsilon}} \right\} \]

where \( \sigma > 0 \) is the scale parameter and \( \tau \) the threshold. One practical application for resource adequacy is to estimate the value that a positive residual will exceed with a given probability. The value is estimated by the quantile function that is obtained by inverting the GEV and GPD distributions as

\[ Z_p = G(x; \mu, \phi, \xi)^{-1} \left( 1 - p \right) \]

\[ Z_p = G(x; \tau, \sigma, \epsilon)^{-1} \left( 1 - p \right) \]

where \( p \) is the probability 1/T, T referring to years, corresponding to the \( (1 - p) \) quantile.

Two important issues before applying EVT distribution relate to the assumption of the data being independent and identically distributed (IID) and the choice of the threshold. For peak load and weather data in particular, it seems unlikely that the probabilities of extremes show no time dependency. Possible ways for treating non-IID data include declustering and/or incorporating seasonality and trends by linking the parameters of the GEV and GPD distribution to covariates to describe the seasonal cycle and the time of summer and by selecting a higher (lower) threshold. According to Smith (2001), due to the asymptotic distribution of the distributions, a non-IID dataset could be properly
evaluated by a GEV distribution. However, for GPD selecting the threshold and hence leaving sufficient data over the threshold becomes a trade-off between bias in estimating the distribution parameters and sampling variability. Since no diagnostic device provides a clear-cut criterion for choosing the threshold, a preferred one fits the GPD over a range of thresholds and selects a threshold focusing on stability in the parameter estimates and their standard errors.

Among the methods available to evaluate the parameters of the GEV and GPD distributions in estimating how well a parametric model fits the data, a Bayesian method is preferable for data with non-normal distribution, especially when the dataset is short and exhibits the presence of a heavy tail. The Bayesian method is more flexible since it allows fitting non-stationary models with linear and non-linear trends in the location, scale and shape parameters. It is also better for its simplicity and computational efficiency. A Bayesian method estimates the parameters by fitting the GEV and GPD distributions to the positive residuals through maximizing the logarithm of the likelihood function of the distributions and computes their intervals. Then, substituting the estimated parameters into the quantile functions, one obtains estimates of the positive residuals quantiles and their intervals.

THE MCMC PROCEDURE

In a Bayesian estimation, after observing the random data x and specifying a prior probability density function \( p(\theta) \) for the random unknown parameter \( \theta = (\mu, \sigma, \xi) \), the Bayes’ theorem (Gelman et al, 2000) updates the prior distribution into the posterior probability density function \( p(\theta \mid x) \) through the likelihood function \( p(x \mid \theta) \) that represents the conditional density of the data given a particular set of values for the parameters. Thus, the posterior density is proportional to the prior density times the likelihood function. The prior distribution represents information about the unknown parameter before analyzing the data. The likelihood function that here corresponds to either the GEV or GPD distribution is the conditional density of the data given a particular set of values for \( \theta \). The posterior distribution represents information about the parameter after analyzing the data; it represents an updating of the prior information to account for the new information contained in the data. Bayes’ theorem simply brings together the prior and data information to produce the posterior distribution.

In practice, the interest is not in the Bayesian estimators \( \theta = (\mu, \sigma, \xi) \) but in the probability statement or predictive inference that depends on them. Predictive inference answers questions about the probability that the value of a variable over a given value will occur or, the inverse problem, the value that the variable will exceed with a given probability. Within Bayesian inference, the posterior distribution contains information that by averaging over the uncertainty in the parameter estimates allows for the two sources of uncertainty: uncertainty originating from randomness in variables and uncertainty originating from statistical estimation and distribution.

SAS® ENTERPRISE MINER™ AND AN EVT EXTENSION NODE
The PROC MCMC fully integrates into the SAS Enterprise environment by means of an EVT extension node. To create and install an EVT Extension node requires two icons, an XML properties file, and a SAS program. The reference documentation providing detail information about creating and installing an Extension node include the SAS® Enterprise Miner™ 6.1 Extension Nodes: Developer’s Guide, Extending SAS® Enterprise Miner™ with User-Written Node, and applications by Schubert (2008) and Cathie (2011).

ICONS
The EVT node has two identical icons; one appearing on the SAS Enterprise Miner Toolbar under the Model menu and the other one on the process flow diagram when dragging and dropping it from the toolbar. Both share the same filename but the former is a 16x16 pixel image gif file and the latter a 32x32. The gif files came from the SAS supplied icons and modified using Microsoft Paint program. The 16x16 gif file goes in the gift16 folder and the 32x32 in the gift32 folder, both located under the SAS configuration directory for a personal workstation installation C:\SAS\EMDesktop\Lev1\AnalyticsPlatform\apps\EnterpriseMiner\ext\.

XML PROPERTIES FILE
The XML properties file provides information about the EVT node such as the location of the icons and the source code and the selection of the distribution, either GEV or GDP, and their initial parameters necessary to run the PROC MCMC. The file consists of a Component element, a Property Descriptors element, and a View element. The bold blue syntax is unique to the EVT extension node and hence needs to be customized. XML is case-sensitive, so it is important to write the element tags as specified in the example. The Component element has the following:

```xml
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE Component PUBLIC "-//SAS//EnterpriseMiner DTD Components 1.3//EN" "Components.dtd">
```
The value for description appears as a tooltip for the toolbar node. The value for displayname names the node being displayed in the tooltip for the node's icon on the toolbar node and in the process flow diagram. The value for group indicates the SEMMA group where the node appears on the node toolbar. The icon value names the icon files. The name value identifies the node ID in the Properties panel. The prefix value appendes to files generated by the node.

The Variable property descriptor elements appearing in the Property panel consist of the following:

```
<PropertyDescriptors>
  <Property type="String" name="Location" initial="CATALOG"/>
  <Property type="String" name="Catalog" initial="SASHELP.EMmYUTIL.EVT.SOURCE"/>
  <Property type="String" name="VariableSet" displayname="Variables" description="Variable Properties">
    <Control>
      <Choice lists=
        <Choice rawValue="GEV"/>
        <Choice rawValue="GPD"/>
    </ChoiceList>
  </Control>
  <Property type="int" name="NBI" displayname="Number burn-in iterations" description="Number burn-in iterations" edit="Y">
    <Control>
      <Range excludeMax="N" excludeMin="Y" max="1000000" min="1"/>
    </Control>
  </Property>
  <Property type="double" name="Thold" displayname="Threshold" description="Threshold" edit="Y">
    <Control>
      <Range excludeMax="N" excludeMin="Y" max="10000.00" min="0"/>
    </Control>
  </Property>
</PropertyDescriptors>
```

For illustration purposes the syntax includes only the variables “Number burn-in iterations” and “Threshold” and excludes “Number MCMC Iterations”, “Thinning rate”, and “Quantile”.

The initial option in the Catalog property refers to the location of the SAS program. The attributes of the Property elements are as follows. The type value refers to one of the four types being supported, string, boolean, int, or double. The name value refers to a name by which the Property element is referenced in the properties file and in the node’s SAS code. The value for displayname specifies the name of the Property element being displayed in the node’s Property panel. The description value provides a brief description of the Property element appearing in the node’s Property panel. The value for initial specifies the initial or default value for the property. The edit value indicates whether the property's value can be modified.

More functionality is added to the Property elements by the Control elements. The Control elements are ChoiceList and Range. A ChoiceList Control presents the user with a predetermined drop-down list containing the EVT distributions choices. Each choice element has a rawvalue that is passed to the node’s SAS program. A Range Control validates a numeric value entered by the user. A range element has several attributes restricting the range of permissible values to enter: a number representing the minimum (min) and maximum (max) of the range of permissible values and whether to include and/or exclude the minimum (include- and excludeMin) or the maximum (include- and excludeMax) as a permissible values.

The View element organizes properties in the Properties panel.

```
<Views>
</Views>
```
The Views element has a single View element and name as its single attribute with its value equal to Train. Within the View element is a set of PropertyRef elements, one PropertyRef element for each Property element in the properties file. Each PropertyRef has a single nameref attribute that has a value corresponding to the name attribute of one of the Property elements.

**SAS CODE**

The SAS Code include four types of code corresponding to equivalent SAS Enterprise Miner process flow actions and triggered by an internally generated &EM_ACTION macro variable that takes action value CREATE, TRAIN, SCORE, or REPORT. Each flow action relates to a macro — %evt_create, %evt_train, %evt_score, and %evt_report. At this time, no need of a Score code in the EVT application.

```sas
%macro evtmain;
  %if %upcase(&em_action) = CREATE %then %do;
    %evt_create;
  %end;
  %end;
  %else %if %upcase(&em_action) = TRAIN %then %do;
    %evt_train;
  %end;
  %else %if %upcase(&em_action) = SCORE %then %do;
    %evt_score; /* empty */
  %end;
  %else %if %upcase(&em_action) = REPORT %then %do;
    %evt_report;
  %end;
%mend evtmain;
%evtmain;
```

**CREATE CODE**

The CREATE code runs when the node is first placed in the flow diagram and initializes properties with defaults and registers data sets. The %EM_PROPERTY macro sets up the attributes of the node by assigning an action value to each property like the GEV or GPD distribution and the %EM_REGISTER macro registers data sets by associating each one with an unique key and then the dataset is subsequently referenced using the macro variable &EM_USER_KEY. This macro needs to be reinitialized by using the %EM_GETNAME macro when referring to registered data sets in the Train, Score, or Report action.

```sas
%macro evt_create();
  %em_property(name = dist, action = train);
  %em_property(name = nbi, action = train);
  %em_property(name = nmc, action = train);
  %em_property(name = thin, action = train);
  %em_property(name = thhold, action = train);
  %em_property(name = quan, action = train);
  %em_register(key = evt, type = data);
  %em_register(key = iter, type = data);
  %em_register(key = kde, type = data);
  %em_register(key = acf, type = data);
%mend evt_create;
```
The Train code executes PROC MCMC and passes on the results to the Report code. The SAS code for the GEV distribution is given by:

```sas
%macro evt_train();
ods graphics on;
%let y = resid_plus;
proc mcmc data = &EM_IMPORT_DATA outpost = &EM_EXPORT_TRAIN
  nmc = &em_property_nmc thin = &em_property_thin nbi = &em_property_nbi
  seed = 3482947 monitor = (mu xi tau zp) maxtune = 32000 propcov = quanew;
by tac;
ods select postsummaries ess mcse geweke postintervals;
parms mu 0 xi .1;
parms tau 1;
prior mu ~ normal(mean = 0, prec = 1e-04);
prior xi ~ normal(mean = 0, prec = 1e-05);
prior tau ~ gamma(1e-03, iscale = 1e-04);
constraint = 1 + xi * ((&y - mu) / tau);
if constraint ge 0 then loglike =
  - log(tau) - (1 + (1 / xi)) * log(constraint) - constraint ** (- 1 / xi);
else loglike = .;
model &y ~ general(loglike);
beginprior;
zp = mu + (tau / xi) * (- log(&em_property_quan) ** (- xi) - 1);
endprior;
run;
ods graphics off;
%mend evt_train;
```

The ODS GRAPHICS ON turns the Output Delivery System (ODS) on to produce posterior diagnostics plots and the ODS GRAPHICS OFF disables it. The %LET statement modify the name of the resid_plus variable (positive residuals) as the dependent variable (y).

The PROC MCMC statement invokes the procedure and the macro variable &EM_IMPORT_DATA in the DATA = option contains the name of the data set imported into the PROC MCMC and supplied by the prior node in the process flow and the macro variable &EM_EXPORT_TRAIN in the OUTPOST = option names the data containing the posterior sample and passed on to the subsequent node in the diagram.

The &EM_PROPERTY_NMC, &EM_PROPERTY_NBI, and &EM_PROPERTY_THIN macro variables from the Property elements in the properties file correspond to the length of the Markov chain in the NMC = option, the number of iterations in the NBI = burn-in option, and the thinning rate in the THIN = option. The total length for the chain excludes the number of iterations in the burn-in option. The SEED = 3482947 sets the seed for random number generator. The MONITOR = (mu xi tau zp) outputs analysis for the selected parameters, where mu, xi, and tau stand for the location (µ), scale (φ), and shape (ε) parameters of the GEV distribution respectively and zp for the qth quantile of the posterior distribution.

The MAXTUNE = 32000 specifies the maximum number of tuning loops to find a good proposal distribution for the parameters and the PROPCOV = QUANEW option uses the estimated inverse Hessian matrix as the initial proposal covariance matrix. Using these options helps solving non-convergence for EVT modeling for the data.

The BY statement instructs the MCMC procedure to process the dataset observations in groups by the levels of the tac variable, which here has only one level, SDGE. The ODS SELECT statement requests post MCMC summaries for the variables in the MONITOR option, sample sizes, Monte Carlo standard errors, the Geweke z-test diagnostic
for convergence, and the credibility intervals. Since the EVT node is registered as a MODEL in the XML properties file, fit statistics are automatically displayed.

To obtain the three types of plots that the MCMC procedure produces – the trace, the autocorrelation, and the marginal posterior density – the integrated graphical capabilities in SAS Enterprise Miner requires passing on the &EM_EXPORT_TRAIN macro variable to the REPORT code and using two macros, the %EM REGISTER and the %EM REPORT. The %EM REGISTER macro in the CREATE node registers the EVT, ITER, KDE, and ACF data sets containing summary statistics derived from the posterior samples and the %EM REPORT macro uses the registered datasets to specify the contents of a results window display.

The PARM statement identifies the model parameters and their starting values. The PRIOR statement specifies normal prior distribution on the location and shape parameters and gamma prior distribution on the scale parameter. The analysis uses relatively flat or noninformative priors on all parameters, with mean 0 or very close to it and precision at either $10^{-4}$ or $10^{-5}$.

The MODEL defines the dependent variable and declares its sampling distribution using the GENERAL function that indicates a nonstandard distribution taking the value of the logarithm of the likelihood function corresponding to the GEV distribution and its constraint. The statement within the BEGIONPRIOR/ENDPRIOR statements calculates the $q$th quantile of the posterior distribution according to $(1 - p)$ where $p$ is the probability and the value of the macro variable &EM_PROPERTY_QUAN.

The syntax for the GPD distribution differs from the GEV distribution in that contains no reference for the mu parameter, the sigma symbol replaces the tau parameter, and has different logarithm of the likelihood function and quantile function. In addition, it includes the macro value &EM_PROPERTY_THHOLD that sets the threshold value.

```plaintext
if (&y - &em_property_thhold) gt 0 then
constraint = 1 + xi * ((&y - &em_property_thhold) / sigma);
if constraint ge 0 then loglikelihood =
- log(sigma) - (1 + (1 / xi)) * log(constraint);
else loglikelihood = .;
beginprior;
zp = &em_property_thhold +
(sigma / xi) * (((&n / &nobs) *(&em_property_quan) ** (- xi)) - 1);
endprior;
```

**REPORT CODE**

The Report code shows the creation of the posterior sample table, the iteration plot, the posterior density plot, and the autocorrelation plot. The macro %EM_CHECKNAME checks for the existence of macro variables from the Property elements in the properties file. The %EM_GETNAME macro retrieves the name of a dataset prior registered with the macro %EM_REGISTER to a given key in the CREATE node and initializes the &EM_USER_KEY macro variable.

```plaintext
%macro evt_report();
%em_checkmacro(name = em_property_dist, global = y);
%em_checkmacro(name = em_property_nbi, global = y);
%em_checkmacro(name = em_property_nmc, global = y);
%em_checkmacro(name = em_property_thin, global = y);
%em_checkmacro(name = em_property_thhold, global = y);
%em_checkmacro(name = em_property_quan, global = y);

%em_getname(key = evt, type = data);
%em_getname(key = iter, type = data);
%em_getname(key = kde, type = data);
%em_getname(key = acf, type = data);

/* posterior sample table */
data &em_user_evt;
  set &EM_EXPORT_TRAIN;
run;
%em_report(
  key = evt,
  viewtype = data,
  block = Tables,

```
The &EM_EXPORT_TRAIN dataset stores the iteration number, the posterior draws for each parameter being monitored, and the log of the prior density, of the likelihood, and of the posterior density. The &EM_USER_EVT macro variable copies the posterior samples from &EM_EXPORT_TRAIN.

The macro %PGM_ITER reads the &EM_EXPORT_EVT datasets, converts columns to rows by the levels of the tac variable, and saves them in the %EM_USER_ITER. The PROC KDE processes the &EM_EXPORT_TRAIN dataset, computes the kernel density estimate of the posterior density on each parameter being monitored, and saves them in...
the &EM_USER_KDE. The macro %PGM_ACF calls PROC ARIMA to obtain the autocorrelations up to 50 lags from &EM_EXPORT_TRAIN for each parameter and saves them in &EM_USER_ACF.

Then, &EM_USER_EVT, &EM_USER_KDE, and &EM_USER_ACF pass on summary statistics to the %EM_REPORT to produce the posterior sample table and the iteration plots, the posterior density plots, and the autocorrelation plots, respectively.

The %EM_REPORT generates the table and graphs as follows. The key value identifies the data set used to produce the table or graph. The value for viewtype indicates to generate a table or a type of graph. The latticetype value specifies the lineplot to display. The values for x and y specify the variables to be used as x-axis and y-axis variables and the values for latticex and latticey as rows and columns in a lattice. The autodisplay value indicates to display automatically the report whenever the Results viewer of the node is opened. The value for description specifies the title bar of the report and for block indicates that the report should appear under the “Tables” or “Plots” menu item.

RESOURCE ADEQUACY

The following Properties panel, icons, and process flow diagram appear as a result:

Figure 2 Properties Panel of the EVT node and the Model menu toolbar with the 16x16 icon added and the 32x32 icon on the process flow diagram

For this analysis, the PROC MCMC specifies an MCMC sample of 50000, a burn-in sample size of 50000, a thinning rate of 5, and a 90th quantile corresponding to a probability of .1. The thinning rate retains the 5th observation in the chain to end up with a 10000 posterior sample. The prior marginal densities with low precision (high variance) lead to flat marginal priors reflecting ignorance about the parameters. Starting with different initial values makes relatively little difference to the posterior estimates for the distribution. A large range of models were considered to account for serial correlation and seasonal dependence and the best model selected based on the deviance information criterion include CDD65 and DOW without declustering or incorporating seasonal or time covariates.

Table 1 shows the summary statistics and number of observations for each of the parameters. The summary statistics table shows that the estimate of the mean of the marginal posterior distribution for the parameter xi is .4953 and the 95% equal-tail intervals and HPD credible intervals with the lower endpoints being negative indicate high probability that xi is positive. Hence, the tail of the distribution converges to the Fréchet distribution, a fat-tailed distribution.
Table 1 GEV Posterior Summaries and Intervals

In Table 2, the Monte Carlo standard error of each parameter is significantly low relative to the posterior standard deviations, indicating that the Markov chain has stabilized and so the mean parameter estimates do not vary over time. The table indicates that the parameters have statistically significant t statistics, assuming approximate normality of the parameter estimates. The p-values in the Geweke diagnostics table shows that the mean estimates of the Markov chain are stable over time. The effective sample sizes table reveals all parameters have similar effective sample sizes. All these diagnostics tests indicate convergence of the Markov chain.

Table 2 GEV Monte Carlo Standard Errors, Geweke Diagnostics, and Effective Sample Sizes

As illustration, Figure 3 shows trace, density, and autocorrelation plots of the posterior samples for the shape parameter (xi) and indicates that convergence looks good; there is a stabilized good mixing in the trace plots, a strong peak on density, and a quick drop-off in the ACF plots.

In summer 2010, the number of times the data exceed the 90th quantile is closer to the number of times the data exceed the median than the mean of the 90th quantile estimated from the GEV distribution. Not that surprising, since
the sample distribution for the 90th quantile has a very asymmetrical skewed to the right posterior distribution with almost all values concentrating to the left of the mean. It seems GEV slightly overestimates the risk; indeed 38 positive residuals are not yet long enough to provide more reliable and closer estimation.

Over the simulation period of 61 years, exceeding the 90th quantile becomes closer to the one from simulated data, but once again, highlighting that some noise in a shorter time series could not be accounted accurately by the GEV distribution. This indicates that the GEV distribution with a longer time series is capable of estimating the appropriate quantile to assess the risk in demand forecast uncertainty.

Using the estimated coefficients from the autoregressive regression and daily summer temperatures from 1950 to 2010 allows simulation of demand under a wide range of historical temperature values. By adding the positive residual 90th quantile from GEV, one obtains demand distributions and from there extreme demand quantiles for low probabilities. One can even resample residuals and temperatures to estimate demand density.

For instance, from the probability distribution using summer historical temperatures estimation, Figure 4 shows the simulated and simulated adjusted demand Kernel and normal densities for 2010 using temperatures from 1950 to 2010 for the autoregressive regression. The adjusted demand density includes the residual 90th quantile GEV in the simulation.

![Figure 4 Normal (red) and Kernel (blue) distributions of simulated demand and of simulated adjusted demand (added the 90th GEV quantile). The lines represent the 90th quantile for the corresponding distributions.](image)

The first panel overlays the normal density estimation on the Kernel’s for simulated demand, while the second panel depicts the corresponding densities for simulated adjusted demand. While the 90th quantile estimates derived under the normal distributions do not differ greatly, there is an obvious difference under the kernel distributions in the 90th extreme quantiles based on the simulated and simulated adjusted demand. The 90th quantile by the simulated adjusted distribution covers more of the extreme values in the actual data set than the corresponding simulated distribution quantile. It covers more extreme observations since it is displaced to the left by the simulated GEV 90th quantile. The gap between the distributions implies that they provide different estimates of the quantiles and the choice between the distributions makes a difference leading to underestimation of demand.

The results from estimating the GPD distribution under different threshold values turn out to be somewhat disappointing and clearly the Markov chain could not converge and reach a stationary distribution. A large range of threshold values and techniques to improve calculation and non-convergence were tried, even standardizing the residuals, but they did not eliminate the problem. The residuals exhibit clustering and making the correct choice of the threshold value proves to be difficult; a high threshold leads to too few residuals to estimate the parameters reliably and a low one to an inaccurate approximation of the distribution. Small number of data over the threshold makes difficult or impossible to estimate precisely the distribution.

The best alternative, after preliminary analyses, seems setting thresholds low enough to capture sufficient data for estimation and conducting sensitivity analyses to confirm the stability of the parameter estimates and the fit of the GPD distribution at those thresholds. On this basis, setting the threshold at 100, almost at the 50th percentile, and using the residuals from an ordinary least squares (OLS) regressive model, eliminates the problem and then the random walk Metropolis algorithm proceeds smoothly. The corresponding results for the OLS regression are r-square of 84.95% and coefficients of -0.09619 for DOW and of 0.04015 for CDD65, meaning 9.17% lower loads on weekends and holidays than on weekdays and 12.8% loads increase for every additional 3 cooling days. The posterior mean estimate of the parameter $\alpha$ is 0.3786 and of the 90th quantile 49.71. Additional results are not presented.
CONCLUSION

Underestimation of demand leads to mismatch supply and demand. Analysis of the underestimation of demand or residuals is an effective method for not only assessing the fit of the model to the data and determining whether the model is useful but also by considering the tail of the residual distribution for understanding how rare and/or extreme temperatures affect demand. Such knowledge provides better resource adequacy planning and more importantly, with extreme demand more accurately assessed, improves the reliability of the system.

The MCMC procedure offers a lot of flexibility and functionality for fitting Bayesian predictive models using nonstandard EVT distributions. A Bayesian EVT method shows how extreme errors in demand forecasting can be easily modeled incorporating uncertainties. Specifying them is relatively easier using DATA steps and macro language programming. One declares the parameters and their initial values in the model, specifies prior distributions of them, and specifies either the GEV or GPD conditional distribution for the response variable given the parameters. Then, the procedure computes posterior estimates, provides convergence diagnostics, and outputs the posterior samples to a data set for further analysis.

The EVT node fully incorporates the MCMC procedure into SAS Enterprise Miner, extending its functionality by making the MCMC procedure and EVT distributions available in data mining. The node requires a SAS program containing the MCMC procedure and a corresponding XML file containing the property elements and their valid values that the property panel displays in the Enterprise Miner user interface. In addition, the node presents table and graph output like any other standard SAS Enterprise Miner node.

The results from the analysis suggest that the GEV distribution provides a reasonable fit in estimating probabilities of positive residual extremes, and that Bayesian estimation quantifies the uncertainties in the statistical estimation of such extremes value probabilities. Additionally, the results highlight the interplay between EVT and Bayesian predictive inferences and integrated risk management because working in deciding upon the GEV or GPD distribution and its corresponding quantile is equivalent to estimating value-at-risk; value-at-risk simply represents the quantile of a distribution. Yet, there remain some research issues about the appropriate choice of thresholds.

REFERENCES


CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Miguel Garcia-Cerrutti PhD
California Energy Commission
1516 9th street, MS-22
Sacramento, CA 95814
(916) 654-4761
mcerrutt@energy.state.ca.us

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.