Jonathan Frame

Cloud to Street Hydrologist Technical Test

August 25th 2021

1 Estimate the return period for the 2019 Tulsa flood event

1.1 Introduction

Return period analysis is one of the most important aspects of hydrological science. The endeavor is not particularly scientific, although some attempts have been made to add rigour, but it is critical for designing infrastructure and natural resources management. In general, a return period analysis is done to understand how likely an event is to occur, often to assess potential risks. Since water managers can not plan for every scenario, they can use return period analysis to plan for scenarios that meet some threshold for likelihood.

1.1.1 Calculating the return period

1.1.1.1 Fitting a probability distribution

Return periods can be calculated in a number of ways, but in general includes fitting a probability distribution to the data, and interpolating an event onto the curve. Without a probability distribution we can only estimate the return period based on the number of records available. For instance, if we have a 50-year record we can assume the largest event during that period is the 50-year event. If we want to extend to a 100-year event, we need to assume the 50 years are a sample that falls on the curve of a probability distribution. We can never be sure what the appropriate probability distribution is, since we are always limited by the data record, but there are commonly accepted distributions, particularly in the United States. We do know for certain that the probability distribution should be heavy tailed. This is because there is always a non-zero probability of larger events occurring.

1.1.2 Defining an event

The peak streamflow runoff is my preferred metric for return period calculations. The trouble with using the precipitation is defining an event, and the trouble with using the inundation area is assessing the role that infrastructure and land use/land cover had in the event.

Although we can calculate the return period of any event, or summary of event (peak or average of streamflow, precipitation or water surface elevation), we run into some trouble with deciding on the characteristic for each type of event. A return period calculation on a peak event requires only the decision of which probability distribution to use, but the return period calculation for an averaged condition requires some decision made on how to determine the average. If we try to make a calculation on not the peak daily rainfall, since a large single day's precipitation might not result in a streamflow as large as *n* straight days of moderate precipitation, we would need to make a decision on the number of days to include in the average. This additional information

adds biases to our analysis. A similar problem arises if we try to use flood inundation for our event. Flood infrastructure may operate differently (or even fail in some circumstances), making the event open for interpretation. Two identical upstream streamflow hydrographs may result in different downstream flood extents if a dam operator is slow to open some outlet gate, or if a river adjacent to a farmer is in the middle of planting or harvesting.

1.1.2 Accounting for non-stationarity

One problem with the idea of exceedance probability is that the timeframe we are typically interested in (the reasonably near future) is not necessarily represented by the past events. There are always changes to a hydrological system, be it climatological, geologica, geographical, industrial, political, etc.

This return-period analysis does not account for nonstationarity -- i.e., the return period of a given magnitude of event in a given basin could change due to changing climate or changing land use. There is currently no agreed-upon method to account for nonstationarity when determining return periods, so it would be difficult to incorporate this in these calculations. The method I use (described below) includes high and low estimates for each return period classification, which are shown in the resulting graphs presented in section 3. Those high and low estimates give us a range of flows to expect with a given return period, but in this exercise we are estimating the return period for a given event.

1.2 Methods

To calculate the return period of the event I used the The U.S. Interagency Committee on Water Data Bulletin 17b (IACWD, 1982). There is an updated methodology, Bulletin 17c, but I chose to use the 17b version because I was not able to find system compatible software for the newer version.

1.2.1 Data

I calculated the return periods on the streamflow and precipitation records provided by C2S, however this was mostly as a source of comparison. I downloaded the annual peak data from the National Water Information System (NWIS), which is a more appropriate dataset than the hourly or daily data for this type of analysis. Bulletin 17b guidelines require annual peak flows and use all available data. It turns out that the annual peaks of the provided hourly data match those in the official peak streamflow dataset on NWIS, but this is not always the case, as some gauging stations record peak streamflow values differently than streamflow time series.

1.2.2 Log Pearson III

I followed guidelines in the U.S. Interagency Committee on Water Data Bulletin 17b (IACWD, 1982). The procedure is to fit all available annual peak flows (log transformed) for each basin to a Pearson Type III distribution using the method of moments:

$$f(y;\ \tau, \alpha,\ \beta)\ =\ \frac{\left(\frac{y-\tau}{\beta}\right)^{\alpha-1} e^{\left(-\frac{y-\tau}{\beta}\right)}}{|\beta|\Gamma(\alpha)}$$

with $(y\text{-}\tau)/\beta > 0$ and distribution parameters τ, α, and β, where τ is the location parameter, α is the shape parameter, β is the scale parameter, and $\Gamma(\alpha)$ is the gamma function.

The Bulletin 17b (IACWD, 1982) guidelines require using all available data. For the hourly discharge data provided 1987-2020 (complete water years), mean daily discharge data provided 1972-2020 (again, complete water years) and for precipitation is 1981-2020 (once more, complete water years). After fitting the return period distributions, each water year is classified according to the return period of its annual peak. I used Matlab code (Burkey, 2009) from MathWorks File Exchange to fit annual peak flow distributions and obtain return period estimates for type of data. This software includes the exact process and logic outlined in Bulletin 17b.

1.3.0 Results

Table 1. Results of return period calculations based on Log Pearson type III probability distribution and the methodology presented in the Bulletin 17b.

Data type (Annual peaks)	Return period (years)	years in sample	Figure reference
Peak annual streamflow Haskell*	**566**	48	5
Hourly streamflow Haskell	293	34	1
Daily average streamflow Haskell	193	50	2
Peak annual streamflow Muskagee*	81	63	6
Peak 10 day cumulative precipitation	4	39	4
Peak daily precipitation	2	39	3

* These data were not provided for this technical test, these were downloaded from National Water Information System

Some python code used to analyze

1.3.0.1 Return period calculated with Log Pearson III from provided hourly and daily streamflow data

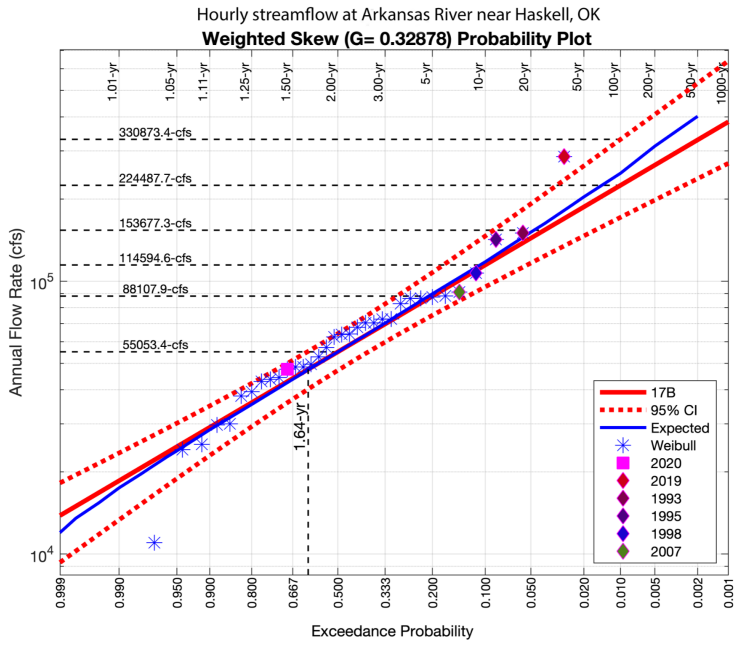

Figure 1. Bulletin 17b result for USGS 07165570 Arkansas River near Haskell, OK from hourly flow records from 1987-2021.

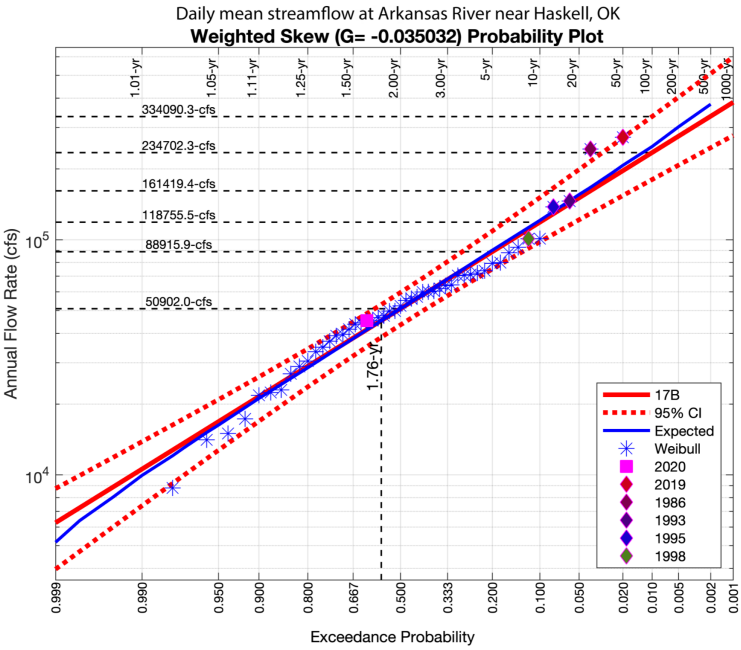

Figure 2. Bulletin 17b result for USGS 07165570 Arkansas River near Haskell, OK from daily averaged flow records from 1982-2021.

1.3.0.2 Return periods based on precipitation data

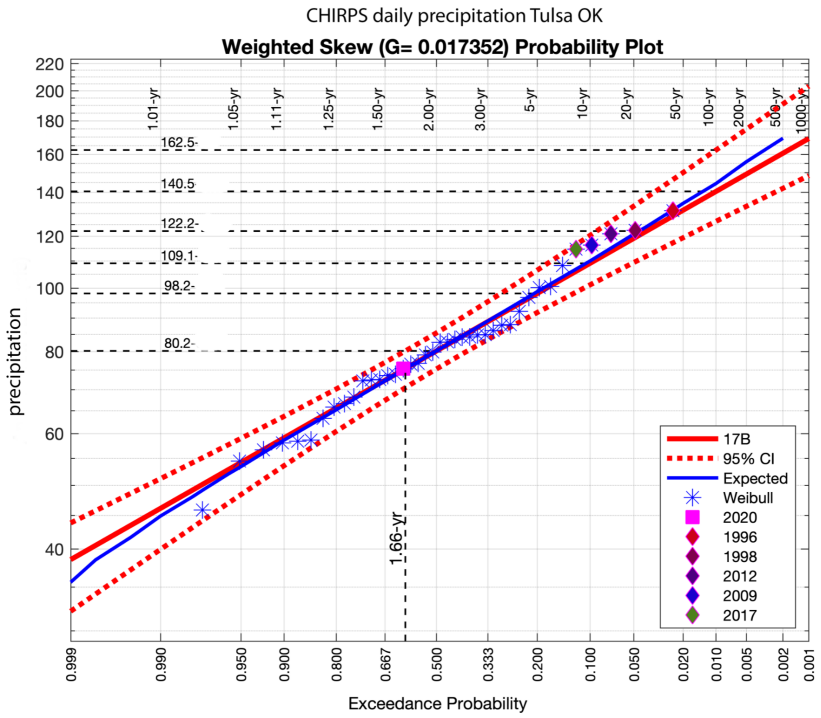

Figure 3. Bulletin 17b results from daily averaged precipitation, peak annual events, 1982-2021.

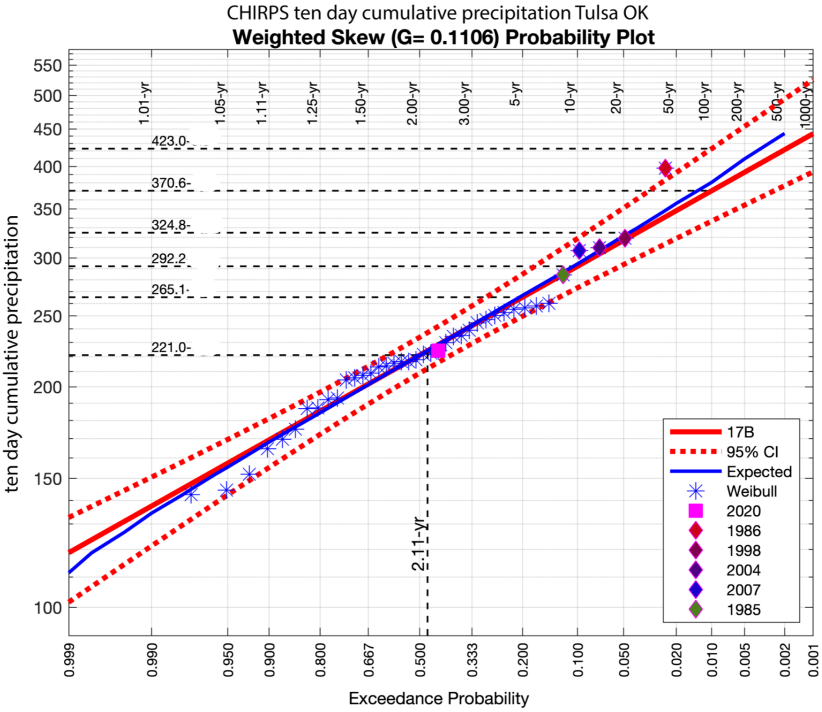

Figure 4. Bulletin 17b results from daily averaged precipitation, peak 10 day cumulative precipitation, 1982-2021.

1.3.0.3 Peak runoff return period calculation from Annual Peak Data

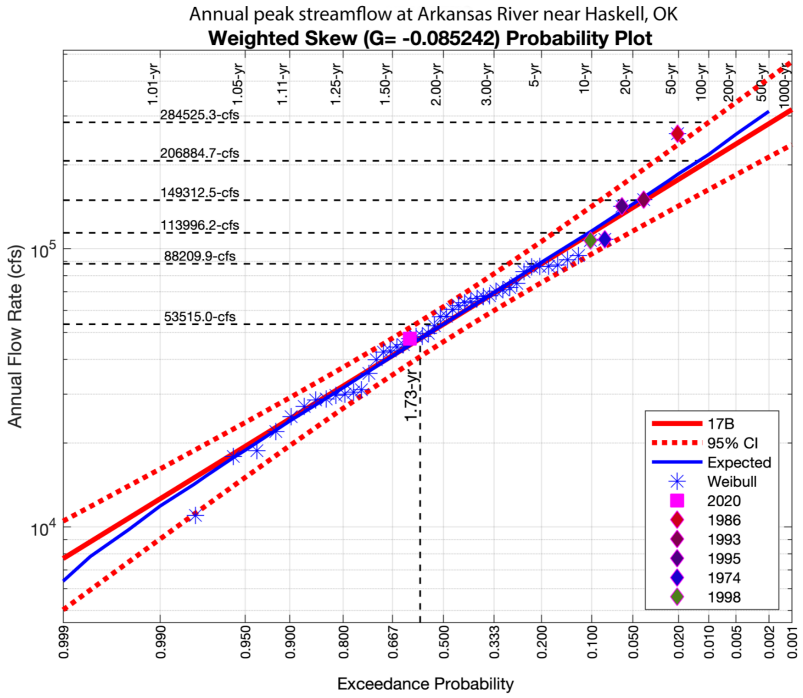

Figure 5. Bulletin 17b result for USGS 07165570 Arkansas River near Haskell, OK

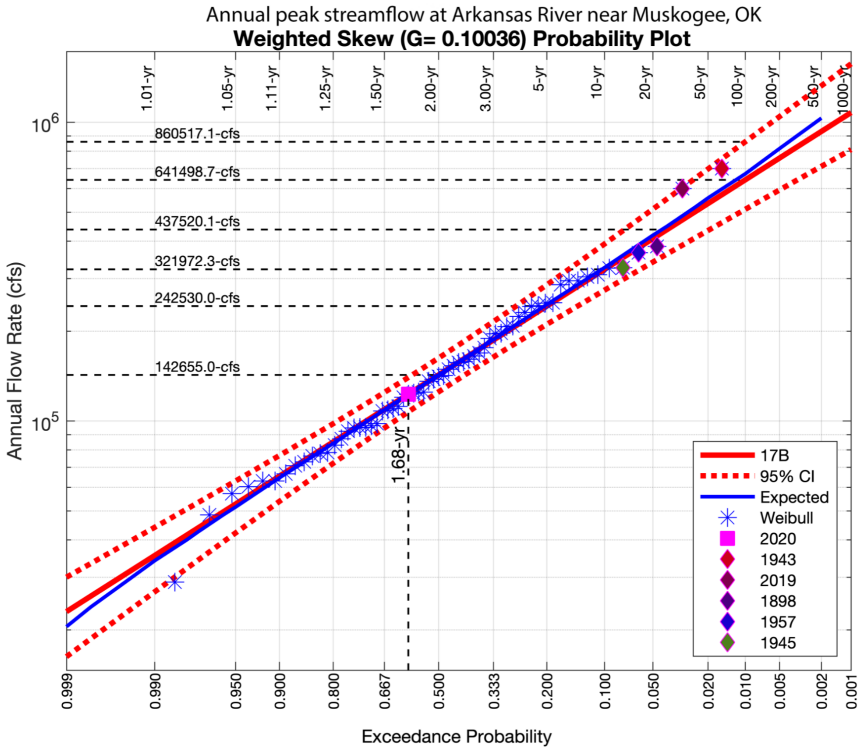

Figure 6. Bulletin 17b result for USGS 07194500 Arkansas River near Muskogee, OK

1.4.0 Discussion and Conclusion

My estimate is that this particular event in Oklahoma has a return period of about 500 years. This is based on the full record of annual peak flows from the USGS gauge station at USGS 07165570 Arkansas River near Haskell, OK.

The return period calculations based on the precipitation were awfully low, especially compared to the peak flows. But it is important to note that one single precipitation gauge is unlikely to represent the precipitation across an entire watershed, and we cannot be sure exactly how to define the single precipitation 'event'.

If I had more time I would have liked to do the following:

- Explore the flood extent image (flood_extent_sentinel2_hydroassessment.tif) to see what flood zone types (tulsaOK_flood_hazard_zones_simple.geojson) are intersected by water (pixel value of 1). This would give an idea of the return period based on the flood extent. This is a very easy task for me using ArcGIS, but unfortunately my personal computer does not have this software, and I could not learn QGIS in a reasonable amount of time for this task.
- Write Python code to calculate return periods from Bulletin 17c. There is official USGS software for this calculation, which is actually quite complex with many logical steps, but that software is only available on Windows computers.
- Train a Generative Adversarial Network (GAN to predict a particular stream's return period from sub-annual peak flows and remote sensing images. I believe there is a lot of lost information when we use only annual peaks to do these analyses. A GAN can extract useful information and use those to estimate a suitable probability distribution, which I believe could be more realistic than our current attempts to fit peak flows to a pre-defined distribution.

2 Understanding flood risk from observations versus models

2.1 How would you integrate satellite observations of flood frequency with modeled flood return periods?

Integrating a satellite image with a modelled flood is an exercise in data assimilation. We know that uncertainties exist in the satellite image as well as the model. One promising way to integrate the two is with a machine learning - based - dynamic state update (Pelissier et al., 2020). In this method we would train a machine learning model to ingest both the satellite image and the model states, and to make *dynamic state corrections* one time step at a time to the model. The one step at a time method here allows the full model state space to respond to the new information, creating not just a corrected model, but a new dynamics model that contains information from the satellite image, essentially data assimilation.

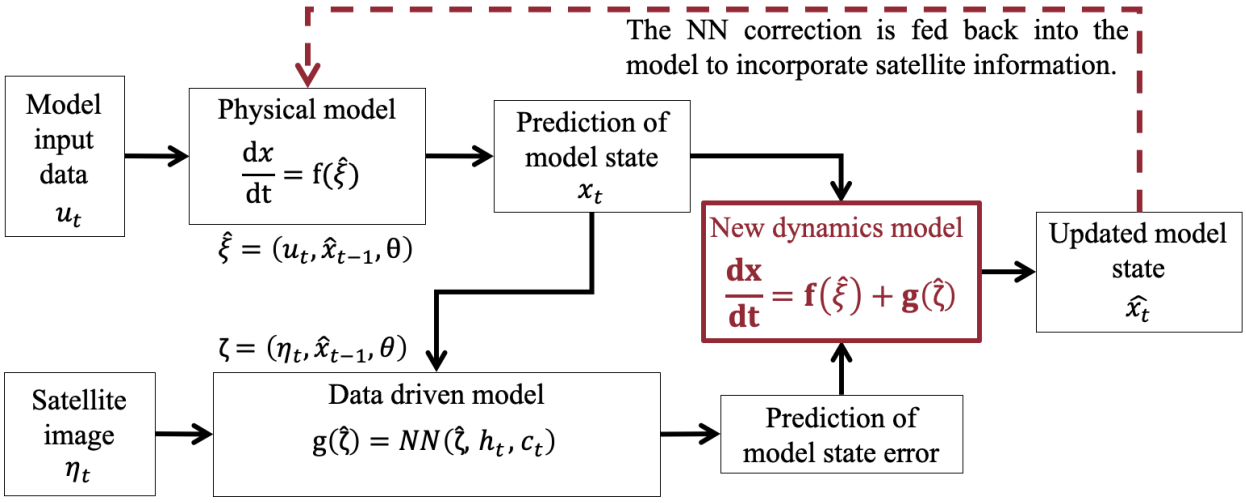

Figure 7. Machine learning model used for dynamic state update of a physical model based on new data. The Neural Network would be trained to ingest satellite data and the state of the physically-based model to produce a state correction which can be used immediately by the physically-based model.

2.2 What would be the best integrated product for a flood manager to understand spatial flood risk in their region using all data available under a changing climate, infrastructure, and land uses?

A valuable product for a floodplain manager to use for understanding spatial flood risks would be a web-application which has flood extents that change with toggle switches, including climate change scenarios, land use/land cover scenarios, precipitation events, etc. So we can start with a 'baseline' flood map, and if we wanted to see the flood map under IPCC scenarios with a post-forest-fire scenario. We can do this with entity aware deep learning by including climate indices (Nearing *et al.*, 2019) and evolving attributes summarized from satellite data in the training data. For instance, if the training set includes basins from many different climatological

conditions, from the entire set of training basins, we can modify these during the predictions to produce alternative scenarios (Figure 1).

LSTM models would be appropriate for this product, because they make good predictions in extreme events (Frame et atl, 2021a) making them good for flood predictions and they can take many different inputs (Frame et al., 2021b) making them good for utilizing data from Satellites.

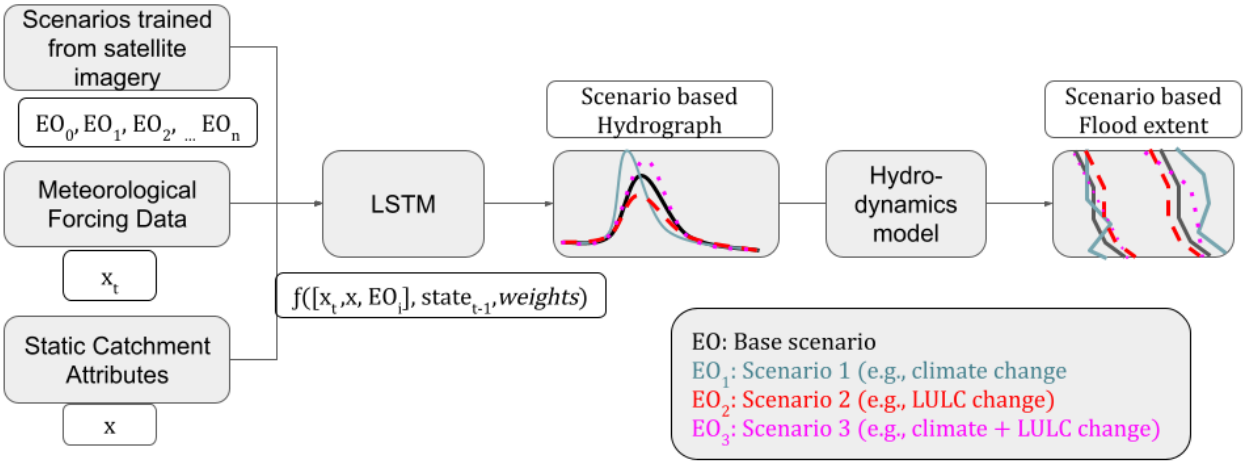

Figure 8. Flowchart of a product for floodplain managers. Using deep learning (Long Short-Term Memory Networks) to analyze hydrologic scenarios with modified inputs based on remote sensing observations.

References:

Jeff Burkey (2009). Log-Pearson Flood Flow Frequency using USGS 17B (https://www.mathworks.com/matlabcentral/fileexchange/22628-log-pearson-flood-flow-frequency-using-usgs-17b), MATLAB Central File Exchange. Retrieved August 25, 2021.

Frame, J.M., Kratzert, F., Klotz, D., Gauch, M., Shelev, G., Gilon, O., Qualls, L. M., Gupta, H. V., and Nearing, G. S. 2021a Deep learning rainfall-runoff predictions of extreme events, Hydrol. Earth Syst. Sci. Discuss. [preprint], https://doi.org/10.5194/hess-2021-423, in review.

Frame *J.M.,* Kratzert F., Raney A., Rahman M., Salas, F., Nearing G.S. 2021b, "Post-processing the National Water Model with Long Short-Term Memory Networks for Streamflow Predictions and Model Diagnostics". Accepted for publication in the Journal of the American Water Resources Association. https://eartharxiv.org/repository/view/124/

IACWD: Inter-Agency Advisory Committee on Water Data. 1982. Guidelines for Determining Flood Flow Frequency: Bulletin 17B, Tech. rep., Washington, D.C.

Pellisier, Frame and Nearing, 2020, "Combining Parametric Land Surface Models with Machine Learning". 2020 IEEE International Geoscience and Remote Sensing Symposium. https://arxiv.org/abs/2002.06141

Nearing G.S., Pelissier C., Kratzert F., Klotz D., Gupta H.V., Frame J.M., *Sampson A.K.* 2019. "Physically

Informed Machine Learning for Hydrological Modeling Under Climate Nonstationarity". Science and Technology Infusion Climate Bulletin. NOAA's National Weather Service. *44th NOAA Annual Climate Diagnostics and Prediction Workshop Durham, NC.*
https://www.nws.noaa.gov/ost/climate/STIP/44CDPW/44cdpw-GNearing.pdf