Professor Ditzler,

I am not an engineer, nor a mathematician. I am hoping to do something relevant to my research (see Abstract) and within my field of hydrology. I really struggle with the mathematical concepts in this course, so I suspect you might find this proposal lacking there!

Thanks, Quinn

Title: A Machine Learning Approach to Streamflow Prediction in the Upper Colorado River Basin

**Abstract:**

Due to their infrequency, the impacts of low-probability hydrologic conditions, such as extreme snowpack and drought, on watershed behavior can be difficult to assess in purely data-driven approaches. Large-scale process-based models can be used to simulate high-resolution data that capture hydrologic dynamics subject to low-probability events; however, large computational demands and complex model construction often limit the usability of these process-based simulations for stakeholders and decision makers. For that reason, this study presents a Machine Learning (ML) approach to learn streamflow dynamics from a large-scale process-based simulation in the Upper Colorado River Basin. Our goal is to create faster and more-interpretable tools that can still capture low frequency events. Long Short-Term Memory (LSTM) statistical emulators are used to predict daily, monthly, bi-annual, and long-term streamflow across a range of high- and low-frequency hydrologic conditions where observations may be too sparse to otherwise make accurate predictions. *Here we explore first the ability of LSTM models to capture simulated streamflow from the process-based model and second the ability to transfer across the domain spatially.*  Our results will help bridge gaps between computationally intensive process-based simulations and purely data driven approaches.

**Approach:**

Data Source:

Variables = temperature, precipitation, hill slope 1983-2019 at Taylor River processed from [NLDAS](https://ldas.gsfc.nasa.gov/nldas) to be time-series of dimension (nxd), where n = # variables (3) and d = number of days on record. (Note, I’ve already accessed and processed this dataset)

Target = streamflow at [Taylor River at almant CO](https://waterdata.usgs.gov/nwis/uv?site_no=09110000) (Note, I’ve already accessed and processed this dataset)

This project will focus on evaluating the performance of LSTM NN on a

1. regression task - target = streamflow values
2. classification task - target = streamflow ‘state’ [< flood stage, > flood stage]

at different time intervals (daily, monthly, bi-annual)

That will include adjusting hyperparameters, neural architecture, train-test-split, and perhaps resampling the input variables.

Metrics of success:

1. Solid generalization (via RMSE) of LSTM model to test sample data at Taylor River at times outside of training sample data
2. Solid generalization (via RMSE) of LSTM model to test sample data at Taylor River at values outside of training sample data (i.e. data at ‘tail ends’ of population probability distribution)
3. Solid generalization (via RMSE) of LSTM model to test sample data at stream locations outside of Taylor River (ie at other geographic locations).

**Novelty:**

Although ML approaches like LSTM may be very familiar to other fields, they are still relatively new to hydrology and there hasn’t been too much done along these lines. This work is novel in the sense that I will be employing this technique in an geographic area where it hasn’t been done before, although my work will likely borrow from existing workflows such as [this](https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/) (a generic article on LSTM in continuous data) and [this](https://github.com/jfzhang95/LSTM-water-table-depth-prediction/) (a hydrology-specific workflow).

**Feedback points:**

1. I’m really interested in encoding ‘memory’ into our data driven models, and I’d love to learn more about this. Right now the only approach I’m familiar with that does this is LSTM.
2. I’d love to learn more about how to build generalizable LSTM models and preventing overfitting in LSTM. How does this relate to the regularization approaches we discussed in regression tasks?