SEAF – A prototype of an expert system for at-site frequency analysis of hydrological annual maxima

Márcio de Oliveira Cândido[1] & Mauro Naghettini[2]
1. Ingeniero Hidrólogo de CPRM (Servício Geológico de Brasil), Belo Horizonte (MG), Brasil
2. UFMG (Universidade Federal de Minas Gerais), Departamento de Engenharia Hidráulica e Recursos Hídricos, Belo Horizonte (MG), Brasil
naghet@netuno.lcc.ufmg.br

Recibido para evaluación: 05 de -Septiembre de 2008 / Aceptación: 01 de Octubre de 2008 / Versión Final: 16 de Octubre de 2008

Abstract

The usually short-sized samples of data recorded at a site and the large uncertainties involved in parameter and quantile estimation are some of the shortcomings encountered in at-site frequency analysis of hydrologic annual maxima. Another drawback is the subjectivity entailed by the arbitrary selection of a candidate probability distribution to model the sample of hydrological annual maxima. Conventional goodness-of-fit tests are not designed to discriminate among candidate models and are not powerful enough to provide the necessary objective backing to such a decision-making process and may possibly lead a novice hydrologist to inadequate choices. In this paper, we describe the authors experience in employing the technology of artificial intelligence and fuzzy-logic theory to build a computer expert system that emulates the reasoning principles used by a human expert to select one or more candidate probability distribution models for at-site hydrologic frequency analysis. The expert system has been applied to 20 relatively large samples of annual maxima of daily rainfall and daily streamflow recorded at gauging stations in the Brazilian southeast. In order to check the system performance, the same samples have been submitted to a panel of actual experts in frequency analysis. The comparison of the results provides evidence that the computer system performs at an expert level and may be utilized to help an inexpert to select one or more appropriate candidate distributions for at-site hydrologic frequency analysis.

Key Words: At-site hydrologic frequency analysis, Expert system, Fuzzy logic, Probability distribution.

Resumen

El tamaño corto de las series de tiempo registrados en una estación, las altas incertidumbres involucradas en la estimación de parámetros y sus cuantiles, son algunas de las desventajas que se encuentran en el análisis de frecuencia in-situ de los máximos anuales hidrológicos. Otra desventaja es la subjetividad presente en la selección arbitraria de una distribución de probabilidad candidata para modelar la muestra de datos hidrológicos máximos anuales. Las pruebas convencionales de ajustes no están diseñadas para seleccionar entre varios modelos y no son lo suficientemente poderosas para proveer los criterios que respalden un proceso de toma de decisiones de tal manera que pueden guiar al hidrólogo novato a escoger opciones no adecuadas. En este artículo describimos la experiencia de los autores en emplear la tecnología de la inteligencia artificial y la teoría de la lógica difusa para construir un sistema experto ()en computador) que simule los principios de razonamiento usados por un experto para escoger uno o varios modelos candidatos de distribución probabilística para estudios in-situ de análisis de frecuencia en hidrología. El sistema experto se aplicó a 20 muestras hidrológicas relativamente largas de máximos anuales de lluvia y caudales diarios registrados en estaciones de monitoreo en el Sureste de Brasil. Para chequear el comportamiento del sistema experto, las mismas muestras fueron entregadas a un panel de expertos en el tema de análisis de frecuencias. La comparación de los resultados mostró que el sistema diseñado se comporta al mismo nivel que el panel de expertos y que se puede utilizar para ayudar a una persona con poca experiencia a escoger una o varias distribuciones de probabilidad candidatas apropiadas en un análisis de frecuencia de datos hidrológicos in-situ.

Palabras clave: análisis local de frecuencias, sistemas expertos, lógica difusa, distribución de probabilidades.

1. Introduction

Analysis and estimation of flood flows are current problems in the domain of water resources engineering. Along history, it is almost natural to perceive the attraction that floodplains and river valleys have exerted on human societies. In fact, river valleys offer very favorable conditions to develop and maintain activities connected to human settlements, such as agriculture, fishing, transportation, and convenient access to local water resources (ASCE, 1996). However, the economic and social benefits resulting from the occupation and use of river valleys and floodplains are frequently offset by the negative effects of flood-induced disasters, such as the loss of lives and the material damage to riverine communities and properties. From a geomorphological standpoint, there is no surprise coming from the fact that rivers occasionally reclaim their own dynamic constructions which are their valleys and plains (Knighton, 1998). However, it is indeed a surprise to acknowledge that human societies occasionally disregard the fact that occupying the floodplains means to coexist with risk.

Flood-risk reduction and mitigation of flood-induced damages can be provided by human actions on the fluvial system, such as the construction of reservoirs and levees, the definition of strategies for regulating land use in flood hazard areas, and the implementation of alert systems and protection measures for riverine structures and properties. Estimation of flood flows is essential not only for planning these interventions, but also for designing and operating the engineering structures to control and manage water, since their structural safety depends much on reliable estimates of flood characteristics. Engineers and hydrologists are often asked to estimate relevant characteristics of flood flows, such as the associated precipitation depths, the peak discharges, the volume and duration of flood hydrographs, the flooded areas, as well as their corresponding critical values for design and/or operation purposes (ASCE, 1996). There are different methods for estimating flood flow characteristics: a few are conceived on a purely deterministic basis, whereas others seek to associate the variable magnitude to an exceedance probability. The latter are current applications of probability theory and mathematical statistics to the field of water resources engineering.

Frequency analysis of hydrologic variables, broadly defined here as the quantification of the expected number of occurrences of an event of a given magnitude, is perhaps the earliest and most frequent application of probability and statistics in the field of water resources engineering (Haan, 2002). In brief, the methods of frequency analysis aim to estimate the probability with which a random variable will be equal or greater than a given value, or quantile, from a sample of observations of the variable in focus (Kidson & Richards, 2005). If these observations are recorded only at a single streamflow (or rainfall) gauging station, the so-called at-site frequency analysis is being carried out. Otherwise, if other observations of the variable, as recorded at distinct gauging stations within a specified region, are jointly employed for statistical inference, then the frequency analysis is said to be regional (ASCE, 1996).

At-site frequency analysis of hydrologic random variables has received much attention from hydrologists over the years. New probability distribution models and improved techniques of statistical inference have been proposed, in the pursuit of more reliable estimates of rare quantiles. However, the relatively short samples of hydrologic maxima seem to impose a limit to the degree of statistical sophistication to be employed in at-site frequency analysis. Along these lines, the regional frequency analysis is certainly an alternative that seeks to balance the limited temporal distribution of hydrologic data, as recorded at a single site, with a more detailed characterization of the variable over space. Potter (1987), Bobée and Rasmussen (1995), and Hosking and Wallis (1997) are among those who presented the many advantages and arguments in favor of regional methods, as compared to at-site frequency analysis. Despite these arguments, it is plausible to expect that a number of decisions on the occurrence of rare hydrologic events will be made still on the basis of a single sample of hydrologic data. This expectation may be justified either by (i) the scarcity or even the absence of hydrologic data within a given region; or (ii) the engineers lack of experience and/or knowledge of using contemporary methods for regional frequency analysis; or (iii) the expedite way with which engineering solutions are occasionally proposed to problems involving hydrologic rare events. It appears then that at-site frequency analysis will stay as a current engineering method for some appreciable time. This impression is, in fact, the chief motivation for the research described in here.

This paper focuses first on the difficulties inherent to at-site frequency analysis of hydrologic annual maxima, particularly those related to picking the suitable probability distribution model (or models) among a number of possible candidates. Next, we describe our experience in employing the technology of artificial intelligence and fuzzy-logic theory to build a computer expert system that emulates the reasoning principles used by a human expert to select one or more candidate probability distribution models for at-site hydrologic frequency analysis. In the sequence, we provide a description of the experiment devised to evaluate the performance of the computer expert system, along with a summary of the results obtained. The last item contains the main conclusions of this paper.

2. Some problems with at-site frequency analysis of hydrologic maxima

According to Kidson and Richards (2005), at-site frequency analysis of hydrologic maxima involves three steps: data choice, model choice and a parameter estimation procedure. Usually, a sample of annual maximum records of the variable is employed for frequency analysis. In principle, the sample records are assumed to be random, independent, and homogeneous occurrences of the hydrologic variable. Frequency analysis of the so-called partial duration series , far from being an uncommon practice among water resources engineers, is however not so often utilized than the analysis of annual maxima series and will not be focused herein.

Once the probability distribution function has been selected, among a number of candidate models that are potentially adequate to fit hydrologic annual maxima, its parameters and variable quantiles should be estimated by conventional methods of statistics, such as the method of moments, the method of the maximum likelihood or the method of L-moments. In the context of at-site frequency analysis of hydrologic annual maxima, the commonly-used probability distribution functions may be grouped in (i) two-parameter models, as exemplified by Gumbel (or Extreme Value Type I), Log-Normal 2P, Gamma, and Exponential distributions; and (ii) models described by more than two parameters, such as the Generalized Extreme Value (GEV), Pearson III, Log-Pearson III, and Wakeby distributions. The reader is referred to Kidson and Richards (2005), ASCE (1996) and Stedinger et al. (1993) for more details on frequency analysis of hydrologic data.

Despite the relatively large set of potentially usable models, there is no consensus, among hydrologists and statisticians, regarding the prescription of a specific parametric form that is universally accepted to model the frequency distribution of hydrologic annual maxima. As opposed to some deductions that can be made in other applications of statistics, such as the ones related to the central limit theorem, the statistical theory of extreme values does not provide the mathematical deductive laws for selecting such a universal model. In fact, in order to deduce the limiting distribution of the maximum value of a large number of variables, the classical theory of extremes departs from the critical assumption that the original variables are independent and identically-distributed (Gumbel, 1958). As pointed out by Perichi and Rodríguez-Iturbe (1985), this assumption is often in direct disagreement with the hydrologic reality, which is composed mainly by complex interactions of seasonal changes and statistical dependence among the original variables. As a result, the task of selecting a suitable probability distribution function for at-site frequency analysis of hydrologic maxima is effectively an *ad-hoc* procedure, with a few clues coming from the perception of how good is the assumed model s fit to the data set of hydrologic annual maxima.

In practice, the short-sized samples of annual maximum records of hydrologic variables bring about subjectivity into the task of selecting a suitable probability distribution model, solely on the basis of goodness-of-fit measures. With samples of sizes typically ranging from 20 to 70, it is impossible to categorically assert that a given probability distribution function, which is considered well fitted to a sample, will represent the true population behavior of the variable. In fact, customary goodness-of-fit tests, such as the χ^2 and Kolmogorov-Smirnov tests, are not powerful enough and, in fact, are incapable of discriminating among different probabilistic models (Kidson and Richards, 2005). Additionally, the methods of statistical inference may yield possibly unreliable estimates of parameters (and of rare quantiles) due mainly to the large uncertainties imposed by short-sized samples.

These difficulties end up in making the judicious selection of a probabilistic model a task for experts, who generally carry it out under the light of a set of heuristic guidelines which are conceived in accord with the knowledge they have acquired along many years of experience and study. Broadly defined, the heuristic approach is an inventive process known to yield incorrect or inexact results at times but likely to yield correct or sufficiently exact results when applied in commonly occurring conditions. The heuristic approach narrows the logical pathways to follow, thus selecting the more convenient ones and reducing a complex task to a smaller group of judgment operations (Chow, 1988). For instance, some expert may suggest using a specified set of candidate probability distributions simply based on the numerical proximity of the population and sample coefficients of skewness. Then, he (or she) may proceed by selecting, from the previously suggested probabilistic models, the one that visually best fits the data records, as plotted on probability paper.

In general, the set of heuristic guidelines can provide some backing for an expert but, when they are used in an indiscriminate way, they may also lead to biased decisions. As an example, admit that the inclusion of a single extraordinary occurrence into the sample, knowingly very rare when compared to the frequencies of other records, has the effect of substantially increase the sample coefficient of skewness. In such a hypothetical case, if the selection of a suitable probabilistic model is based solely on how close are the population and sample coefficients of skewness, then the decision would be to prescribe models that are adequately asymmetric but wrongly chosen because the presence of a flagrant outlier was too determinant to the choice. This example shows that, far from being a decision guided by a systematic set of objective rules, the judicious selection of a probability distribution function to model hydrologic maxima is essentially multi-criteria and heuristic, and susceptible to be made by experts.

Despite the increasing number of existing computer programs that have been developed for at-site frequency analysis of hydrologic records, they are not capable of guiding the choice of the most adequate group of probability distribution functions that fit the sample data. Their results consist mostly of numerical and/or graphical displays of quantile estimates, which may eventually lead an inexperienced hydrologist to face difficult choices regarding which distribution should be adopted, given that two or more distributions can easily pass conventional goodness-of-fit hypothesis tests.

An inadequate choice for the probability distribution function may result in quantile estimates of the characteristic (or decision) variable that may either compromise the economic feasibility of a water resources project or to expose it to an intolerable risk of failure. In this scenario, the judicious selection of a probability distribution for modeling a sample of hydrologic maxima is imperative and cannot be achieved by a single simple algorithm. With effect, such a selection requires the aggregation of pieces of objective and subjective analysis, which may lead to different results as depending upon the reasoning pattern adopted by each expert.

In general, every time a solution to a problem requires the combination of subjective criteria, the utilization of expert systems may be a useful alternative in order to achieve some standardization and agility to the processes of analysis and decision-making. An expert system is referred to as a computer program that was designed and developed to carry out certain tasks that are human in nature or are derived from human knowledge (Jackson, 1998). It should be capable of assisting the decision-making process, on the grounds of knowledge that is substantiated from an information base, thus emulating the reasoning behavior of a human expert.

As for the expert system here described, the information base may be constructed by collecting pieces of specific knowledge and actual decisions taken by qualified professionals with a vast experience in hydrologic frequency analysis. Then, these pieces of information are intertwined to build the logic of decision-making under subjective conditions. The items to follow summarize our experience in conceiving, implementing, and testing the SEAF expert system to assist the judicious selection of a probability model for at-site frequency analysis of hydrologic maxima, having as information base a specified collection of heuristic rules extracted from trends and results of recent research in the area.

3. The SEAF Expert System for At-Site Hydrologic Frequency Analysis

The SEAF expert system (SEAF is the acronym for the Portuguese words *Sistema Especialista de Análise de Freqüência*) aims to implement a set of criteria for selecting one or more suitable probability distribution functions to model hydrologic annual maxima recorded at a single site, with the joint use of elements from fuzzy logic theory and statistical estimation based on L-moments. It is worth it to note that SEAF has not been designed to identify the true probability distribution of the data population. Instead, its objective is to select, among a number of candidate models, the probability distribution functions that appear to be the most appropriate to fit the sample under analysis.

The set of candidate distributions, from which SEAF extracts its possible choices, is formed by the following models: Normal (NOR), 2-parameter Log-Normal (LNR), Extreme Value Type I or Gumbel (GUM), Generalized Extreme Value (GEV), Exponential (EXP), Generalized Pareto (GPA), Pearson Type III (PE3), and Log-Pearson III (LP3). With the exception of the Normal distribution, which is used in SEAF only as a paradigm for making some auxiliary decisions, the set of candidate models encompasses the ones that are most employed in frequency analysis of hydrologic maxima.

SEAF has been developed to work under Microsoft Windows 98 and above. The system architecture is composed of two parts: the first was developed in Borland Delphi Professional and refers to the user graphical interface and to the necessary numerical calculations. The second was developed in the FuzzyCLIPS language and refers to the tasks of storing, interpreting, and analyzing the so-called system knowledge base. FuzzyCLIPS is a fuzzy logic extension of the CLIPS expert system shell from NASA and was developed by the Integrated Reasoning Group of the Institute for Information Technology of the National Research Council of Canada (http://www.iit.nrc.ca/IR_public/fuzzy/fuzzyClips/fuzzyCLIPSIndex2.html).

In brief, SEAF first reads a text file containing the sample of hydrologic maxima and extracts from it the numerical information to be used in the next steps. Then, SEAF analyses the extracted information, under the light of a knowledge-based internal set of heuristic rules, and transforms it into a number of decision statements. In this context, by interfacing with the FuzzyCLIPS shell, SEAF employs the technology of artificial intelligence and fuzzy logic to deal with the inherent uncertainties and to emulate the reasoning principles of a human expert. In order to evaluate the plausibility of each candidate distribution function, the system makes its decision statements on the grounds of a set of rules based on contemporary knowledge. These rules are reasonably similar to those a human expert would utilize to select a probability distribution function to model a sample of hydrologic annual maxima.

As opposed to dichotomical facts, such as true and false or positive and negative , a human expert often faces the situation where uncertainty is inherent and must be represented in the form of a vague concept or statement. This is the rationale of fuzzy logic, where a hypothetical set *A*, as defined in the universe of *X*, is characterized by a membership function ì$_A$ which associates the *x* elements to real numbers within the interval [0,1]. Therefore, the membership function ì$_A$(*X*) is a quantitative expression of how likely a given *x* belongs to *A*. For the sake of clarity, consider the following fuzzy pattern the sample coefficient of skewness *g* tends to zero . Objectively, no one would be able to define a threshold below which the sample coefficient of skewness can be considered as equal to zero. If the sample coefficient of skewness is equal to 0.001, for instance, then one would intuitively be able to state with a high degree of confidence that the skewness tends to zero, which translates into ì$_A$= 1. However, if the sample skewness is equal to 0.3, then ì$_A$ is less than 1 and will decrease according to a specified function, as the sample estimate departs from zero. By interfacing with FuzzyCLIPS, SEAF makes use of bell-shaped and cumulative (or S-type) membership functions to assign confidence levels to each candidate distribution function. The reader is referred to Jackson (1998) for more information on combining expert systems and fuzzy logic theory and to Cândido (2003) for specific details on coupling SEAF and FuzzyCLIPS.

In order to describe the sample data variability and to infer the parameter estimates and distributional shape, SEAF utilizes the

L-moment statistics as introduced by Hosking (1989). The L-moments of order r, denoted by λ_r, can be written as linear combinations of the corresponding probability weighted moments (PWM), these denoted by β_r and defined by the following mathematical expectation

$$\beta_r = \mathrm{E}\left\{X[F(X)]^r\right\} \tag{1}$$

The estimators for the first four L-moments can be calculated in terms of the PWM estimators from

$$\begin{aligned}
\hat{\lambda}_1 &= \hat{\beta}_0 \\
\hat{\lambda}_2 &= 2\hat{\beta}_1 - \hat{\beta}_0 \\
\hat{\lambda}_3 &= 6\hat{\beta}_2 - 6\hat{\beta}_1 + \hat{\beta}_0 \\
\hat{\lambda}_4 &= 20\hat{\beta}_3 - 30\hat{\beta}_2 + 12\hat{\beta}_1 - \hat{\beta}_0
\end{aligned} \tag{2}$$

where $\hat{\beta}_r$ represent the unbiased PWM estimators for a given ordered sample $X_{1:n} \le X_{2:n} \le \ldots \le X_{n:n}$ of size n. Formally,

$$\hat{\beta}_r = \frac{1}{n}\sum_{j=r+1}^{n} \frac{(j-1)(j-2)\ldots(j-r)}{(n-1)(n-2)\ldots(n-r)} X_{j:n},\ r \le n-1 \tag{3}$$

As compared to conventional moments, L-moments generally yield more robust and accurate estimates of distribution parameters and quantiles of a random variable. L-moments and L-moment ratios are also practical and convenient measures of distributional shape. For instance, λ_1 is a measure of location, λ_2 a measure of scale, the ratio $\tau = \lambda_2/\lambda_1$ is analogous to the conventional coefficient of variation, the ratios $\tau_3 = \lambda_3/\lambda_2$ and $\tau_4 = \lambda_4/\lambda_2$ represent measures of skewness and kurtosis, respectively. These quantities can be estimated from a sample by using the L-moment estimators given by equations 2 and 3. The reader is referred to Vogel and Fennessey (1993), Zvi and Azmon (1997), and Hosking and Wallis (1997) for further details on the relative superiority of L-moments, as compared to conventional moments.

The following sequential steps outline the necessary numerical calculations and the set of heuristic rules forming the knowledge base of SEAF reasoning procedures:

1. SEAF checks the sample of annual maxima for serial independence, homogeneity, and presence of outliers by applying the following significance tests: Kendall, Mann-Kendall, and Grubbs and Beck, respectively. The results of these tests are only informative and do not affect the process of selecting and classifying the candidate probability distributions. SEAF informs user on the eventual presence of serial correlation, heterogeneity, and outliers, and asks user whether or not to proceed with the analysis and whether or not to remove low and/or high outliers. It is worth it to emphasize that SEAF does not take any action concerning the outliers, since it is a rather controversial issue to remove them from the analysis unless there is absolute certainty they refer to measurement errors and other data inconsistencies;

2. SEAF calculates sample L-Moments and L-Moment ratios: $l_1\left(\hat{\lambda}_1\right)$, $l_2\left(\hat{\lambda}_2\right)$, t_2 ($\hat{\tau}$), t_3 ($\hat{\tau}_3$), and t_4 ($\hat{\tau}_4$), for the sample of size n, according to formulation described by Hosking and Wallis (1997);

3. SEAF estimates parameters for the following 2-parameter distributions: *dist*=Normal, Log-Normal, Exponential, and Gumbel, as well as their respective Monte Carlo-simulated finite-sample variances $Var^{dist}\left(\tau_3\right)$;

4. SEAF estimates parameters for the following 3-parameter distributions: *dist*=Pearson III, Log-Pearson III, Generalized Extreme Value (GEV), and Generalized Pareto (GPA), as well as their respective Monte Carlo-simulated finite-sample variances $Var^{dist}\left(\tau_4\right)$;

5. By assuming t_3 as Normally distributed according to N[τ_3^{dist}, $Var^{dist}\left(\tau_3\right)$], SEAF defines the confidence interval $\left[\tau_3^{0.025}, \tau_3^{0.975}\right]$ for each 2p distribution;

6. On the basis of the interval $\left[\tau_3^{0.025}, \tau_3^{0.975}\right]$ and on the sample L-skewness t_3, SEAF selects all plausible 2p distributions and associates to each one a preliminary confidence level, which is calculated according to a membership function of the bell-shaped type with the 0.5-level threshold;

7. By assuming t_4 as Normally distributed according to N[τ_4^{dist}, $Var^{dist}\left(\tau_4\right)$], SEAF defines the confidence interval $\left[\tau_4^{0.025}, \tau_4^{0.975}\right]$ for each 3p distribution;

8. On the basis of the interval $\left[\tau_4^{0.025}, \tau_4^{0.975}\right]$ and on the sample L-kurtosis t_4, SEAF selects all plausible 3p distributions in a manner similar to that described in item 6;

9. SEAF applies the probability correlation coefficient goodness-of-fit test (also known as Filliben s test), as described in Stedinger et al. (1993), for the 2-parameter and 3-parameter plausible distributions and defines the confidence interval $[\rho_{0.05}^{dist}, 1]$, on the basis of Monte Carlo-simulated finite-sample variances of the correlation coefficient ñ between estimated and empirical quantiles;

10. On the basis of the interval $[\rho_{0.05}^{dist}, 1]$ and on the test statistic r^{dist}, SEAF selects all plausible distributions and assigns to each one a secondary confidence level, which is calculated according to a membership function of the cumulative type (or S type) with the 0.5-level threshold;

11. If skewness $g_x<0$ (or $g_{\ln x}<0$), SEAF will remove Pearson III (or Log-Pearson III) from the analysis;

12. SEAF checks the sign of GPA (and GEV) shape parameter k; if k>0 SEAF will remove GPA (and GEV) distribution from the analysis;

13. SEAF checks whether or not the presence of low outliers significantly modify parameter and upper-tail estimation. Remove from the analysis those distributions for which the sample minimum is inferior to an arbitrarily chosen level of 90% of their respective location parameter estimates;

14. SEAF applies the following parsimony criterion: $CI_{adjusted} = 1-(1-CI_{average})(n-1)/(n-p)$, where *CI* denotes the confidence level and *p* the number of estimated parameters, in order to discriminate among distributions of the same family, such as GPA versus EXP, GEV versus GUM, and LP3 versus LNR; and

15. SEAF ranks distributions according to their average confidence levels and provide general decision statements.

A .zip file containing the SEAF installation package is downloadable from the URL http://www.ehr.ufmg.br/downloads.php. The package also includes the file example.dat which is a text file, with a total of 52 lines as corresponding to 52 records of annual maxima, to serve as an application example of SEAF. Once SEAF is installed and running, it will first open a welcome screen as illustrated in Figure 1. The second step in running SEAF is to open a new project, by providing its name, a brief description of it, as well as the name and location of the .dat file (e.g. example.dat) containing the maxima records. SEAF will then create a project file (e.g. example.prj) which may be directly opened in later runs of the computer program.

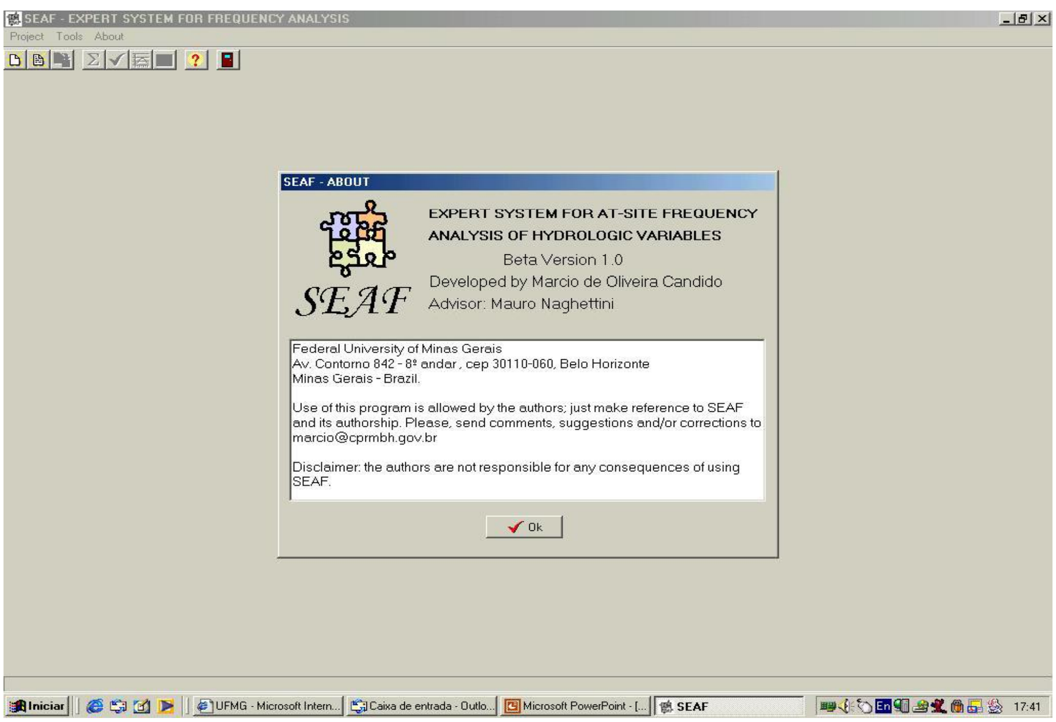

Figure 1 SEAF welcome screen.

After reading the data file, SEAF shows a histogram of the variable and provides a summary of the main descriptive statistics, in both arithmetic and log spaces, as depicted in Figure 2.

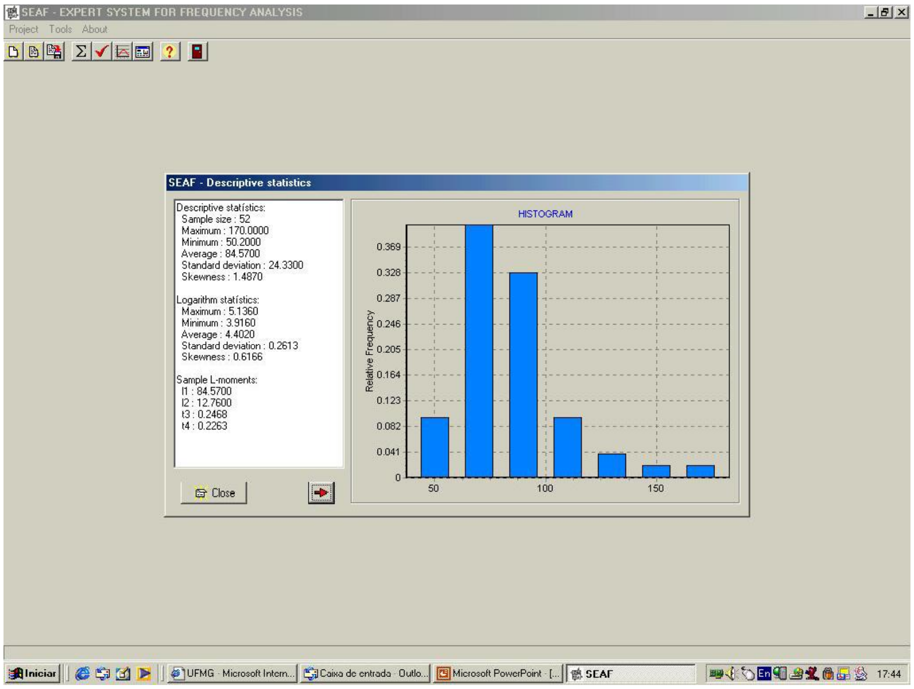

Figure 2 Histogram and descriptive statistics for data in example.dat .

By clicking on the arrow key shown in SEAF screen, the program will proceed by testing the hipothesis of data independence and homogeneity, and checking for the presence of low and high outliers in the sample. This is shown in Figure 3.

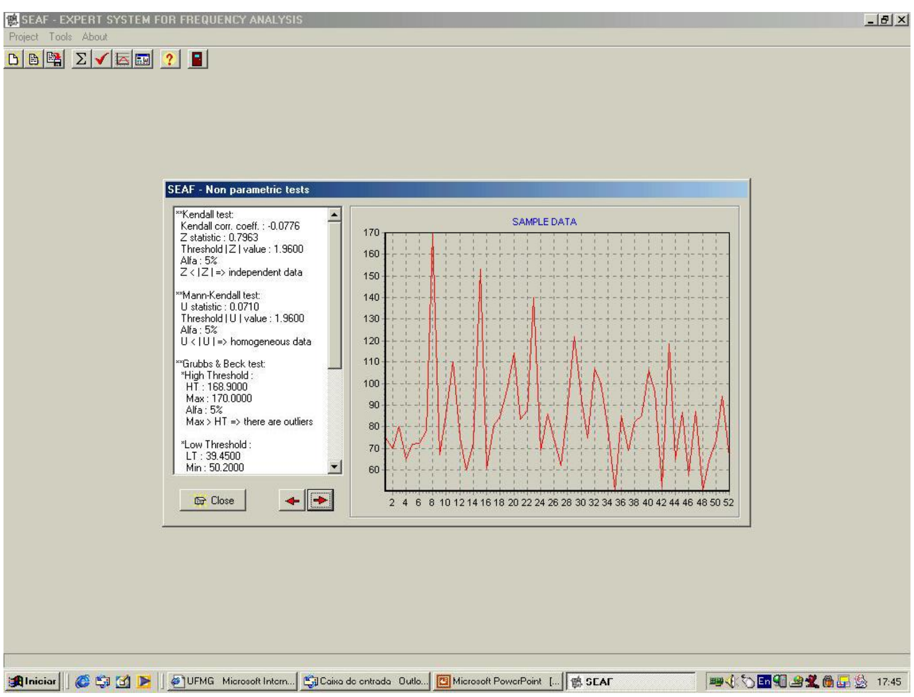

Figure 3 SEAF hypothesis testing.

By clicking the right-arrow key shown in its screen, SEAF will display, on the left side, the parameter estimates for all candidate probability distributions, as calculated by the method of L-moments which was outlined in the previous section. In addition, as depicted in Figure 4, SEAF will show a probability plot diagram, with the log-transformed return period (in years) on the abscissa axis and the variate quantiles on the ordinate axis; note that, in the context of maxima and for a given quantile, the return period corresponds to the inverse of the exceedance probability. The probability plot, as illustrated in Figure 4, shows both the empirical and the theoretical quantiles. The former are plotted by associating the sample records, as classified in descending order, to their respective empirical return periods which are calculated by the ratio $(N+1)/m$, where *N* is the sample size and *m* denotes the order of classification. The theoretical quantiles correspond to the inverse of the cumulative distribution function in focus, among the eight candidate models. For instance, in Figure 4, the GEV theoretical quantiles are shown; with the cursor on the chart and by clicking the mouse right button, SEAF will open a list of all available candidate models to choose from (see Figure 4).

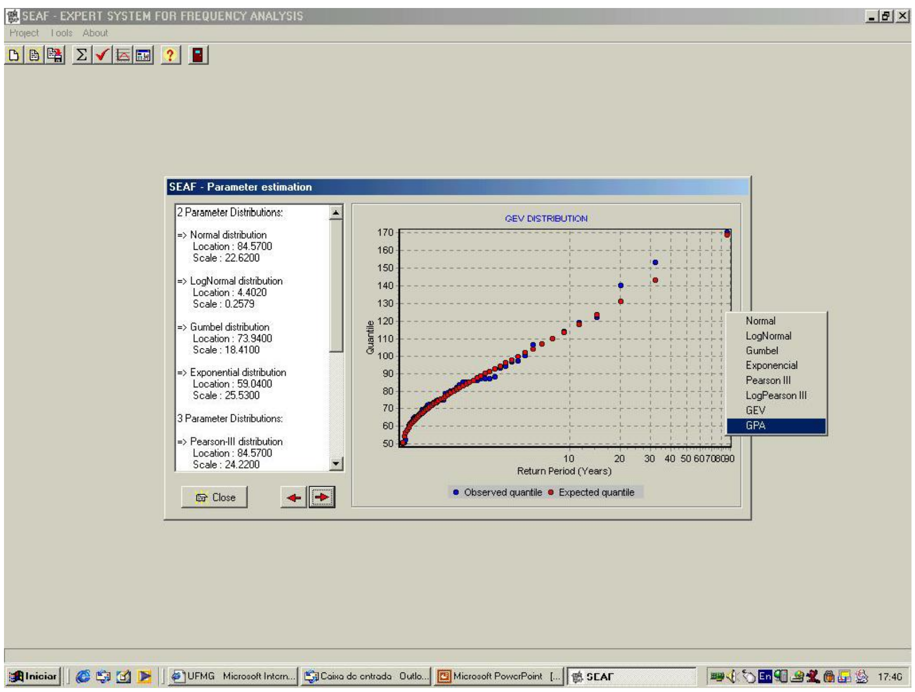

Figure 4 SEAF screen with parameter and quantile estimates.

Following a mouse click on the right arrow, the next screen will display all numerical calculations done by SEAF, prompting the user with the option of saving the summarized information to a user-specified text file. Another mouse click on the right arrow will cause SEAF to automatically call the FuzzyCLIPS shell and start the built-in reasoning procedures (see Figures 5 and 6). These are:

- *Selecting distribution on the basis of L-moments variance test*: this procedure corresponds to steps 3 to 8, as described in the previous section;
- *Selecting distribution on the basis of Filliben s test*: this procedure corresponds to steps 9 and 10, as described in the previous section;
- *Looking for reasons to reject any of previously selected probability distributions*: this procedure corresponds to steps 11 to 13, as described in the previous section;
- *Verifying parsimony of selected distributions*: this procedure corresponds to step 14, as described in the previous section; and
- *Recommended distributions*: this corresponds to step 15, as previously described.

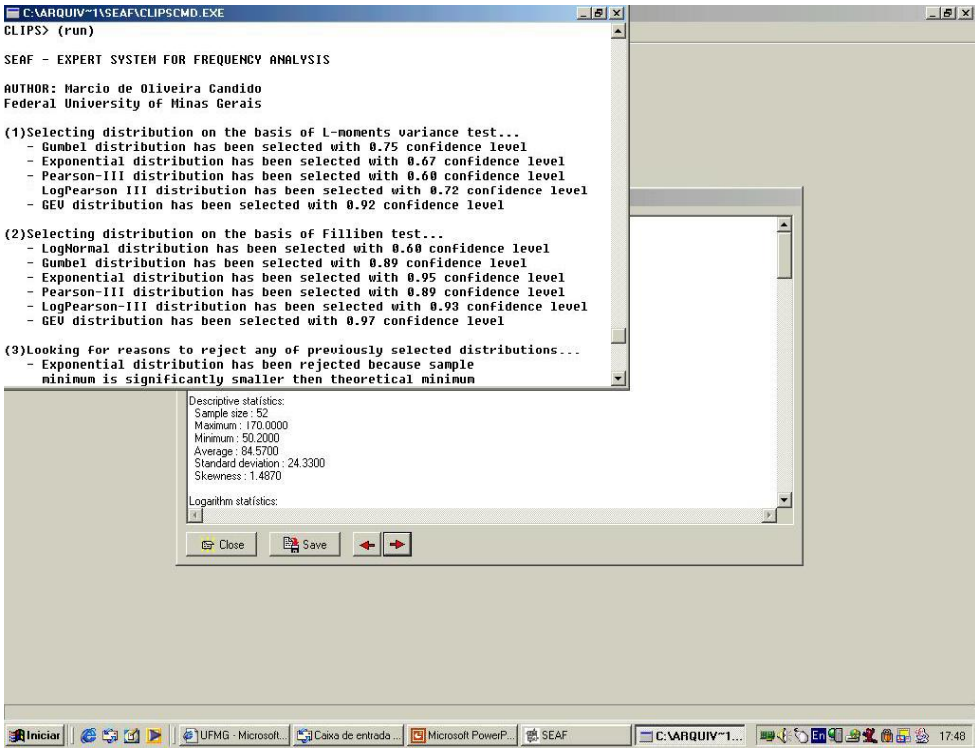

Figure 5 SEAF calls FuzzyCLIPS shell.

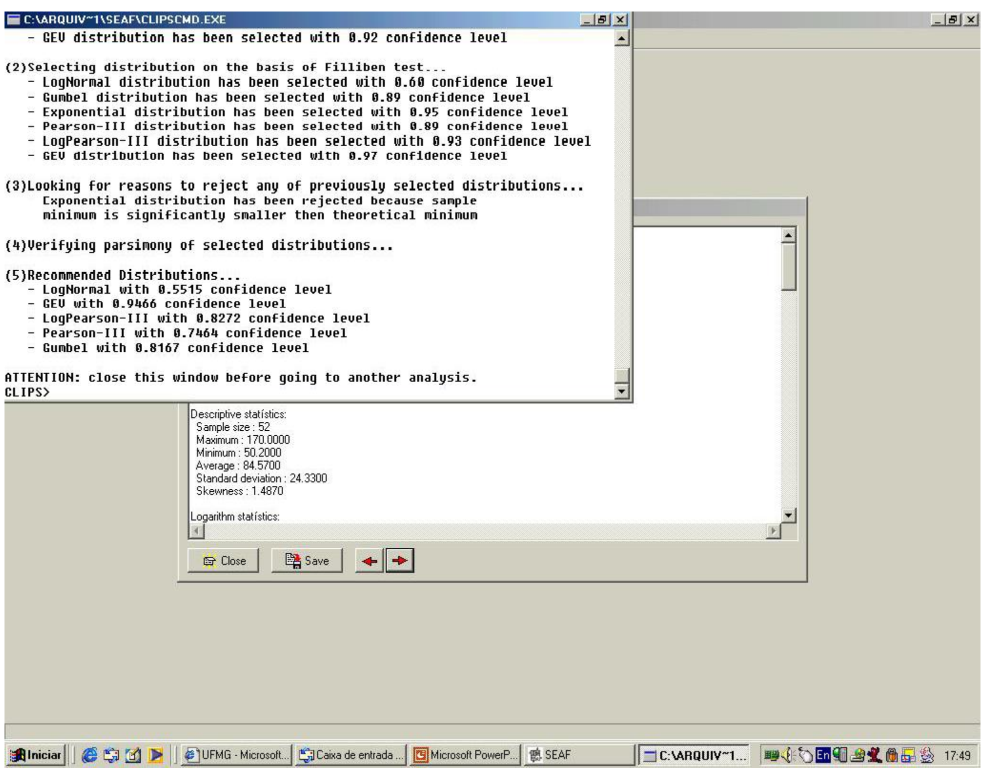

Figure 6 SEAF displays recommended distributions with their respective confidence levels.

4. Analysis of SEAF performance

In order to evaluate the performance of SEAF, 10 samples of annual maximum daily rainfall depths together with 10 samples of annual maximum daily stramflows have been submitted to analysis by the system. These samples correspond to data recorded at 20 gauging stations located in southeastern Brazil, within and near the borders of the Brazilian state of Minas Gerais, which are illustrated in Figure 7. The main criterion to select these gauging stations was to have available at least 35 years of continuous records of consistent and good quality daily observations. In fact, too-short samples could yield larger estimation errors on parameter and quantiles, and would eventually induce SEAF to make a wrong choice. In the case of streamflow data, an additional criterion was to consider only the gauging stations with insignificant diverted and/or regulated upstream flows. Also, it is worth it to note that the streamflow gauging stations are not located in nested river basins to avoid eventual inconsistencies among quantiles estimated by different probability distributions along the same river.

Figure 7 Borders of Brazil and the state of Minas Gerais.

All 20 samples have passed the hipothesis tests of data independence and homogeneity. As for the presence of outliers, the test of Grubbs and Beck has identified atypical data points in 8 of the 20 samples. After checking for eventual data inconsistencies and visually inspecting the corresponding probability plots for these 8 samples, the action taken has been that of maintaining the outliers in the samples because no objective reasons were found to discard them. Then, all samples have been submitted to SEAF built-in statistical tests, as described in steps 6, 8, and 10 of the previous section. The distributions that have passed these tests, for each one of the 20 samples, are marked with the √ symbol in Table 1.

Table 1. SEAF approved distributions for each of the 20 samples

Sample (station code)	Normal (NOR)	Log-Normal (LNR)	Gumbel (GUM)	Exponential (EXP)	Pearson III (PE3)	Log-Pearson III (LP3)	Gen. Extreme (GEV)	Gen. Pareto (GPA)
01544012		√	√	√	√	√	√	
01645000		√	√	√	√	√	√	√
01943000	√	√	√		√	√	√	√
01944004		√	√		√	√	√	√
01944007				√	√	√	√	√
02044012		√	√	√	√	√	√	√
02045005		√	√	√	√	√	√	√
02244038			√	√	√	√	√	√
01943009		√		√	√	√	√	
02243004		√	√	√	√	√	√	√
40025000				√	√	√	√	√
40050000		√	√	√	√	√	√	√
40100000		√		√			√	
40680000						√	√	
41250000		√	√	√	√	√	√	√
40800001		√	√	√	√	√	√	√
56028000		√		√	√	√	√	√
56075000				√	√	√	√	√
56415000		√	√	√	√	√	√	√
56500000		√		√			√	
Number of approvals	1	15	12	17	17	18	20	14

Table 1 shows that, on the basis of SEAF statistical tests only, 5 of the 8 candidate models are capable of fitting three quarters of the samples. In addition to that and excluding the Normal distribution, all candidate models are capable of fitting at least 12 of the 20 samples and no less than three different models can fit 90% of the samples. The GEV model is capable of fitting all samples. The 3-parameter probability distributions have been approved more often than the 2-parameter models because their third parameter actually confers more flexibility to fitting sample data. However, it is important to note that adding a third parameter means adding more estimation uncertainty to a fitted model. All these are arguments in favor of using other criteria, in addition to SEAF goodness-of-fit tests, to select and classify probability distribution models for hydrologic annual maxima.

Table 2 lists, for each sample, the distributions that have been rejected by SEAF either by applying (i) the criterion of occasional upper bounds and/or lower sample bounds, as corresponding to steps 11, 12 , and 13 (see description in the preceding section) or (ii) the criterion of parsimony, or step 14, as previously described. The first criterion was the single cause of 84% of the rejection cases. Note also that the distributions EXP and GPA altogether correspond to 57.8% of all rejection cases.

Table 2. SEAF rejected distributions

Sample (station code)	Criterion				
	Upper/Lower Bounds				Parsimony
01544012	EXP				
01645000	EXP	GPA	GEV		LP3
01943000	GPA	GEV			LP3
01944004	GEV	LP3			
01944007					GPA
02044012	GPA	PE3	EXP		
02045005	GPA	EXP			GEV
02244038					GPA
01943009	PE3	EXP			
02243004	GPA	EXP			
40025000	GPA	EXP	PE3		
40050000	EXP	LP3	PE3		GEV
40100000	EXP				
40680000	GPA				
41250000	GPA	LP3	PE3	EXP	
40800001	EXP	GPA	GEV	LP3	
56028000	PE3				
56075000	GPA				
56415000	GPA	EXP			LP3
56500000	EXP				

The distributions that have not been previously rejected are then ranked with respect to their corresponding average confidence levels, as described in step 15 of the preceding section. Table 3 presents, for each sample, the system first to fifth choices, as ranked according to the decreasing confidence levels that have been assigned by SEAF. Note in Table 3 that, at the end of SEAF reasoning procedures, some samples can be fitted only by no more than two or three candidate models.

Table 3. SEAF first to fifth choices

Sample (station code)	Selected Distribution				
	1st	2nd	3rd	4th	5th
01544012	GEV	LP3	GUM	PE3	LNR
01645000	GUM	PE3	LNR		
01943000	LNR	GUM	PE3	NOR	
01944004	GUM	LNR	PE3		
01944007	GEV	EXP	LP3	PE3	
02044012	GEV	LP3	GUM	LNR	
02045005	PE3	GUM	LP3	LNR	
02244038	EXP	PE3	GEV	LP3	GUM
01943009	GEV	LP3	LNR		
02243004	GEV	LP3	PE3	GUM	LNR
40025000	GEV	LP3			
40050000	GUM	LNR			
40100000	GEV	LNR			
40680000	GEV	LP3			
41250000	GEV	LNR	GUM		
40800001	GUM	LNR	PE3		
56028000	GPA	EXP	GEV	LP3	LNR
56075000	GEV	LP3	EXP	PE3	
56415000	GEV	LNR	GUM	PE3	
56500000	LNR	GEV			

In order to evaluate the SEAF performance, two experiments have been conducted. The first consisted of submitting the same 20 samples that have been examined by SEAF to frequency analysis by a panel of human experts; the results were then compared to SEAF s choices. The second experiment consisted of submitting synthetic samples of typical sizes to frequency analysis by SEAF. These samples have been drawn from hypothetically-constructed populations of a random variable with an assumed known probability density function. The main objective of the second experiment was to test the SEAF robustness to sampling errors.

The panel of human experts was formed by four members of recognized knowledge and experience in frequency analysis of hydrologic data: two statisticians, a professional hydrologist and a university professor of water resources engineering. As previously consented, the panel members have agreed not to be nominally identified at any phase of this experiment. We have sent to each panel member an e-mail containing the 20 samples of hydrologic data, along with their respective probability plots (observed quantiles versus empirical probabilities) and the main sample statistics, as calculated by conventional moments and by L-moments. A brief descrition of the experiment and of the set of heuristic rules used by SEAF has been attached to each e-mail.

The panel member was asked to analyze each one of the 20 samples and to select the distribution candidate that best fitted the data, according to his/her own set of heuristics rules, along with a brief justification of his/her decision. The results are summarized in Table 4. These show that, for any of the 20 samples, no consensual selection has been exercised by the experts. In fact, this was an expected outcome since the lack of consensus, as mentioned earlier in this paper, is inherent to hydrologic frequency analysis.

Expert 1 has performed his/her analysis with the help of Genstat, a commercial software described in <http://www.vsni.co.uk/software/genstat/>, in which the model s lack-of-fit is measured by the deviance statistic D for that model; this statistic is proportional to $-2\ln L$, where L denotes the likelihood function of the model fitted to the sample data. In general, the larger the deviance the poorer the model s fit; this property can be used to compare different models. In his/her words, expert 1 has combined the objective criterion of Genstat s deviance statistic with a visual appraisal of the probability plot of the sample data. He/she pointed out that the universe of Genstat s candidate models was slightly different than that of SEAF. Because of that, he/she has considered only the distributions that were common to both universes.

Expert 2 has expressed his/her opinion that it is unfeasible to identify a single model as the true probability distribution of the population, on the basis of sample sizes of the order of 50. As a result, he/she advocates using regional information to select the probability distribution that is appropriate to a particular data sample. In what concerns the 20 samples under analysis, he/she assumed not to exist climatic or geomorphic differences among the gauging stations that could cause significant heterogeneity from the viewpoint of the frequency of annual maxima. As a result, his/her results have been obtained by means of separate regional analysis of streamflow and rainfall data, both under the premise that the respective sites were all located within a single homogeneous region. The method of regional L-moments, as described by Hosking and Wallis (1997), has been used to regionalize the frequency distribution of data.

Table 4. First-choice distribution models as selected by the panel of human experts.

Sample	Expert #				Number of coincidental choices among experts'	
	1	2	3	4	≥ 2	≥ 3
01544012	LNR	GEV	GEV	EXP	GEV	
01645000	LNR	GEV	GUM	GUM	GUM	
01943000	LNR	GEV	GUM	LNR	LNR	
01944004	LNR	GEV	GUM	EXP		
01944007	LNR	GEV	GEV	EXP	GEV	
02044012	LP3	GEV	GEV	EXP	GEV	
02045005	LP3	GEV	GUM	LP3	LP3	
02244038	LP3	GEV	GEV	LP3		
01943009	LNR	GEV	GEV	GEV	GEV	GEV
02243004	LNR	GEV	GEV	EXP	GEV	
40025000	LP3	GEV	GPA	GEV	GEV	
40050000	LNR	GEV	GEV	EXP	GEV	
40100000	LNR	GEV	GEV	LP3	GEV	
40680000	LP3	GEV	GEV	LP3		
41250000	LP3	GEV	GEV	GEV	GEV	GEV
40800001	LNR	GEV	GUM	GUM	GUM	
56028000	LP3	GEV	GPA	LP3	LP3	
56075000	LNR	GEV	GEV	GEV	GEV	GEV
56415000	LNR	GEV	GEV	LP3	GEV	
56500000	LNR	GEV	GEV	GEV	GEV	GEV

Expert 3 has employed a collection of subjective criteria that is rather similar to the one embedded in SEAF. His/her results were justified mainly on the basis of the visual appraisal of probability plots and on the graphical comparison of sample and theoretical L-moment ratios, as plotted on L-moment diagrams (see Hosking and Wallis (1997) and Vogel and Fennessey (1993) for details on L-moment diagrams). In addition, he/she used Filliben and Kolmogorov-Smirnov significance tests to check how good the model fits the sample data. Expert 4 has conducted his/her analysis in a way which is similar to that used by expert 3. However, he/she paid more attention to the upper-tails of the fitted probability distributions.

The results from the experts have been assembled into two groups for evaluating SEAF s performance. The first group corresponds to the samples with more than one coincidental choices among the models selected by experts, as shown in Table 4 under the column headed by ≥ 2". By comparing the results of the first group with those obtained by SEAF, one can verify that 76.5% of experts choices are coincidental with the system first choices. If SEAF s first two choices are considered, this percentage increases to 82.4%. If all distributions approved (or selected) by SEAF are considered, the percentage goes up to 94.1%. The second group corresponds to the samples with three coincidental choices among the models selected by experts, as shown in Table 4 under the column headed by ≥ 3 . By performing the same comparisons as in group 1, the percentages change to 75%, 100%, and 100%, respectively. These are summarized in Table 5.

Table 5. Comparisons among SEAF s and experts choices

SEAF Choice	Percentage of coincidental choices with Expert #				Percentage of coincidental choices	
	1	2	3	4	≥ 2	≥ 3
1st	10.0	55.0	70.0	35.0	76.5	75.0
1st or 2nd	50.0	60.0	85.0	50.0	82.4	100.0
All selected	85.0	70.0	90.0	65.0	94.1	100.0

In the second experiment, the Monte Carlo method has been used to generate synthetic samples with the same numerical values of L-moments and L-moment ratios of the observed samples. For each sample of supposedly known L-moments and L-moment ratios, eight synthetic series of typical size have been generated as corresponding to every candidate distribution under analysis. In total, 160 synthetic series have been drawn from hypothetically-constructed populations of a random variable with an assumed known probability density function. Then, they have been submitted to analysis by the system with the purpose of evaluating both the number of correct decisions and the robustness of SEAF s selections. The results of this experiment are summarized in Table 6.

Table 6. Evaluation of SEAF performance from synthetic samples

Synthetic Samples	Percentage of correct decisions by SEAF				
	Distribution Model			Distribution Family	
	1st Choice	2nd Choice	All Selected Distributions	1st Choice	2nd Choice
NOR	50.0	70.0	95.0	90.0	100.0
LNR	25.0	65.0	95.0	30.0	75.0
GUM	40.0	55.0	95.0	70.0	80.0
EXP	20.0	30.0	95.0	35.0	60.0
PE3	15.0	45.0	90.0	20.0	50.0
LP3	6.3	31.3	81.3	25.0	50.0
GEV	56.3	75.0	81.3	62.5	93.8
GPA	0.0	57.1	57.1	28.6	85.7
TOTAL	**28.8**	**53.2**	**89.2**	**46.8**	**73.4**

By examining SEAF s results for synthetic samples, one can verify that in 53.2% of the cases the system identifies, in its first or second choices, the probability distribution from which the sample has been drawn. If all distributions selected by SEAF are considered, this percentage increases to 89.2% of the cases. It is worth it to note that in 73.4% of the cases, the probability distribution that generates the sample and/or its close relative from the same family of distributions are classified by SEAF as a first or a second choice.

It is also worth it to note that most of the samples that have been drawn from the GPA distribution exhibit positive shape parameters or, in other words, are drawn from upper-bounded distributions. This explains the very poor performance of the system in what regards this distribution, particularly in its first choice, since SEAF has been built to reject upper-bounded probability distributions. However, with the exception of the GPA distribution, in almost 90% of the cases, the population model from which the samples are drawn is among the distributions that have been selected by SEAF.

Regarding SEAF s first choices only, one can verify that its percentage of correct decisions is below 50%. An exception is made for the samples drawn from NOR and GEV distributions which accounted respectively for 50% and 56.3% of correct decisions. Table 7 summarizes the first-choice analysis for all candidate models. The results in Table7 do not show a clear evidence of a possible SEAF tendency or preference for any particular model.

Table 7. SEAF s first choices for the synthetic samples

Distribution from which the samples are drawn	SEAF First Choice							
	NOR	LNR	GUM	EXP	PE3	LP3	GEV	GPA
NOR	**50.0**	5.0	5.0	0.0	40.0	0.0	0.0	0.0
LNR	5.0	**25.0**	10.0	0.0	40.0	5.0	15.0	0.0
GUM	0.0	25.0	**40.0**	0.0	5.0	0.0	30.0	0.0
EXP	0.0	0.0	5.0	**20.0**	35.0	5.0	20.0	15.0
PE3	5.0	5.0	30.0	15.0	**15.0**	0.0	20.0	10.0
LP3	0.0	18.8	12.5	6.3	12.5	**6.3**	37.5	6.3
GEV	0.0	12.5	6.3	0.0	12.5	12.5	**56.3**	0.0
GPA	0.0	0.0	14.3	28.6	0.0	14.3	42.9	**0.0**
TOTAL	8.6	12.2	15.8	7.2	22.3	4.3	25.2	4.3

5. Conclusions

In this paper, we described the structure of a prototype expert system, namely SEAF, built to select one or more appropriate candidate distributions for at-site frequency analysis of annual maxima. We also described two experiments to verify whether the SEAF prototype system performs at an expert level. According to the results of these experiments, SEAF performs satisfactorily. Despite the lack of consensus among experts and the intrinsic complexities of hydrologic frequency analysis, we conclude that, when SEAF is utilized with samples of sizes not too much smaller than 30, the system is able to give helpful directions for novice hydrologists on selecting one or more appropriate probability distributions among a number of candidate models. However, it is worth it to remind that the heuristic rules, as built in SEAF, are just approximations of a certain reasoning pattern that reflects some of our particular convictions and preferences. Given the subjectivity that is invariably present in these rules, other convictions and preferences can equally be built into other similar systems and lead to different decisions from those made by SEAF.

In spite of the helpful directions SEAF is able to give for the novice hydrologists and also for the practitioner engineers, caution is always necessary in the decision making process derived from at-site frequency analysis of hydrologic data. First, it is worth it to remind that the true distributions describing the probabilistic behavior of the phenomenon in question are not known and even if they were, their analytical forms would probably be too complex, or would have too many parameters, to be of practical use. In this context, the steps involved in conventional hydrologic frequency analysis, particularly the selection of a reasonable and simple probabilistic model, is in fact a practical issue to ultimately obtain risk estimates of satisfactory accuracy for the problem in question. However, the short samples of hydrologic annual maxima usually do not offer solid grounds for extrapolating a fitted model to return periods much larger than two or three times the number of years of record. If estimates for larger return periods are required, other methods, such as regional frequency analysis, are required.

Acknowledgements

The authors wish to acknowledge the partial support of FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais) to this research through grant CRA APQ4683-5.04/07 (PPM). The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

References

ASCE, 1996. Hydrology Handbook (2nd Edition) - Chapter 8 Floods. ASCE Manuals and Reports on Engineering Practice No. 28. American Society of Civil Engineers. New York, USA.

Bobée, B. and P. Rasmussen, 1995. Recent advances in flood frequency analysis. U.S. National Report to IUGG, 1991-1994, Rev. Geophysics, V. 33 Suppl. (<http://earth.agu.org/revgeophys/bobee01/bobee01.htm>).

Davis, E. G. and M. C. Naghettini, 2001. Estudo de Chuvas Intensas. Projeto Rio de Janeiro. CPRM. Belo Horizonte, Brazil.

Cândido, M. C., 2003. SEAF Um Protótipo de um Sistema Especialista para Análise de Freqüência Local de Eventos Hidrológicos Máximos Anuais. MSc dissertation. Universidade Federal de Minas Gerais, Belo Horizonte, Brazil.

Chow, K. C. A., 1988. Towards a knowledge-based expert system for flood frequency analysis. PhD thesis. Queen s University. Kingston, Canada.

Gumbel, E. J., 1958. Statistics of Extremes. Columbia University Press. New York, USA.

Haan, C. T., 2002. Statistical methods in Hydrology (2nd Edition). Iowa State Press. Ames, USA.

Hosking, J. R. M., 1989. Some theoretical results concerning L-moments. IBM Research Report - RC 14492. IBM Research Division. Yorktown Heights, USA.

Hosking, J. R. M. and J. R Wallis, 1997. Regional Frequency Analysis - An Approach Based on L-Moments. Cambridge University Press. Cambridge, United Kingdom.

Jackson, P., 1998. Introduction to Expert Systems. Addison Wesley Longman. Essex, United Kingdom.

Kidson R. and K. S. Richards, 2005. Flood frequency analysis: assumptions and alternatives. Progress in Physical Geography, V. 29, N. 3, pp. 392-410.

Knighton, D., 1998. Fluvial Forms and Processes - A New Perspective. Arnold, London, United Kingdom.

Muhara, G., 2001. Selection of flood frequency model in Tanzania using L-moments and the region of influence approach. 2nd WARFSA/WaterNet Symposium: Integrated Water Resources Management - Theory, Practice, Cases. Cape Town, South Africa.

Perichi, L. R. and I. Rodríguez-Iturbe, 1985. On the Statistical Analysis of Floods. A Celebration of Statistics, Atkinson A. C. & S. E. Fienberg (eds), pp. 511-541. Springer-Verlag. New York, USA.

Potter, K. W., 1987. Research on flood frequency analysis: 1983-1986. Rev. Geophys. , V. 26, N. 3, pp. 113-118.

Stedinger, J. R., R. M. Vogel and E. Foufoula-Georgiou, 1993. Frequency Analysis of Extreme Events. Chapter 18 in Handbook of Hydrology, Maidment D. R. (ed). McGraw-Hill, New York, USA.

Vogel, R. M. and N. M. Fennessey, 1993. L-moment diagrams should replace product moment diagrams. Water Resources Research, V. 29, N. 6, pp. 1745-1752.

Zvi, A. B. e B. Azmon, 1997. Joint use of L-moment diagram and goodness-of-fit test: a case study of diverse series. Journal of Hydrology, No. 198, pp. 245-259.