A Simple Data-Driven Model for Streamflow Prediction

Valliappa Lakshmanan[1,2], Jonathan J. Gourley[2],Zac Flamig[1,2],Scott Giangrande[3*]

Abstract

It is sometimes useful to create a statistical model to simulate streamflow based on precipitation estimates over a basin. Because the model is independent of physical descriptions of the basin or initial states such as soil moisture conditions, soil infiltration characteristics, or land surface roughness, it can be used to justify the complexity required in a hydrologic model to adequately simulate streamflow.

In this paper, we developed a data-driven streamflow prediction model using observations of rainfall and runoff over the heavily instrumented Ft. Cobb basin in western Oklahoma. The statistical model was developed for a dataset of ten hydrologic events in which there was complete coverage by the KOUN polarimetric research radar and streamflow observations on three subcatchments in the basin. The data-driven model was evaluated for each event by considering the rainfall/runoff observations from the other nine, independent events. In this paper, we focus on model results from Tropical Storm Erin which produced streamflow having a return period greater than 100 years. Future work will compare results from the data-driven approach to more complex and distributed-parameter models. Inadequacies highlighted by the data-driven approach will justify the added complexity of physics and spatially distributed parameters represented in the conceptual models.

1. Introduction

The Ft. Cobb watershed was added to the Agricultural Research Service's (ARS) watershed research network in 2005 to address research objectives related to constituents that impair water quality and wildlife habitat. The basin is 813 km^2 in area and features a Micronet, a network of 15 stations that measure air temperature, rainfall, relative humidity, solar radiation, soil temperature at four depths, and soil water content at three depths. A majority of the basin is within 100 km of the KOUN radar, the WSR-88D prototype equipped with the capability to transmit and receive at horizontal and vertical polarization. We have archived polarimetric radar datasets for ten hydrologic events occurring over a three-year period, including an extreme event

[*] Corresponding author: lakshman@ou.edu [1]The Cooperative Institute of Mesoscale Meteorological Studies (CIMMS), University of Oklahoma, [2]The National Severe Storms Laboratory, Norman, OK, [3]McGill University, Montreal, Canada

from a tropical storm that had a return period greater than 100 years. The duration of rainfall for these events ranged from 6-61 hours, and the characteristics of the storms included intense convective supercells with severe hail, squall lines with trailing stratiform rain, and tropical rain. The high-density rain gauge network was used to evaluate remote-sensing and in-situ rainfall algorithms. Six different polarimetric rainfall algorithms employing reflectivity, differential reflectivity, specific differential phase, and combinations have been proposed and are now compared to reflectivity-only and gauge-only algorithms. The first level of comparison simply evaluated differences between the algorithms' rainfall estimates and collocated rain gauge accumulations. The next evaluation considers the hydrologic sensitivity of using these different rainfall inputs into a suite of hydrologic models. In this paper, we concentrate on fitting a finite impulse response function to nine events and using that to build an empirical hydrological model to be evaluated on TS Erin.

The term *genetic algorithm* (GA) is applied to any search or optimization algorithm that is based on Darwinian principles of natural selection. A key concept in genetic algorithms is that of a chromosome. A chromosome contains a group of numbers that completely specifies a candidate during the optimization process (Goldberg 1989; Lakshmanan 2000). The most fit members (the ones for which fitting error is least) are more likely to be propagated into the next generation. Propagation takes two forms: crossover, where the new chromosome consists of parts of two chromosomes in the current generation, and mutation where a chromosome in the current generation is subtly modified. The crossover points and mutated numbers are chosen randomly. It has been shown (Goldberg 1989) that the process of selecting the most fit members of a generation to propagate results in a steadily improving population i.e. optimization. In addition to the genetic algorithm, we used simulated annealing (Metropolis et al. 1953) to move the top 10% of a generation to their nearest local minima.

2. Method

The basin was divided into three bands according to terrain elevation, and the rainfall within each band was estimated using polarimetric radar using six different techniques, including those proposed by Ryzkhov et al. (2005) and Bringi and Chandrasekar (2001). A rain gauge analysis field was also supplied to the model by analyzing the Micronet rain gauge accumulations using inverse-distance weighting with a ”leave one out” optimization step. Average rainfall within each band are used to simulate streamflow at three basin subcatchments.

A model that simply tries to predict the streamflow based on observations at a snapshot in time will not work because the streamflow is based on rainfall over a time interval. It is not enough to use rainfall estimates at just the current time.

Hence, we chose to model the streamflow as a finite-impulse response (FIR) model (Brown and Hwang 1997). The model chosen was of this form:

$$\hat{S}(t) = \sum_{k=0}^{n} \sum_{j=0}^{t-\tau_k} (t-j-\tau_k)\beta_k e^{\frac{t-j-\tau_k}{-\alpha_k}} E_k(j) \quad (1)$$

where $\hat{S}(t)$ is the estimate of the streamflow at time t and E_k is a time series of measurements of rainfall over the k^{th} band. There are three parameters for each of n bands: β_k, α_k and τ_k. The impact of a band is given by β_k, the time delay between rainfall at the k^{th} band and the time that water from it reaches the stream is given by τ_k while the decay rate is given by α_k.

It may help the reader to gain a feel for Equation 1 if we were to consider the response of the function to an impulse response i.e. suppose there were to be rainfall of E_0 at time $t = 0$ and no rainfall after that. The streamflow that would result because of this would begin at $t = \tau_0$, reach a peak and then decay exponentially as shown in Fig. 1. Since any digital series of measurements can be considered a sum of such $E_0(j)$, the resulting streamflow is the sum of curves of the form shown in Fig. 1. The smaller the α, the slower the curve decays. The larger the β, the faster and higher it ramps up. The larger the τ, the greater the delay before the effect of rainfall shows up in the streamflow.

We used $n = 3$, i.e. 3 bands in all the work presented here. To fit this model to streamflow observations, it is necessary to estimate the parameters α_k, β_k and τ_k that minimize some measure of error. We chose to use the mean-square-error as the criterion to minimize.

A conceivable way to accomplish this would be to take advantage of the fact the Fourier transform of a convolution of two data sets is the product of the Fourier transforms of the two data sets. The Fourier transform of $\hat{S}(t)$ follows the form of:

$$F(e^{-t/\alpha}u(t)) = \frac{1}{1/\alpha + j2f\pi} \tag{2}$$

where $F()$ is the Fourier transform, $u(t)$ the unit step function and j the square root of -1. $u(t)$ is required in order to maintain causality. As can be seen, the Fourier transform of the model would be a complex, nonlinear function of f. Thus the problem becomes one of generalized non-linear optimization of a complex dataset.

A more tractable approach is to work within the time domain itself. To do so, it is enough to realize that if α_k and τ_k are known, the parameters b_k can be estimated using linear regression. A genetic algorithm was used to create reasonable values of α_k and τ_k. Then, linear regression was used to compute the best estimate of β_k. The resulting square error was used to determine the fitness of α_k and τ_k, for future generations of the genetic algorithm.

3. Results

The Ft. Cobb watershed is 813 km^2 in area and has elevations varying from 383-565 m. Each of the three subcatchments were subdivided into three topographic bands where each band has equal area.

The watershed is fortuitously situated within two observational networks. Two Oklahoma Mesonet sites (FTCB and HINT) are located on the sides of the watershed and are capable of measuring standard meteorological variables as well as soil temperature and moisture at three depths. In addition, fifteen ARS-owned and Mesonet-operated Micronet stations are located completely within the watershed. River discharge is measured by three USGS stream gauges in the basin having catchment areas of 75, 154, and 342 km^2; the latter two are small enough to be considered flash flood basins.

A majority of the watershed lies within 100 km of the WSR-88D prototype of an S-band polarimetric radar (KOUN) located in Norman. 15-min rainfall accumulations were estimated from the radar over the entire watershed and then averaged within each of the topographic bands. Thus, at every time instant, there were three estimates of rainfall within each subbasin. Time series of rainfall estimates were used to simulate streamflow.

Training of the model was carried out separately on nine events:

- 12-16 Jun 2005: Maximum rainfall accumulation of 96 mm.

- 30 Sep - 3 Oct 2005: Maximum rainfall accumulation of 60 mm. Several severe hail reports were associated with this event.
- 14-17 Jun 2007: Maximum rainfall accumulation of 201 mm. Several severe hail reports were associated with this event.
- 20-23 Jun 2007: Maximum rainfall accumulation of 72 mm.
- 27 Jun - 2 Jul 2007: Maximum rainfall accumulation of 120 mm.
- 2-4 Mar 2008: Maximum rainfall accumulation of 32 mm, with several severe hail reports and tornadic storms.
- 31 Mar - 7 Apr 2008: Maximum rainfall accumulation of 74 mm.
- 7-13 May 2008: Maximum rainfall accumulation of 81 mm.
- 9-13 Jun 2008: Maximum rainfall accumulation of 78 mm.

The empirical hydrologic model is an ensemble of nine functions of the form of Equation 1. The fitted functions closely matched the observed streamflow, as illustrated in Fig. 2 for the 12-16 June 2005 calibration case.

The ensemble was independently evaluated on the 18-20 Aug 2007 case (Tropical Storm Erin). Maximum rainfall accumulation was 307 mm in 24 hr which resulted in significant flooding claiming two lives.

The set of fitted functions based on the individual FIR functions with rainfall inputs from the synthetic algorithm of (Ryzkhov et al. 2005) encompassed the extreme event quite well for the 342 km^2 catchment (upper-left panel in Figure 3), which is impressive considering no events of this extreme magnitude were included in the training data set. Figure 3 also indicates the skill in simulating runoff apparently decreases with smaller basin sizes. From this initial study, we can infer that additional complexity in hydrologic modeling e.g. inclusion of initial soil moisture conditions, distributed soil types and land cover roughness values, etc. may be warranted at the flash flood scale whereas simplistic, parsimonious approaches are sufficient for larger basins.

In considering the hydrologic model sensitivity to different inputs, simulations using the synthetic algorithm were more skillful than those using the conventional, R(Z) algorithm and rain gauge inputs. These differences were more noticeable at smaller basin scales. This suggests the impact of upgrading the NEXRAD network with polarimetric capabilities will be significant to hydrologic prediction, especially at the flash flood scale.

We plan to extend this work to evaluate hydrologic modeling skill by iteratively adding complexity to the FIR model and determining which variables or parameters are needed at each subcatchment scale. In addition, we will compare results to those from conceptual models having distributed parameters and continuous soil moisture accounting. Ultimately, we will arrive at an ensemble of hydrologic simulations with uncertainty bounds based on performance-weighted inputs and models.

Acknowledgments

Funding for this research was provided under NOAA-OU Cooperative Agreement NA17RJ1227. The statements, findings, conclusions, and recommendations are those of the authors and do not necessarily reflect the views of the National Severe

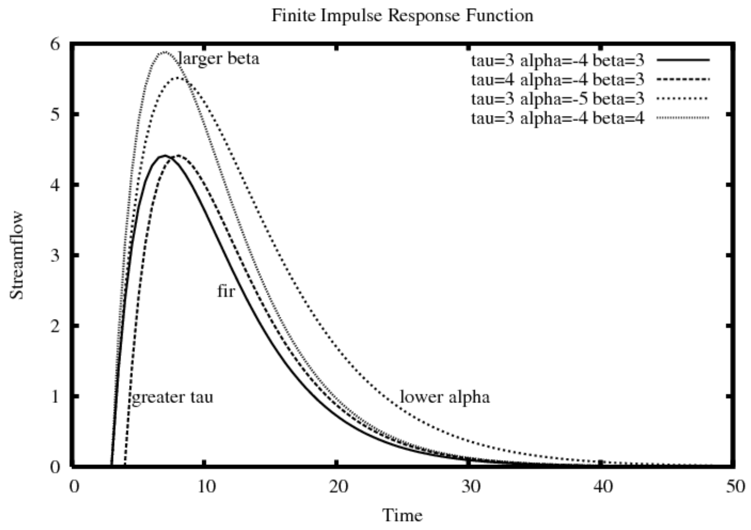

Figure 1: The finite impulse response (FIR) transfer function. The solid curve shows the effect of a unit of rainfall at $t = 0$ in band $k = 0$ and zero rainfall everywhere else. The dashed and dotted curves show the effects of increasing beta, decreasing alpha and increasing tau from the baseline parameters. The parameters themselves are shown in the legend

Storms Laboratory (NSSL) or the U.S. Department of Commerce.

References

Bringi, V. and V. Chandrasekar, 2001: *Polarimetric Doppler Weather Radar: Principles and Applications*. Cambridge University Press, 636 pp.

Brown, R. and P. Hwang, 1997: *Introduction to Random Signals and Applied Kalman Filtering*. John Wiley and Sons, New York.

Goldberg, D., 1989: *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Publishing Company, Inc., 432 pp.

Lakshmanan, V., 2000: Using a genetic algorithm to tune a bounded weak echo region detection algorithm. *Journal of Applied Meteorology*, **39**, 222–230.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, 1953: Combinatorial minimization. *J. Chem. Phys.*, **21**, 1087–1092.

Nash, J. and J. Sutcliffe, 1970: River flow forecasting through conceptual models Part i: A discussion of principles. *J. Hydrology*, **10**, 282–290.

Ryzkhov, A., S. Giangrande, and T. Schuur, 2005: Rainfall estimation with a polarimetric prototype of WSR-88D. *J. Appl. Meteor.*, **44**, 502–515.

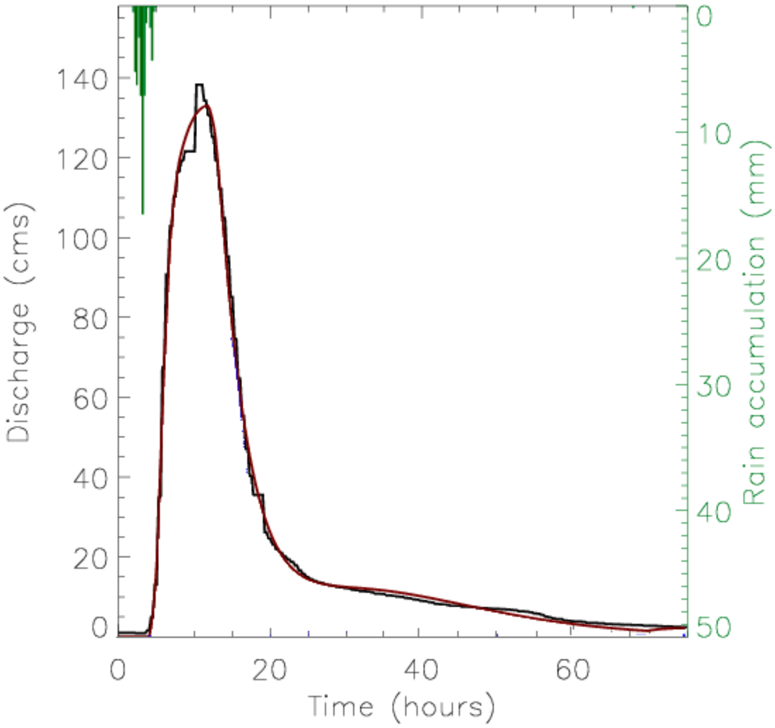

Figure 2: The FIR function is capable of being fit to closely match observed streamflow based on input rainfall. The input rainfall (green), observed streamflow (black) and simulated streamflow (dark red) at a USGS stream gauge with a 342 km^2 catchment area for the 12-16 Jun 2005 event are shown.

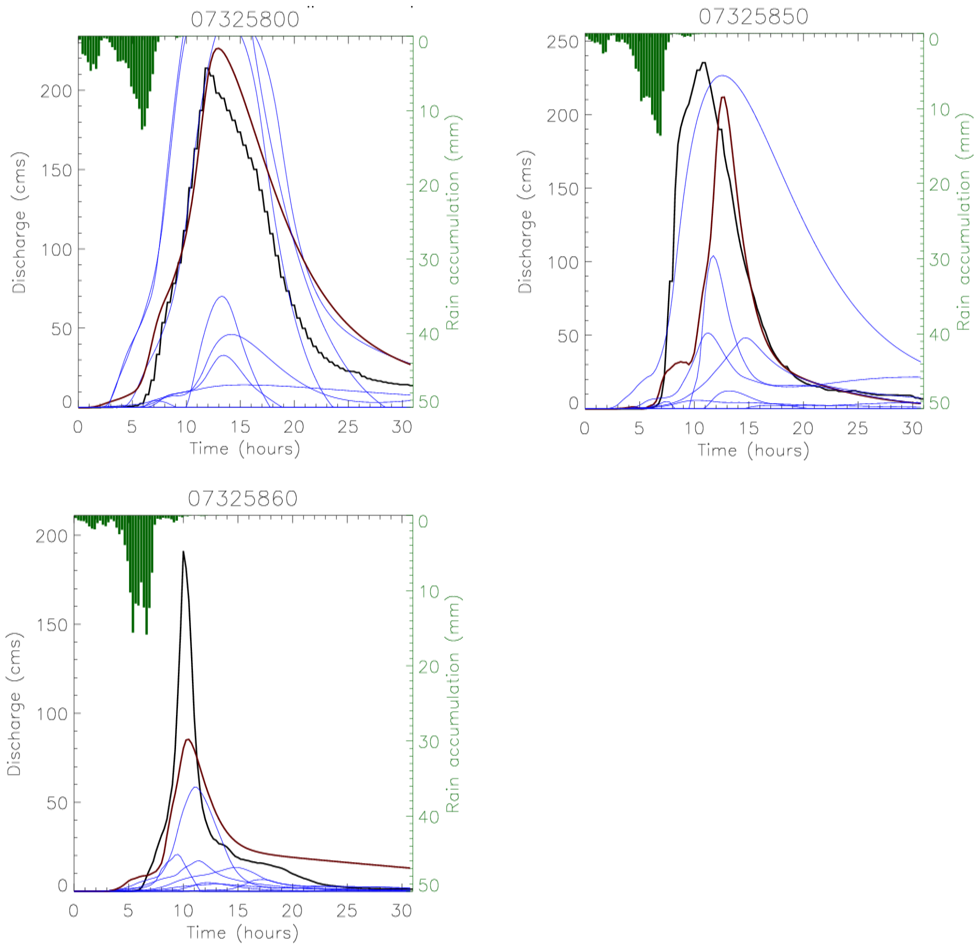

Figure 3: Radar-derived rainfall from the synthetic algorithm (green), observed streamflow (black), and the ensemble of data-driven FIR simulations of streamflow (blue) for an independent test event (Tropical Storm Erin) at 3 USGS stream gauges (USGS ID noted at top of each panel). The best member of the ensemble, according to Nash-Sutcliffe score (Nash and Sutcliffe 1970), is shown by the dark red line. Note that even though extreme events were not part of the training sample, this data-driven approach does manage to capture the streamflow that would result from 308 mm of maximum rainfall observed in TS Erin.