# Towards Long-Lead Flood Prediction: Discovering The Spatiotemporal Co-occurrence Patterns of Extreme Precipitation Clusters

Chung-Hsien Yu
University of Massachusetts Boston
100 Morrissey Blvd.
Boston, MA 02125
csyu@cs.umb.edu

Dong Luo
University of Massachusetts Boston
100 Morrissey Blvd.
Boston, MA 02125
dongluo@gmail.com

Wei Ding
University of Massachusetts Boston
100 Morrissey Blvd.
Boston, MA 02125
ding@cs.umb.edu

## ABSTRACT

## Categories and Subject Descriptors
H.2.8 [**Database Applications**]: Data mining

## General Terms
Algorithms

## Keywords
Flood Prediction, Water Cluster, Spatial-Temporal Cluster

## 1. INTRODUCTION
Due to the chaotic nature of atmospheric circulation, it is still a challenge for current scientist to accurately predict extreme weather phenomena such as hurricane, tornado, or severe flood. Traditionally, building a atmospheric models to simulate atmospheric circulation in the near future is the most popular and common way of predicting weather.[1] However, there are limitations for these weather forecasting models which use the deterministic method of simulation because the accuracy drop dramatically when simulating the longer lead-time.[3] Therefore, the acceptable prediction range of this approach is within five days.

Most recently, data mining techniques have been widely adapted to study extreme weather phenomena and deeply understand the formation and correlated factors of these phenomena. [2] [6] [4] [7] Furthermore, the forecast systems based on data mining framework has the potential of delivering long-lead weather prediction, such as the study done by Wang et al. [7]. In Wang's research, precipitation "blocking" is assumed to be the trigger of severe flood. In other words, if the accumulated precipitation within certain time period at certain location passes the unusual high level, this atmospheric regime is considered as blocking. Therefore, instead of doing the daily weather forecast, data mining approach can be applied to predict precipitation blocking which has the high risk of resulting extreme flood.

In our research, we also target our long term goal on predicting flood event in long range of time. We start with the new definition of the atmospheric regime of blocking regarding to the precipitation. Rather than using the fixed length for every temporal blocks proposed in [7], our proposed method identify the blocks, or called cluster in our case, with different length of time. The clusters captured by our method has the better representation of true nature of precipitation blocking. When given a certain threshold, we then be able to separate the extreme precipitation cluster (EPC), which has the higher potential in causing flood than other normal precipitation cluster (PC). This way, we can focus on investigating these clusters to find the conditions under which the PCs turn into EPCs by extracting the most correlated data for being used in the data mining processes. Besides, this clustering process also eliminated the issue of imbalanced data while using the entire atmospheric data in finding the extremes.

Additionally, the atmospheric features are notoriously complex while using them to build a forecast model or applying data mining techniques. Therefore, selecting the most effective and conductive features to be used in forecast system is also a difficult task. Traditionally, the job of choosing the features from atmospheric variables for simulating atmospheric circulation heavily rely on the domain experts with decades of observation and field experiments. [3] However, even with the understanding of the formation of certain atmospheric regime, the scientist still have hard time finding the initial stage or location to start the modeling simulation. For example, the formation of the tropical cyclone always starts with an "eye" at certain location, but it is nearly impossible to predict when and where there will be an eye of cyclone one week in advance. On the other hand, in order to predict whether a atmospheric regime will occur at certain location, the modeling need to run the simulation starts from every possible initial states or locations which is computational expensive. [5]

We propose using "Spatiotemporal Co-occurrence Patterns" to predict extreme precipitation clusters which is a very efficient way to located the initial state or the precursors. Our main idea is that the extreme precipitation cluster is contributed by the perceptible water clusters (PWCs) since the rain falls must come from the perceptible water retained in

the atmosphere. Thus, we assume that there are relationships between one EPC and certain PWCs and there are patterns of how the PWCs transfer to EPC. The same method and parameters used for finding EPC is used to identify the PWCs so it can be consistence. For example, let's assume that there is a PWC formed at Gulf Coast today. Under certain circumstances, this PWC always move to Iowa one week later and then start drop heavy rain for weeks. This will form a EPC at Iowa and might eventually cause a flood event. Therefore, this Spatiotemporal Co-occurrence Pattern is described as following: "If there is a PWC formed at Gulf Coast, under certain conditions, there will be an EPC occurrence at Iowa one week later." With this assumption, it is possible to predict the heavy rainfalls at Iowa one week ahead.

In order to evaluate correlation between the appearances of PWCs and EPCs, we then next define support and confidence measurements based on the temporal co-occurrence between one EPC and PWCs. By evaluating the support and confidence, we are able to chose the most correlated locations for further investigation on the "certain conditions" which might causing the transformation of PWCs to EPC through data mining techniques. As a result, the searching space of initial states can be reduced by more than half with our approach.

To evaluate our proposed methods and approaches, we use 20-year worth of historical atmosphere data of north hemisphere to predict the EPCs at Iowa in our experiments. Our results shows that not only our proposing methods of predicting EPC is more efficient than the methods using daily-based forecasting but also is able to do long range prediction with about 80% on accuracy.

Overall, our contributions of this research paper are listed as follows:

- We proposed a novel method in identifying the atmospheric regime of precipitation blocking and perceptible water blocking. With our definition of blocking as cluster, we are able to distinguish the extreme precipitation cluster (EPC) from the regular precipitation cluster (PC).
- By focusing on EPCs and PCs, we are able to use most correlated data extracted from the huge atmospheric data sets for forecasting. This also eliminate the problem of imbalanced data while using the entire data set to predict the extreme.
- In finding the relationship between EPCs and PWCs, we further propose the concept of **Spatiotemporal Co-occurrence Patterns** which is identified by the degree of association between the EPCs and PWCs in space and time. This degree of association is measured by our new invented support and confidence score. With these measurement, we are able to reduce the feature space for predicting the weather extreme events by more the half via pruning the irrelevant locations.
- We evaluate our approach by applying it on the real world atmosphere data set. The results show that our method identifies the EPCs which resulting the flooding at state of Iowa but also efficiently predicts future EPCs in the long term future (7 to 15 days ahead) with about 80% accuracy.

The rest of this paper is organized as follows. The related works are discussed in Section 2. In Section 3, we introduce our definition of the precipitation cluster (PC) and the extreme precipitation cluster. With the definitions of the PC, we then apply the same cluster idea in defining PWC. Next, we propose Spatiotemporal Co-occurrence Pattern to describe the association between EPC and PWC in Section 4, including the definition of our proposed support and confidence measurements. In Section 5, we apply our approach to evaluate and prove our concept through our designed experiments using the real world data set. The results and conclusions are discussed in Section 6.

# 2. RELATED WORK

# 3. EXTREME PRECIPITATION CLUSTER

From the basic understanding, the extreme flooding is always caused by the torrential rain and this type of torrential rain always last several days. Therefore, if we can identify the abrupt increase precipitation during certain amount of time and treat this period of time as a "block" or "cluster". Then, this cluster can be used to indicate the potential of future flood event. Therefore, we introduce a new definition of **Extreme Precipitation Cluster** to describe this phenomena.

DEFINITION 1. ***Precipitation Cluster(PC):*** *A PC is a time series data , $p_1, p_2, \ldots, p_n$, consisting of n precipitation data at certain location. The precipitation data right before the start and right after the end of a PC, $p_0$ and $p_{n+1}$, must be less than a low-bound threshold $\theta$ and every precipitation data included in this PC must be greater than $\theta$. In addition, $n \geq \pi$, where $\pi$ is a user-defined threshold used to set the minimal length of a PC.*

As a result, a PC can be used to represent a contiguous rainfalls during a certain period of time at one certain location. With this definition, there is no overlapping PCs over the searching space.

DEFINITION 2. ***Extreme Precipitation Cluster(EPC):*** *An EPC is also a PC. If the average precipitation of a PC is greater than a threshold $\alpha$. This PC is defined as an EPC.*

Thus, an EPC represents the extreme condition of abrupt increase in rainfalls during a certain period of time.

## 3.1 Searching for Precipitation Cluster

# 4. SPATIOTEMPORAL CO-OCCURRENCE PATTERNS

## 4.1 Support and Confidence

# 5. CASE STUDY: STATE OF IOWA

## 5.1 Identifying Extreme Precipitation Cluster

## 5.2 Co-Occurrence Locations

## 5.3 Predicting Future Extreme Precipitation Clusters

# 6. CONCLUSION

# 7. REFERENCES

[1] H. Cloke and F. Pappenberger. Ensemble flood forecasting: a review. *Journal of Hydrology*, 375(3):613–626, 2009.

[2] X. Li, B. Plale, N. Vijayakumar, R. Ramachandran, S. Graves, and H. Conover. Real-time storm detection and weather forecast activation through data mining and events processing. *Earth Science Informatics*, 1(2):49–57, 2008.

[3] J. Lubchenco and T. R. Karl. Predicting and managing extreme weather events. *Print edition*, 65(3):31–37, 2012.

[4] A. McGovern, D. John Gagne, N. Troutman, R. A. Brown, J. Basara, and J. K. Williams. Using spatiotemporal relational random forests to improve our understanding of severe weather processes. *Statistical Analysis and Data Mining*, 4(4):407–429, 2011.

[5] D. J. Stensrud, J.-W. Bao, and T. T. Warner. Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Monthly Weather Review*, 128(7):2077–2107, 2000.

[6] T. A. Supinie, A. McGovern, J. Williams, and J. Abernathy. Spatiotemporal relational random forests. In *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*, pages 630–635. IEEE, 2009.

[7] D. Wang, W. Ding, K. Yu, X. Wu, P. Chen, D. L. Small, and S. Islam. Towards long-lead forecasting of extreme flood events: A data mining framework for precipitation cluster precursors identification. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 1285–1293, New York, NY, USA, 2013. ACM.