Determining Extreme Flows Using Entropy Theory

Drew Allan Loney PhD PE, USACE ERDC, Vicksburg, MS drew.a.loney@usace.army.mil
Aaron Byrd PhD PE, USACE ERDC, Vicksburg, MS aaron.r.byrd@usace.army.mil
Joseph Gutenson PhD, USACE ERDC, Vicksburg, MS joseph.l.gutenson@usace.army.mil
Edward Race PE, USACE ERDC, Vicksburg, MS edward.e.race@usace.army.mil

Abstract

Accurate estimation of extreme weather and hydrologic events with greater than 100 or 500-year recurrence intervals is essential in planning and maintaining civil infrastructure. Recent hydrologic events, such as Hurricane Harvey, have led researchers and the public to question the accuracy of existing recurrence interval estimates. Typical gage sites have a record spanning about 40 years and many have a smaller reference frame. These gage records can give the hydrologist a reasonable approximation of flows with short recurrence intervals. However, accurate extrapolation of this data to the less frequently observed, long recurrence interval events in the distribution tail is a challenge and is often subject to the assumptions built into the statistical models.

Shannon Entropy theory applied to the watershed extremes analysis has significant promise for improving frequency characterization over existing methods. The entropy approach provides an independent estimate of event frequency from current statistical methods that rely solely on the local or regionalized historical record. This approach is also insensitive to the assumptions used to construct the tail region of existing statistical approaches from which extreme events are taken. Moreover, as the concepts of available states and maximum entropy underpin the given method, more information can be extracted about the potential parameter states available to the watershed.

Extension of the method to low frequency events requires subsequent research to demonstrate that the large magnitude states predicted by the method are in fact realizable. Further research is required to compare the entropy derived recurrence estimates to that from traditional historical record statistical analysis, particularly in the low frequency region that is poorly characterized by the latter. In addition, while the developed method is expected to generalize broadly across other watershed parameters beyond discharge, the validity of this method must also be demonstrated under these cases.

Due to the large impacts of recent extreme weather events, there is a significant demand to improve the prediction of extreme event frequency to facilitate planning and construction of mitigating infrastructure. Building on Shannon Entropy theory, the technique described in this research demonstrates that it is possible to estimate the potential maximum states available to a watershed from only the historical minimum, mean, and maximum annual time series through the derivation of an entropy parameter. If the time series are weakly correlated, the marginal probability distribution of each parameter can be used to combinatorially produce a likelihood surface for the maximum event magnitude. Collapsing the likelihood surface results in a return period curve which estimates the likelihood of an annual maximum event magnitude. The researchers hypothesize that the described procedure yields an improved estimate of extreme event magnitudes given its use of watershed states and reduced reliance on long-tail behavior to derive the likelihood of low frequency events.

Introduction

Accurate estimation of extreme weather and hydrologic events with greater than 100 or 500-year recurrence intervals is essential in planning and maintaining civil infrastructure (Ross & Lott, 2003) (Kvocka, Falconer, & Bray, 2016). The Geological Survey (USGS) Gages II dataset contains gage locations in the United States for which there are either 20+ complete (not necessarily continuous) years of data or as of water year 2009 are active (USGS, 2011). The average Gages II gage site has an average record length of about 40 years as of 2009. Figure 1 describes the distribution of the number of complete years for each gage in the Gages II dataset. These gage records can give the hydrologist a reasonable approximation of flows with short recurrence intervals. However, accurate extrapolation of this data to the less frequently observed, long recurrence interval events in the distribution tail is a challenge and is often subject to the assumptions built into the statistical models.

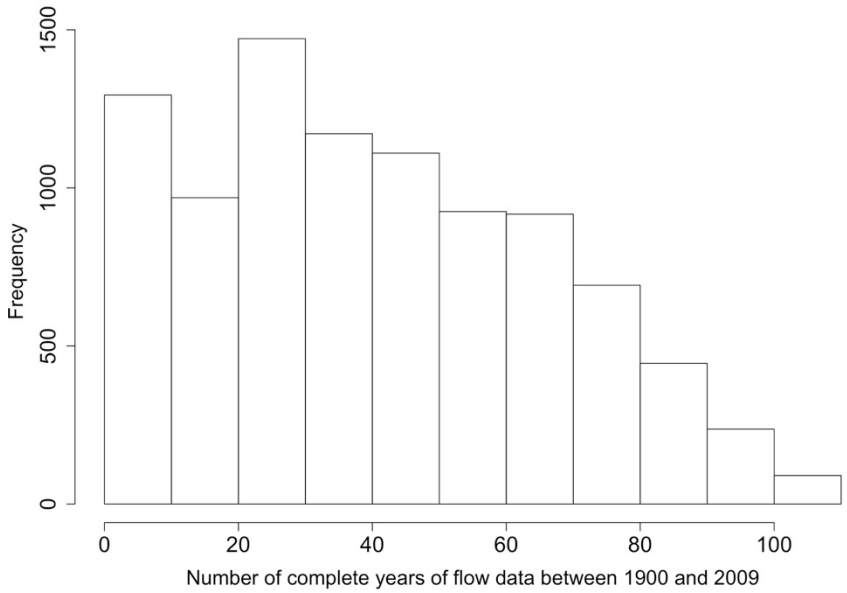

Figure 1: Histogram of the number of complete years of flow data at USGS Gages II locations.

The typical approach to determining the likelihood of extreme hydrological events is through statistical analysis of the historical record, which may have some regionalized information incorporated to extend its temporal coverage or improve its accuracy. (Büchele, et al., 2006; Castellarin, Merz, & Blöschl, 2009) This approach entails that the hydrologist fit a statistical distribution to the historical dataset, typically through the use of probability plots, and extrapolate the likelihood of large magnitude parameter occurrences. The fitting process relies upon the assumption that the historical data points are independent, temporally stationary, and follow the same statistical distribution (Chow, Maidment, & Mays, 1988). Moreover, the distribution used to characterize a historic dataset can vary and is, to a certain degree, subjective. For example, the United States (U.S.) Water Resources Council (USWRC) developed Bulletins 15, 17B, and 17C, with each being a revision of the previous. (Water Resources Council, 1967; Interagency Advisory Committee on Water Data, 1982; Advisory Committee on Water Information, 2015) The implementation of the most recent update, Bulletin 17C, utilizes a log-Pearson Type III (LP-III) distribution with regional skew coefficients to estimate flood flows and frequencies using the magnitudes of historic yearly maximum discharges. Because the Bulletin

17C methodology assumes a representative historic record, observations of extreme events within the historical record is crucial in accurately estimating future extremes. Bulletin 17C methodology can under-predict the occurrence of large event magnitudes when large events are not present in the historical record.

The probable maximum event is another common method to characterize extreme events. The probable maximum precipitation (PMP) is the greatest accumulation of precipitation over a given area and time of year that is meteorologically possible (World Meteorlogical Organization, 2009). To determine the PMP, modelers force numerical weather models with inputs that will lead to the greatest amount of precipitation for a given period and watershed. Many regulatory agencies, such as the U.S. Army Corps of Engineers (USACE), establish design criteria for dam design to account for the Inflow Design Flood (IDF). To find the inflow design flood (IDF) from the PMP, the engineer must model both runoff and routing processes. Because of the synthetic nature of the PMP process, there can be difficulty determining the return period or likelihood of these events.

The traditional methods of statistical extremes analysis, such as analysis of the historical record and PMP, leave knowledge gaps that can be supplemented by other approaches to characterize hydrological extremes. The purpose of this paper is to demonstrate a method of estimating the frequency of hydrological extremes based on the Shannon Entropy (AghaKouchak, 2014) (Singh, Hydrologic Synthesis Using Entropy Theory, 2011) (Singh, Byrd, & Cui, Flow Duration Curve Using Entropy Theory, 2014). Entropy methods are thought to circumvent many of the limitations of traditional statistical approaches by considering the full parameter space of a variable beyond the single realized historical sequence given by the gage record. This work begins by introducing and extending the entropy analysis given by Singh et al (2014) into an entropy-based methodology for extremes characterization. This research demonstrates the application of the entropy-based method for the characterization of extreme flow behavior within the Brazos River watershed. Finally, this research outlines additional work required to validate the predictive capability of entropy-based approaches for extremes analysis.

Methodology

Entropy is a system property that, when taken as a random variable, describes uncertainty in the state of a system through the range of the potential states available to that system. Shannon (1948) derived an entropy formulation which measures system entropy in terms of the state likelihood and the number of states, given as:

$$H(p) = -K \sum_{i=1}^{N} p(x_i)\, log\left(\frac{p(x_i)}{m(x)}\right) \quad (1)$$

where $H(p)$ is the entropy function, N is the number of system states, $p = \{p_i, i = 1, 2, \dots, N\}$ is the probably distribution giving the likelihood of the system being in a given state, K is a property of the logarithm base, and $m(x)$ is a function that maintains invariance under changes of coordinates (Shannon, 1948) (Singh, Hydrologic Synthesis Using Entropy Theory, 2011) (Singh, Byrd, & Cui, Flow Duration Curve Using Entropy Theory, 2014). If the number of states is large and the likelihood is diffused across the states, H will be relatively large in magnitude. If the number of states is small and the likelihood is concentrated on only a few states, H will be relatively small in magnitude. The magnitude of H therefore captures the system uncertainty as a single value by aggregating the number and likelihood of each available system state.

Entropy has seen significant use in water resources analysis, an overview of which is provided by Singh (2011). Applications of entropy theory have centered on improving the ability to model specific physical processes, such as infiltration, soil moisture, flow duration curves, and groundwater as well as capture various physical geometries. (Zehe, Blume, & Bloeschl, 2010; Eltahir & Gong, 1996; Hou, Huang, Leung, Lin, & Ricciuto, 2012; Woodbury & Ulrych, 1996; Moramarco & Singh, 2010) Most cases utilize maximum entropy theory to determine distributions which represent the likely state of a process or geometry of interest from the known information while not imposing artificial constraints on the parameter range. Cases also consider single subsystems and variables due to the complexity arising from multiple interacting physical processes.

The present work extends entropy theory into the estimation of extremes for watershed analysis. As traditional entropy theory considers a sample to be representative of the likely parameter state, the historical record can be considered as a single realization of all likely event sequences for a watershed. Other event sequences are similarly plausible given perturbations in the weather conditions or alternative historical sequences of hydrometeorlogical phenomena. This premise supposes that there exists a finite number of potential states that the parameter may assume that, when sampled repeatedly, would create a realization of an event sequence. Exploration of the underlying watershed states and production of valid historical sequences from the potential states has been the subject of much ongoing work due to the difficulty in accurately constructing and sampling the state space. Analysis of extremes using entropy theory greatly simplifies characterization of the potential states by capturing only the range of possible states and neglecting the exact behavior of any sequence realization.

The characterization of flow duration curves by Singh et al. (2014) forms the foundation on which entropy methods can be extended for extremes analysis of watershed parameters. Singh et al. applied Shannon entropy together with the principle of maximum entropy to evaluate flow duration curves by relating the range of states and the annual mean flow to the anticipated maximum flow. The proposed method follows directly from Singh et al (2014) and the result maximum entropy analysis as given by:

$$\frac{\bar{Q}}{Q_{max}} = \frac{1}{M} + \frac{\frac{Q_{min}}{Q_{max}} \exp\left(-\frac{MQ_{min}}{Q_{max}}\right) - \exp(M)}{\exp\left(-\frac{MQ_{min}}{Q_{max}}\right) - \exp(M)} \tag{2}$$

where $\bar{Q}$ is the average yearly flow, Q_{max} is the maximum yearly flow, Q_{min} is the minimum yearly flow, and M is the entropy parameter. Under the assumption that $Q_{min} \cong 0$, Equation 2 reduces to:

$$\frac{\bar{Q}}{Q_{max}} = \frac{1}{M} - \frac{\exp(-M)}{1 - \exp(M)} \tag{3}$$

The entropy parameter, M, is a dimensionless constant that is directly proportional to the range of possible maximum flow states with respect to the average and maximum flow states. A larger value for M indicates a larger range of discharge rates for that year. Utilizing the available historical record, M is calculated from the annual flow characteristics.

The analysis of extremes begins with Equation 3. Let $f(\bar{Q})$ and $g(M)$ be continuous random variables fit to the historical records for $\bar{Q}$ and M, respectively. Furthermore, let $F(\bar{Q})$ and $G(M)$ be defined as the cumulative distributions of the fits $f(\bar{Q})$ and $g(M)$. The exact distribution used

to model each may be chosen such that each parameter is characterized with the least error. Given the assumptions that:

1. The magnitude of M is uncorrelated with the magnitude of $\bar{Q}$
2. The range of M uncorrelated with the magnitude of $\bar{Q}$

it follows that any value of M is equally likely to apply to any value of $\bar{Q}$. If the assumptions are not upheld, the remainder of the analysis can still be conducted so long as the form of the dependency between M and $\bar{Q}$ is properly formulated. $Q_{max}(M, \bar{Q})$ may then be rewritten as function of both random variables using Equation 3, yielding:

$$Q_{max} = Q_{max}(G^{-1}(M), F^{-1}(\bar{Q})) \tag{4}$$

A surface characterizing Q_{max} can be produced by independently sampling $G^{-1}(M)$ and $F^{-1}(\bar{Q})$ with a sufficiently small interval to make subsequent operations independent of the sampling resolution. The Q_{max} surface represents the maximum extent of the event magnitude range for each combination of M and $\bar{Q}$. The frequency at which a given Q_{max} value occurs therefore gives the likelihood of encountering a given Q_{max} in a given year. A frequency distribution $p(Q_{max})$ for Q_{max} is obtained by binning and counting the occurrence of values on the Q_{max} surface. One may then optionally fit a distribution to the $p(Q_{max})$ histogram to simplify subsequent operations. This method results in a frequency estimate for Q_{max} which, while informed by the historical record, does not limit the potential states of the watershed to those present within the historical record and is separate from traditional annual maximum statistical estimation methods.

Results and Discussion

The proposed entropy method for extremes estimation is illustrated with discharge information from USGS Station 08082500 on the Brazos River near Seymour, Texas. The Brazos River was selected as it is well-gaged with 98 gage stations along its reach. The selected station has a contributing drainage area of 15,467 km^2 (5,972 mi^2) as shown in Figure 2 (USGS, 1990). Much of the watershed is cropland and pasture, particularly in the upper reaches (USGS, 2008).(USGS, 2008). The climate of the Brazos River basin is typical for the southcentral United States ranging from temperate to subtropical (Brazos River Authority, n.d.). The watershed provides a diverse range of flows, including extremes related to droughts and hurricanes. This gage station provided data including minimum, average, and maximum yearly flows available from 1924 to 2016.

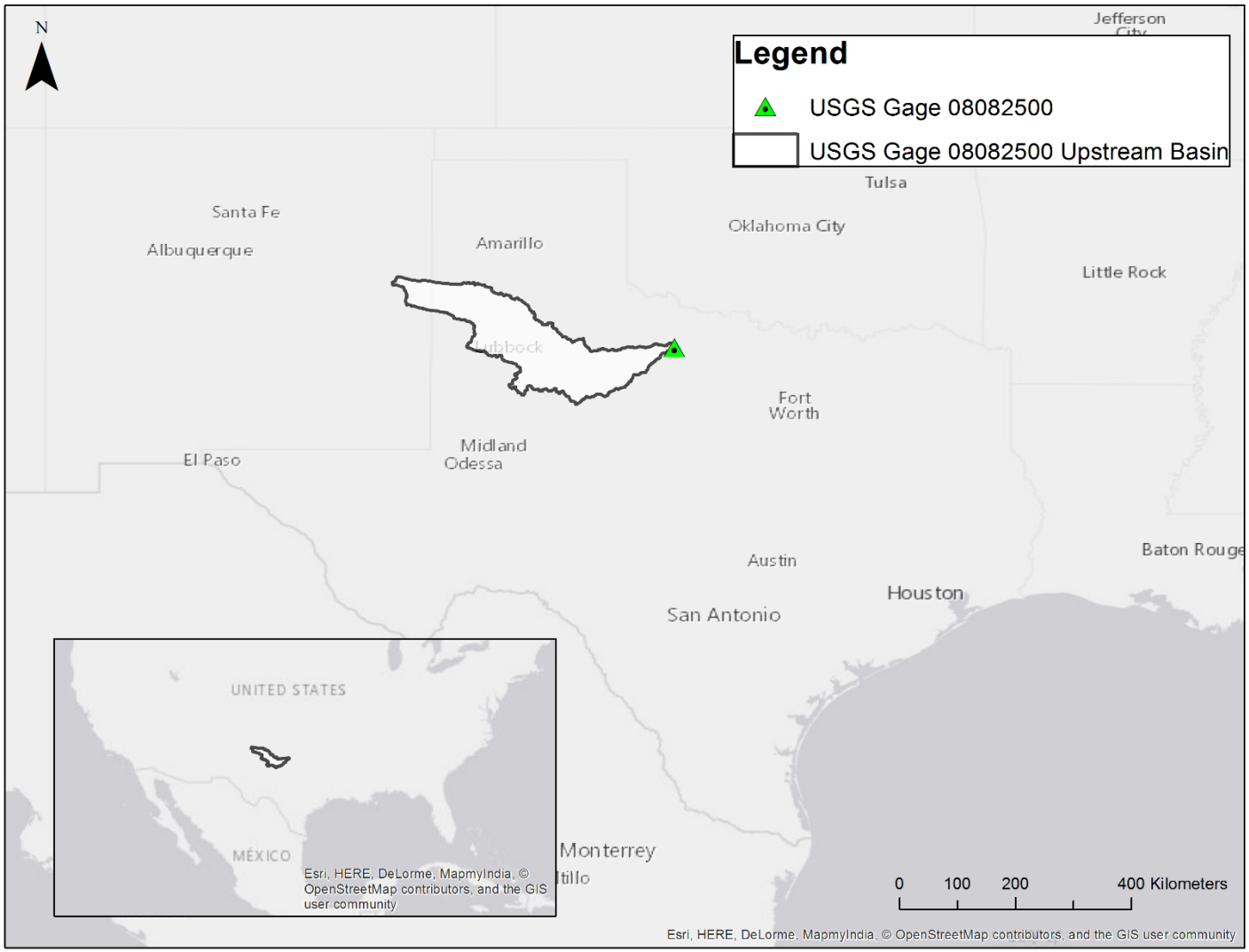

Figure 2: Map of upstream drainage area of gage 08082500

Historical values of M were calculated for each year from the mean and maximum following Equation 3. The independence of M and $\bar{Q}$ was verified with Figure 3 and by calculating the correlation between the series. As the correlation coefficient is low at 0.077, the parameters can be considered sufficiently independent for non-joint characterization.

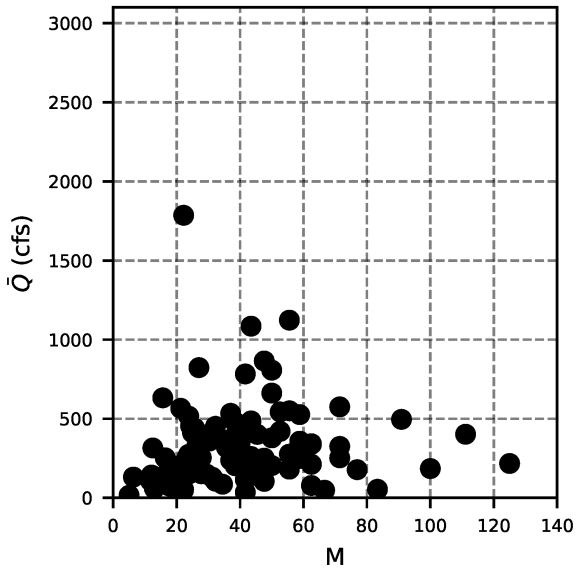

Figure 3: Test for correlation between $\bar{Q}$ and M within the Brazos River basin. The parameters are uncorrelated with a σ=0.077. $\bar{Q}$ is given in units of cubic feet per second (cfs).

Distributions were then fit to M and $\bar{Q}$ to enable parameter sampling. Best fits were achieved using the gamma distribution to characterize M and the Burr distribution to characterize $\bar{Q}$, achieving R^2 of 0.775 and 0.986, respectively. The full forms of both fits are given in Equations 5 and 6 with depictions of each illustrated in Figure 4.

$$f(M) = \frac{\left(\frac{M-c}{b}\right)^{a-1} \exp\left(-\frac{M-c}{b}\right)}{b\Gamma(a)} \qquad \begin{matrix} a = 3.494 \\ b = -0.369 \\ c = 11.774 \end{matrix} \tag{5}$$

$$g(\bar{Q}) = cd\left(\frac{\bar{Q}-m}{k}\right)^{-(c+1)}\left(1+\left(\frac{\bar{Q}-m}{k}\right)^{-c}\right)^{-(d+1)} \qquad \begin{matrix} c = 2.918 \\ d = 0.556 \\ k = -1.661 \\ m = 359.303 \end{matrix} \tag{6}$$

The surface characterizing Q_{max} was generated by sampling the inverse cumulative probability distributions for M and $\bar{Q}$ on the interval $0 \leq x \leq 0.999$ and solving Equation 3, the results of which are given in Figure 5. The dotted lines within the figure are probability contours which divide the surface into 25 regions of equal likelihood. The spacing between the contours

indicates the frequency at which Q_{max} values would be drawn from that region of the surface. Likely Q_{max} values bias strongly toward low mean flows and the lower portion of the entropy value range as these are the average conditions experienced within the Brazos watershed. Additionally, the historic Q_{max} are plotted with the color of each dot corresponding to the magnitude. The magnitude of the calculated surface strongly reflects the historical data which gives confidence that the entropy model correctly characterizes the historically sampled portion

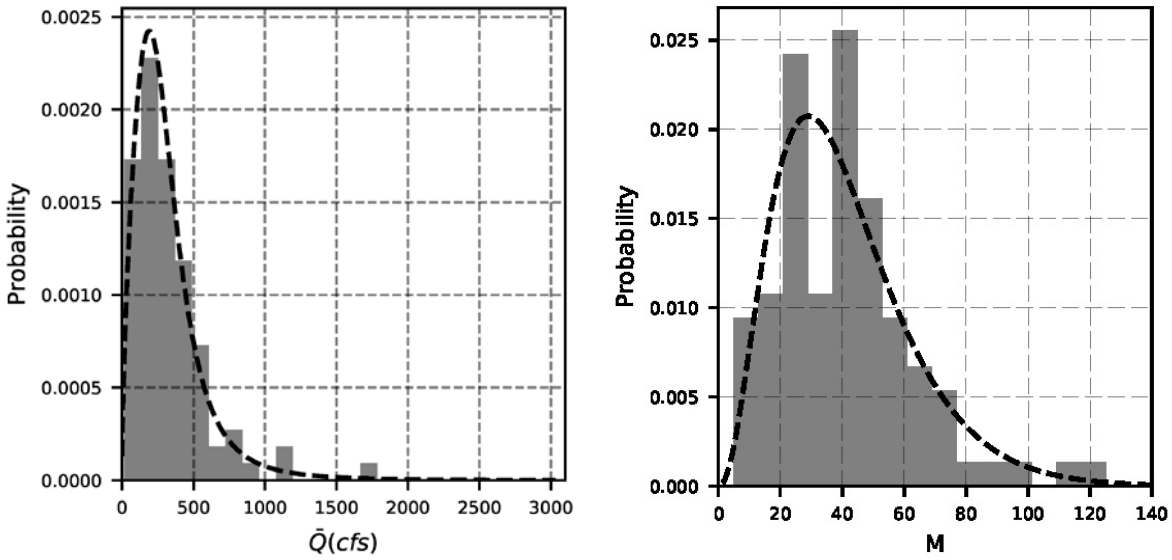

Figure 4: Left – Histogram and fit of the gamma distribution to $\bar{Q}$. Right – Histogram and fit of the Burr distribution to M.

of the state space. However, the historic measurements occupy only the lower magnitude, high sampling the likelihood region of the predicted state space. The predicted existence of a higher magnitude region suggests that there are discharge states available to the watershed that have not been sampled by the historical event sequence. If the accessibility of watershed parameters to some portion of the predicted high magnitude state space can be established, extreme event frequency estimates could be improved using the new state information over the historic record alone.

If accessibility to the full state space can be established, the surface characterizing Q_{max} can be applied in two useful manners. First, the annual likelihood of a given Q_{max} can be determined by finding the frequency of Q_{max} values on the surface, as given in Figure 5

significantly different behavior for low frequency events. Whereas the historical fit asymptotes to a constant, the entropy fit continues to increase with larger return periods.

. A Log-Pearson Type III (LP III) (R^2=0.965) as in Equation 7 characterized the occurrence of Q_{max} in this circumstance with least error (Figure 6). The distribution obtained for the occurrence of Q_{max} represents an improved estimate of the extreme event frequency as it more fully captures the range of states and the various watershed mechanisms which lead to similar parameter magnitudes.

$$P(Q_{max}) = \frac{1}{\sqrt{w\pi}\sigma Q_{max}} \exp\left(-\frac{(\log_{10} Q_{max} - \mu)^2}{2\sigma^2}\right) \qquad \begin{matrix}\mu = 9.041 \\ \sigma = 1.009\end{matrix} \qquad (7)$$

The surface can compare the statistical analysis of historical event sequence, as done in Figure 7. The empirical and LP III fit frequency estimates of the historical record are given by the circles and solid line, respectively. The empirical and LP III fit frequency estimates of the values occurring on Q_{max} surface are given by the squares and dotted line, respectively. The LP III fit to the historical record poorly characterizes the empirical historical estimate beginning at the 10-year return period. This causes extrapolations to longer return period events to have a large uncertainty range. The LP III fit to the Q_{max} surface frequency performs better but also under predicts the empirical distribution. Additionally, the Q_{max} frequency estimate demonstrates

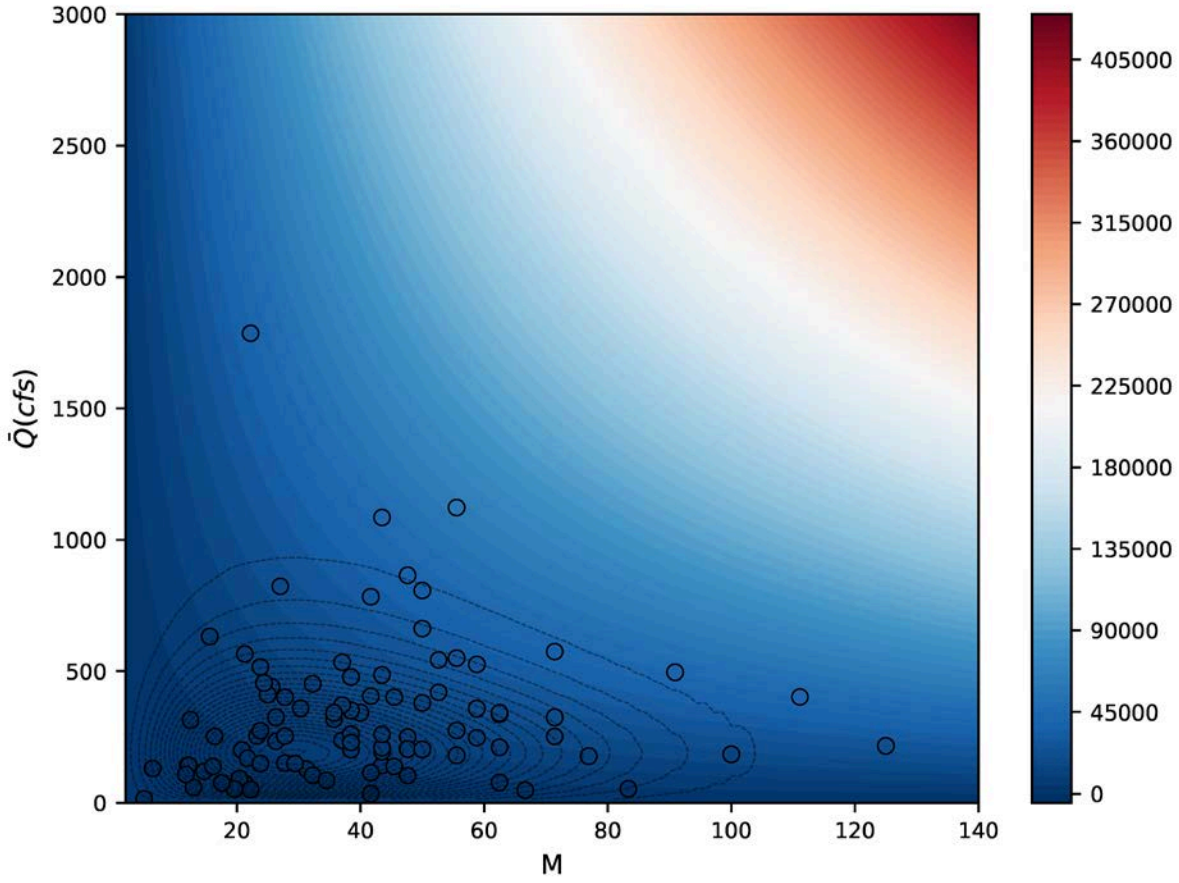

Figure 5: Predicted variation of Q_{max} as a function of M and $\bar{Q}$ produced by sampling the distributions and solving Equation 3. The color bar gives the magnitude of Q_{max} across the parameter space. The dotted contours divide the surface into 25 equally likely bins. Historical measured values are superimposed on the surface as the circle.

significantly different behavior for low frequency events. Whereas the historical fit asymptotes to a constant, the entropy fit continues to increase with larger return periods.

The difference between the historical and entropy fits is a product of the greater number of high magnitude states predicted by the entropy analysis. If these predicted states prove accessible to

the system, the difference between the fit curves would also characterize the information gained through the entropy analysis. In addition, the results portrayed in Figure 6 demonstrate that the empirical distribution of the Q_{max} surface and the LP III fit to the Q_{max} surface capture the largest event at Gage 08082500. The largest observed Q_{max} is approximately a 100-year return period event. These results indicate potential that the predicted states are accessible to the system and that the entropy-based surface is a viable indicator of extreme event magnitude. However, because of poor performance from 10-year to 100-year return period, additional research is necessary.

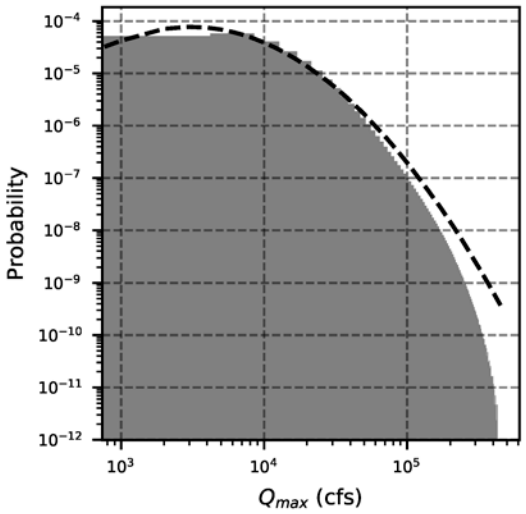

Figure 6: Histogram and fit of the LP III distribution to Q_{max}, giving the annual likelihood of Q_{max} being the maximum discharge within the Brazos River watershed. The distribution was created by counting the frequency of Q_{max} values on the surface in Figure 4.

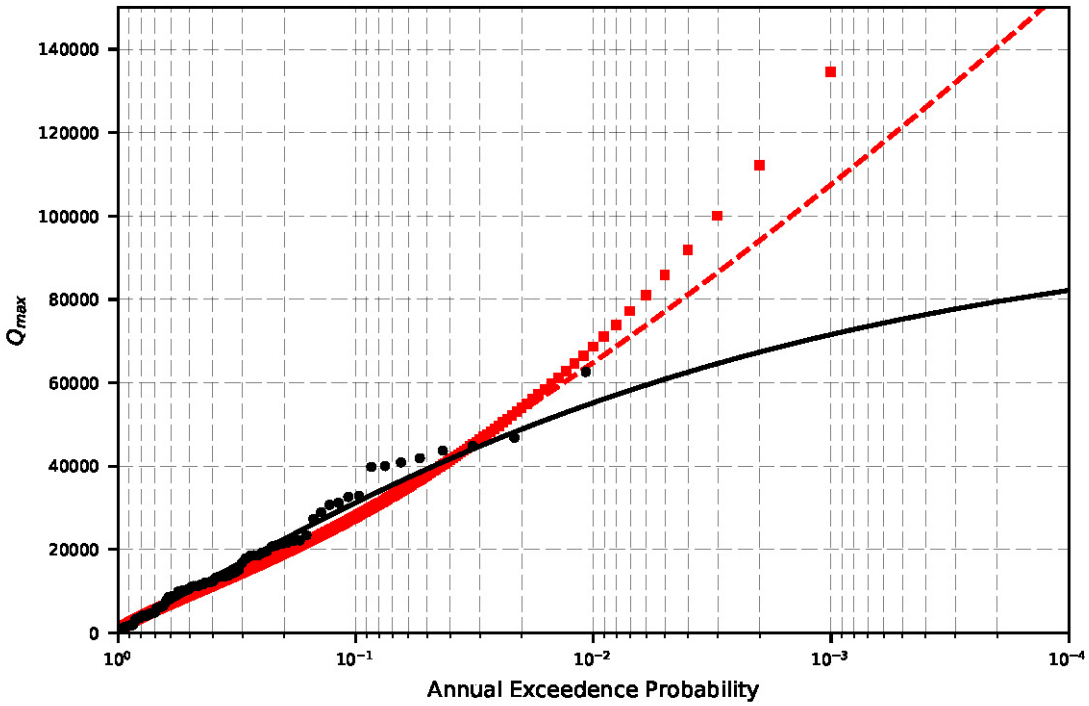

Figure 7: Estimation of the annual exceedance probability from the Q_{max} surface. The circles and solid line give the empirical and LP III fit for the historical record, respectively. The squares and dotted line give the empirical and LP III fit for the values occurring on the Q_{max} surface, respectively.

Conclusion

This work formulated and demonstrated an entropy-based approach to hydrologic analysis within the Brazos River watershed for estimating extreme event frequency. Expanding on the work of Singh et al. (2014), a method for characterizing the mean state and range of states available to watershed parameters was developed. These distributions were then sampled to create a magnitude surface that can be used to characterize the likelihood of the annual peak event magnitude.

Entropy theory applied to the watershed extremes analysis has significant promise for improving frequency characterization over existing methods. The outlined approach provides an independent estimate of event frequency from current statistical methods that rely solely on the local or regionalized historical record. This approach is also insensitive to the assumptions used to construct the tail region of existing statistical approaches from which extreme events are taken. Moreover, as the concepts of available states and maximum entropy underpin the given method, more information can be extracted about the potential parameter states available to the watershed.

At a minimum, an entropy-based extremes analysis can be useful to verify and provide an uncertainty estimate to the historical event sequence. Extension of the method to low frequency events requires subsequent research to demonstrate that the large magnitude states predicted by the method are in fact realizable. In addition, while the outlined method is expected to generalize broadly across other watershed parameters beyond discharge, the validity of this method must also be demonstrated under these cases.

References

Advisory Committee on Water Information. (2015). *Guidelines for Determine Flood Flow Frequency (draft).* Reston, VA: US Geological Survey.

AghaKouchak, A. (2014). Entropy-Copula in Hydrology and Climatology. *Journal of Hydrometeorology, 15*(6), 2176-2189.

Büchele, B., Kreibich, H., Kron, A., Thieken, A., Ihringer, J., Oberle, P., . . . Nestmann, F. (2006). Flood-risk mapping: contributions towards an enhanced assessment of extreme events and associated risks. *Natural Hazards and Earth System Sciences, 6*(4), 485–503.

Brazos River Authority. (n.d.). *What is the Brazos River?* Retrieved June 25, 2017, from Brazos River Authority: https://www.brazos.org/About-Us/Education/Water-School/ArticleID/265

Castellarin, A., Merz, R., & Blöschl, G. (2009). Probabilistic envelope curves for extreme rainfall events. *Journal of Hydrology, 378*(3-4), 263-271.

Chow, V. T., Maidment, D. R., & Mays, L. W. (1988). *Applied Hydrology.* Tata McGraw-Hill Education.

Eltahir, E., & Gong, C. (1996, May). Dynamics of wet and dry years in West Africa. *Journal of*

Climate, 9(5), 1030-1042.
Hou, Z., Huang, M., Leung, L., Lin, G., & Ricciuto, D. (2012, August 10). Sensitivity of surface flux simulations to hydrologic parameters based on an uncertainty quantification framework applied to the Community Land Model. *Journal of Geophysical Research-Atmospheres, 117*.
Interagency Advisory Committee on Water Data. (1982). *Guidelines for Determining Flood Flow Frequency.* Reston, VA: US Geological Survey.
Kvocka, D., Falconer, R., & Bray, M. (2016). Flood Hazard Assessment for Extreme Flood Events. *Natural Hazards, 84*(3), 1569-1599.
Moramarco, T., & Singh, V. (2010, October). Formulation of the Entropy Parameter Based on Hydraulic and Geometric Characteristics of River Cross Sections. *Journal of Hydrologic Engineering, 15*(10), 852-858.
Rajsekhar, D., Mishra, A., & Singh, V. (2013). Regionalization of Drought Characteristics Using an Entropy Approach. *Journal of Hydrologic Engineering, 18*(7), 870-887.
Ross, T., & Lott, N. (2003). *A Climatology of 1980-2003 Extreme Weather and Climate Events.* US Department of Commerce. Asheville: National Climatic Data Center.
Shannon, C. E. (1948). *The Mathematical Theory of Communications I and II.* Bell Systems Tech Journal.
Singh, V. J. (1996). The Use of Entropy in Hydrology and Water Resources. *Hydrological Processes, 11*, 587-626.
Singh, V. J. (2011). Hydrologic Synthesis Using Entropy Theory. *Journal of Hydrologic Engineering, 16*(5), 421-433.
Singh, V. J., Byrd, A. R., & Cui, H. (2014). Flow Duration Curve Using Entropy Theory. *Journal of Hydrologic Engineering, 19*(7), 1340-1348.
USGS. (1990). *Largest Rivers in the United States.* Department of the Interior. USGS.
USGS. (2008, October 8). *Brazos Watershed*. Retrieved 06 25, 2017, from EDNA Derived Watersheds for Major Named Rivers: http://edna.usgs.gov/watersheds/ws_ws_chars.php
USGS. (2011). *Geospatial Attributes of Gages for Evaluation Streamflow.* Reston, Virginia: United States Geological Survey (USGS).
Water Resources Council. (1967). *A Uniform Technique for Determining Flood Flow Frequencies.* Washington, D.C.: Water Resources Council.
Woodbury, A., & Ulrych, T. (1996, September). Minimum relative entropy inversion: Theory and application to recovering the release history of a groundwater contaminant. *Water Resources Research, 32*(9), 2671-2681.
World Meteorlogical Organization. (2009). *Manual on Estimation of Probable Maximum Precipitation (PMP).* Geneva, Switzerland: World Meteorlogical Organization.
Zehe, E., Blume, T., & Bloeschl, G. (2010, May 12). The principle of 'maximum energy dissipation': a novel thermodynamic perspective on rapid water flow in connected soil structures. *Philosophical Transactions of the Royal Society B-Biological Sciences, 365*(1545), 1377-1386.