Leveraging machine learning for identify hydrological extreme events under global climate change

Ying-Jung Chen Deweese
Georgia Institute of Technology
Atlanta, GA, U.S.A
ydeweese3@gatech.edu

Abstract

Hydrological extreme events, such as droughts and floods, are highly destructive natural disasters and its occurrence is expected to increase under the future climate change. Accurate and efficient approach to detect such events will provide timely information to assist management strategies for minimizing socio-economic damages. Despite the threshold approach has established to detect extreme events, the missing data from hydroclimate data and accurately identifying these events are still major challenges. The advent of machine learning models can help to identify the occurrence of droughts and floods events accurately and efficiently. Therefore, this proposed study will develop a machine learning model with semi-supervised anomaly detection approach to identify hydrological extreme events with ground-based data. As a test case, we will use 45-years record of hydroclimate data in coastal California, where was the driest region in 2012-2015, following with flash floods events. The expected results will increase communities' awareness for hydrological extreme events and enable environmental planning and resource management under climate change

1.Introduction

Droughts and floods are hydrological extreme events, and most devastating types of natural disasters under the impact from climate change. These events lead to severe socio-economic impacts in climate sensitive regions of the world. Management strategies that minimize the socio-economic impacts of floods and droughts can be more effective when we can detect the occurrence of droughts and floods accurately. Under future climate projection scenarios, the frequency of extreme droughts and floods has been increased [6]. Thus, there is an urgent need to accurately detect the occurrence of droughts and floods recently under global climate change.

Droughts and floods are different physical processes but identifying these events can be done similarly. For example, we can identify droughts and floods events using extreme value theory with threshold approaches. Floods can be detected by using the peak streamflow volume over threshold or as annual maxima events, while droughts can be detected by using hydrological events below a given threshold of deficit stream volume or annual minimum events. However, this approach is sensitive to the selection of threshold level [2] over different years and may produce detection error for extreme events. Additionally, hydroclimate data from either ground-based or satellite products tends to suffer from missing data issue. The Long Short Term-Memory (LSTM) autoendocers approach has provided the advance power for reconstructing data signals to remedy missing data issue [7].

Given that hydrological extreme events are anomalous events, machine learning (ML)-based anomaly detection can be a promising approach to identify such events. ML-based anomaly detection methods provide an automated and accurate manner, which can save more time than threshold-based approach and human manually identify extreme events. ML-based anomaly detection methods are largely classified into supervised and unsupervised learning approaches.

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2021.

Due to the dynamic of floods and droughts, there are supervised learning (support vector machine, extreme machine learning) and unsupervised learning (LSTM auencoders) approaches were used commonly to detect these extreme events [1,3]. Although these approaches can detect extreme events efficiently, there is still needs to improve the accuracy of the performances. However, a recent study that aimed to detect extreme events in water usage shows that a semi-supervised learning outlier detection can outperform the unsupervised learning approaches [5]. This suggests that floods and droughts detection could be improved further with semi-supervised learning outliner detection.

The main goal of this proposal is to develop a ML-based anomaly detection model with semi-supervised approaches to improve the extreme events detection. We will design this model in such way that a) we can easily apply this model to any other region and b) we can run this model with future climate projection scenarios. Our proposed modeling would serve as an early warning system for natural disaster response.

Contributions: We will examine the unsupervised and semi-supervised anomaly detection approaches for extreme events from ground-based hydroclimate data. We will perform a semi-supervised anomaly detection by obtaining labels from annual maxima and minima approaches. We will also compare several ML-based anomaly detections with threshold approaches to check the performance of extreme events identification.

2. Modeling Approaches

We will examine the central coastal California watersheds, where were the driest areas during the California drought (2012-2015) and followed with many flash floods events. For this, we obtained the rainfall gauges data from Santa Barbara County (https://rain.cosbpw.net/) and streamflow gauge data from USGS (https://waterdata.usgs.gov/ca/nwis/) and SBC LTER project (https://sbclter.msi.ucsb.edu/data/catalog/). Both rainfall and streamflow gauges data are with the time from 1980 to 2015. We also obtained other meteorological data such as evapotranspiration and relative humidity from CIMIS stations from 1990 to present (https://cimis.water.ca.gov/).

2.1 Machine Learning Model Approach

In this proposal, we will examine several modeling approaches for anomaly detection. One hybrid LSTM network-based anomaly detection is proposed since the decoding function can reconstruct the input feature signals with fixed data length (Figure 1). Then, using the threshold approach as a benchmark, we will compare different ML models to select the best-performing model that identifies extreme events accurately. The overall description of our proposed modeling system is shown in Figure 2.

The modeling approaches for anomaly detection used in this study are as follows:

- **Threshold approaches:** typically, the peakflow volume over threshold and the deficit flow volume below a given threshold are used to identify extreme events. This approach requires ensuring independence between events.
- **LSTM autoencoders (unsupervised anomaly detection)**: for anomaly detection, this model is trained to reconstruct signals by minimizing this objective function. When data points have high reconstruction errors, this can be treated as anomalous data based on a threshold value [4].
- **LSTM based semi-supervised anomaly detection:** this approach first retrieves a few labels as extreme events based on annual maxima and minima approaches. A LSTM network is used for reconstructing data signals into fixed data lengths. For anomaly detection, a constraint-based clustering is performed with a few labels on reconstructed data. Then, we assign an initial score based on each data position with respect to the

grouped data distribution. If there are some labels, each data's anomaly score will be updated based on nearby data's labels.

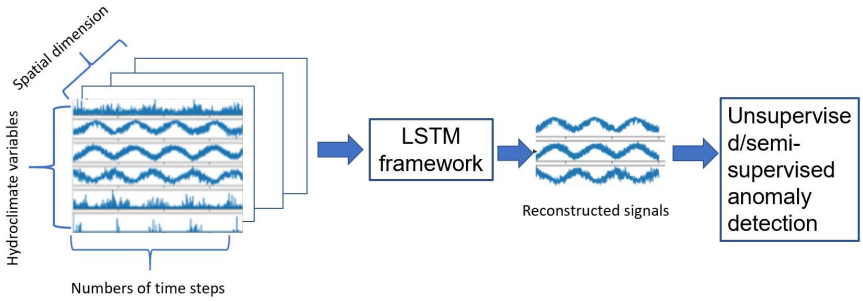

Figure 1. Architecture of the proposed hybrid machine learning model framework with two components 1) LSTM mode for temporal decoding and 2) ML based Anomaly detection.

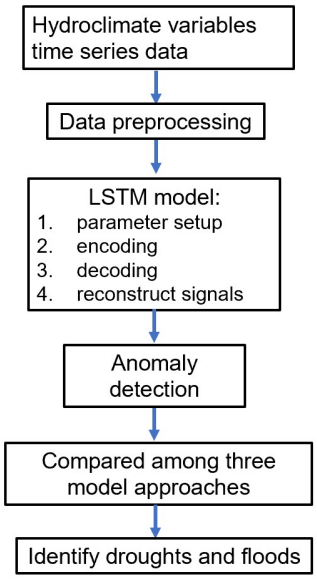

Figure 2. Flowchart of proposed study

3. Conclusion

This proposal offers a novel modeling approach for detecting hydrological extreme events by using semi-supervised anomaly detection with LSTM neural networks. Our semi-supervised anomaly detection approach will be based on the real extreme event signal labels with a clustering model. This approach with a few labels represents droughts and floods can improve the anomalous data identification than unsupervised approach This proposal will establish a modeling framework that extracts the extreme events signals (droughts and floods) from long-term hydroclimate data under global climate change in an accurate and efficient manner. With the improved detection accuracy and its low computing cost, our modeling framework can be used as a decision-making tool of management strategies for environmental planners, agriculture, insurance sectors and investors. Also, communities' awareness is increased by informing the potential disaster risks from climate-fueled drought and flood.

Acknowledgement

The author would thank Dr. Yunha Lee for editing and providing feedback for this project proposal.

References

[1] Allen-Dumas, M. R., Xu, H., Kurte, K. R., & Rastogi, D. (2020). Toward Urban Water Security: Broadening the Use of Machine Learning Methods for Mitigating Urban Water Hazards. *Frontiers in Water*, *2*, 75.

[2] Brunner, M. I., Slater, L., Tallaksen, L. M., & Clark, M. (2021). Challenges in modeling and predicting floods and droughts: A review. *Wiley Interdisciplinary Reviews: Water*, *8*(3), e1520.

[3] Deo R C and Sahin M (2015) Application of the extreme learning machine algorithm for the prediction of monthly Effective Drought Index in eastern Australia Atmos. Res. 153 512–25.

[4] Ergen, T., & Kozat, S. S. (2019). Unsupervised anomaly detection with LSTM neural networks. *IEEE transactions on neural networks and learning systems*, *31*(8), 3127-3141.

[5] Vercruyssen, V., Meert, W., Verbruggen, G., Maes, K., Baumer, R., & Davis, J. (2018). Semi-supervised anomaly detection with an application to water analytics. In 2018 ieee international conference on data mining (icdm) (Vol. 2018, pp. 527-536). IEEE.

[6] Zhao, Y., Weng, Z., Chen, H., & Yang, J. (2020). Analysis of the Evolution of Drought, Flood, and Drought-Flood Abrupt Alternation Events under Climate Change Using the Daily SWAP Index. *Water*, *12*(7), 1969.

[7] Zhang, J., & Yin, P. (2019, November). Multivariate time series missing data imputation using recurrent denoising autoencoder. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 760-764). IEEE.