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

Chung-Hsien Yu
Department of Computer Science
University of Massachusetts Boston
csyu@cs.umb.edu

Dong Luo
Department of Computer Science
University of Massachusetts Boston
dongluo@gmail.com

Wei Ding
Department of Computer Science
University of Massachusetts Boston
ding@cs.umb.edu

David L. Small
Department of Civil and Environmental Engineering
Tufts University
David.Small@tufts.edu

Shafiqul Islam
Department of Civil and Environmental Engineering
Tufts University
Shafiqul.Islam@tufts.edu

ABSTRACT

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

General Terms
Algorithms

Keywords
Flood Prediction, Precipitation Cluster, Spatiotemporal Pattern

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.[4] Therefore, the acceptable prediction range of this approach is within five days.

Most recently, data mining techniques have been widely adopted to study extreme weather phenomena and deeply understand the formation and correlated factors of these phenomena. [3] [7] [5] [8] 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. [8]. In Wang's research, precipitation "blocking" is assumed to be the trigger of severe flood. In other words, if the accumulated precipitation within a certain time period at a 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 events 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 [8], our proposed method identify the blocks called cluster in our case with different length of time. The clusters captured by this method has the better representation of true nature of precipitation blocking. When given a certain threshold, we are be able to separate the extreme precipitation cluster (EPC), which has the higher potential in causing flood than other normal precipitation clusters (PC). By doing this method, 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. This clustering process is also eliminated the issue of imbalanced data while the entire atmospheric data are used as finding the extremes.

Additionally, the atmospheric features are notoriously complex while used to build a forecast model or applied to data mining techniques. Therefore, selecting the most effective and conductive features to be used in forecast system is a difficult task. Traditionally, the job of choosing the features from atmospheric variables for simulating atmospheric circulation heavily relies on the domain experts with decades of observations and field experiments. [4] However, even with the understanding of the formation of certain atmospheric regime, the scientists still have difficulty finding the initial stage or location to start the modeling simulation. The formation of the tropical cyclone, for example, always starts with an "eye" at a certain location, but predicting when and

where will be an eye of cyclone one week in advance is nearly impossible. On the other hand, in order to predict whether a atmospheric regime will occur at a certain location, the modeling needs to run the simulation starting from every possible initial states or locations which is computational expensive. [6]

We propose using "**Spatiotemporal Co-occurrence Patterns (STCP)**" to predict extreme precipitation clusters which is a very efficient way to locate the initial state or the precursors. Our main assumption is that the extreme precipitation cluster(EPC) is contributed by the **Precipitable Water Clusters** (PWCs) because the rain falls must come down from the precipitable water retained in the atmosphere. Using the same idea of finding EPC and then applying it on precipitable water data, the PWCs can be identified. Since relationships between one EPC and PWCs are assumed, and there exist patterns of how the PWCs transfer to EPC. For instance, let's assume that when there is a PWC formed at Gulf Coast today, under certain circumstances, this PWC always moves to Iowa one week later and then start dropping heavy rain for weeks. This rain fall will form a EPC at Iowa and could 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 pattern in mind,the heavy rainfall is predictable 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, the most correlated locations will be selected for further investigation on the so-called "certain conditions" which could be causing the transformation of PWCs to EPC through data mining techniques. As a result, by using this proposed method, the searching space of initial states can be reduced more than half.

To evaluate our proposed methods and approaches, 40-year worth of historical atmosphere data of northern hemisphere to predict the EPCs at Iowa was used in our experiments. The results show that not only is the proposed method of predicting EPC more efficient than the methods using daily-based forecasting, but also it is able to do a 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 precipitable water blocking. With this definition of blocking as a cluster, the extreme precipitation cluster (EPC) from the regular precipitation cluster (PC)is distinguished.
- By focusing on EPCs and PCs, we are able to use most correlated data extracted from the huge atmospheric data sets for forecasting. This way eliminates the problem of imbalanced data while the entire data set is utilized 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, the feature space for predicting the weather extreme events is reduced 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 PWCs resulting the flooding at state of Iowa and efficiently predicts future EPCs in the long term future (7 to 15 days ahead) with the 80% accuracy.

The rest of this paper is organized as follows. The related works are discussed in Section 2. In Section 3, we first 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 (STCP) 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 a 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 θ and every precipitation data included in this PC must be greater than θ. In addition, $n \geq \pi$, where π 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 a 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 high-bound threshold α. This PC is defined as an EPC.*

Thus, with the chosen of appropriate threshold α, an EPC can represent the extreme condition of abrupt increase in rainfalls during a certain period of time.

3.1 The Thresholds

Then, the next question is what are the appropriate thresholds should be chosen to identify the PCs and EPCs. Since we are try to identify the abnormal situation, the percentile measure over the entire precipitation data of the study location is used to decide the thresholds. For example, if the average precipitation of a PC is greater than the 90% percentile value of the entire precipitation data, then this PC is defined as EPC. Based on the same idea, 20% percentile can be used as the low-bound threshold θ to find PCs. Thus, the values of these two thresholds, α and θ, are between 0 and 1. Including another threshold π, the further discussion on how these thresholds should be chosen will be discussed in Section 5, .

3.2 Precipitable Water Clusters

By definition, the precipitable water measure is the total water vapor contains in the atmospheric column bottomed with a ground surface. This measurement is used to indicate the potential of rainfalls of a certain area. In other words, the precipitable water will start turning into precipitation under certain conditions such as the change of temperature up in the atmosphere. [2] Accordingly, the high amount of precipitable water will produce high amount of rain.
Therefore, the "blocking" phenomena of precipitable water at a certain location is also studies in our research. The follow is our formal definition of **Precipitable Water Cluster (PWC)**.

DEFINITION 3. ***Precipitable Water Clusters(PWC):*** *A PWC is a time series data , $w_1, w_2, \ldots, w_n$, consisting of n precipitable water data at a certain location. The precipitable water data right before the start and right after the end of a PWC, w_0 and w_{n+1}, must be less than a low-bound threshold θ and every precipitable water data included in this PWC must be greater than θ. In addition, $n \geq \pi$, where π is a user-defined threshold used to set the minimal length of a PC. Also, the average precipitable water of a PWC is greater than a high-bound threshold α.*

To be consistent, the same thresholds are used while searching PWCs as EPCs in our study.

4. SPATIOTEMPORAL CO-OCCURRENCE PATTERNS

With our definitions of EPC and PWC as well as the assumption that PWC has the high possibility of transforming to EPC, the concept of **Spatiotemporal Co-Occurrence Pattern (STCP)** is then proposed to describe this type of transformation. This main assumption is that there exist patterns of how PWC transform to EPC. This pattern of transformation progresses over the spatiotemporal space. The formal definition of a Spatiotemporal Co-Occurrence Patterns is given as follow:

DEFINITION 4. ***Spatiotemporal Co-Occurrence Pattern (STCP):*** *A STCP of a location A is a transformation pattern which describes a PWC located at location B during time period t_1 progress and then transform to an EPC at location A during time period t_2 under certain circumstances. Location B is defined as the "**Initial State**" and $t_2 - t_1$ as the "**Lead-Time**" of this STCP.*

Thus, if all STCPs of a location are identified, then the occurrences of EPCs at this location can be foreseen by detecting the occurrence of PWC at each initial state with certain lead-time ahead.
Now, the challenge is that how the STCPs of a certain location can be identified. We resolve this issue by proposing the measurements of support and confidence.

4.1 Support and Confidence

To evaluate the relationship between EPC and PWC, the two measurement, support and confidence, are defined and described as follows.

DEFINITION 5. *Given an EPC P and a PWC W, both P and W are time series with length of q and r respectively. Therefore, $P = \{t_{a+1}, t_{a+2}, \ldots, t_{a+q}\}$ and $W = \{t_{b+1}, t_{b+2}, \ldots, t_{b+r}\}$, $t_i \in T$ where $T = \{t_1, t_2, \ldots, t_s\}$. T is the collection of the entire time series space and s is the length of this time series.*
*Given a lead-time l, $\acute{P} = \{t_{a+1-l}, t_{a+2-l}, \ldots, t_{a+q-l}\}$. This means that $\acute{P}$ is a time series obtained by shifting P by l forward. Then, a measure function, denoted as **support(P,W)**, returns length($\acute{P} \cap W$) as the **support score** between P and W.*

Basically, this support score is used to indicate the possibility of whether an EPC is contributed by a certain PWC by measuring the "overlapping" length of this PWC and the shifted EPC over temporal space. For example, if there is an EPC between July 14 and July 19 and a PWC between July 2 and July 11, with the given lead-time of 7 days, the support score measured from this EPC and PWC is 5 (days), the overlapping between July 7 and July 11.
With this *support*() function, the total support score related to a target location is then defined as follow:

DEFINITION 6. *Given a location A, there are total j EPCs identified during the period of time T and they are $\{P_1, P_2, \ldots, P_j\}$. Meanwhile, the other location B has total of k PWCs identified and they are $\{W_1, W_2, \ldots, W_k\}$. Then, the total support score of location B in respect to location A is:*

$$\sum_{x=1}^{j} \sum_{y=1}^{k} support(P_x, W_y)$$

By choosing a target location, the total support scores of other locations over the study spatial space then can be obtained. The locations with the high total support scores have higher possibility of being the initial states of STCPs in respect to the chosen target location.
However, there might be the locations with high total support scores due to the long length of PWCs, not due to the transformation of PWC to EPC. Therefore, another measurement called confidence is then introduced to indicate this situation.

DEFINITION 7. *The confidence of location B is defined as:*

$$\frac{\sum_{x=1}^{j}\sum_{y=1}^{k} support(P_x, W_y)}{\sum_{y=1}^{k} length(W_y)}$$

The range of this confidence is between 0 and 1. When the confidence equals to 1, it means that every PWCs of one location always transformed to EPCs at target location after a certain lead-time.

By investigating these two measurements, the searching space for initial states of the STCPs is reduced efficiently. This reduction is done by setting thresholds on support and confidence because the higher the support and confidence the higher the possibility of a location being a initial state.

4.2 Identify The Patterns

Thus, our proposed approach not only provide a way of identifying the initial locations of spatiotemporal patterns of extreme precipitation clusters but also further apply data mining technique to extracted these patterns which are used to built a model to predict the future EPCs with lead time of more than 7 days. In our research, Decision Tree and AdaBoost are chosen as our main data mining method to learn the patterns. The advantage of using these two kind of supervised learning methods is that the feature selection feature is build-in with them. In other words, we can eliminate more of those factors and locations that are not associated with the transformation patterns.

To start this pattern learning process, we collect the historical data of those atmospheric factors, such as temperature data at certain altitude ,which might contribute to the patterns and only the data belongs to the initial locations selected by our proposed method are included. These factors are used as the features to construct an instance. In addition, the patterns we try to catch is over spatiotemporal space so the spatial and temporal dimensions are also considered when constructing an instance. This 3-dimensional feature space is consist of the atmospheric factors of different location and different time periods. For example, if there are 9 atmospheric factors chosen and there are 500 initial locations and 7 days are included, then there will be $9 \times 500 \times 7$ features included in an instance.

Since this is a supervised learning process, we need to define the class label for each instance. Our goal is to find whether certain pattens will cause EPCs after a certain period of time. When the lead-time is set as 7 days, the class label of an instance should be a positive class if EPC occurs at target location 7 days after the last day of this instance, otherwise it is a negative class. Thus, an instance with feature space between July 1 and July 7 is a positive class if there is a EPC occurred on July 14.

Through this supervised learning as well as feature selection processes, we then can obtain a spatiotemporal pattern of how a EPC is formed. The pattern learned by Decision Tree might be described as follow: " When the temperature at location A dropped to a certain degree at day 1 and then the temperature at location B increased to a certain degree at day 4 along with the continuous high wind and water vapor at location C from day 2 to day 5, then there will be a EPC on day 15 at target location".

Considering that it is possible that there are more than one spatiotemporal pattern of forming EPCs at one location, the AdaBoost method is then adopted to construct a predictive model by consolidating all the patterns learned by Decision Tree. During this "boosting" process, only the highly correlated patterns are chosen to form the model. Using this predictive model, the potential EPCs can be predicted in advance with a long lead-time or more than 7 days. Extended from the EPCs prediction, people can be alerted about the possible occurrence of extreme flood in advance.

5. CASE STUDY: STATE OF IOWA

To evaluate our approach, the 30 years of precipitation data at State of Iowa was chosen for the investigation on the EPCs occurred in Iowa. Next, the precipitable water data of northern hemisphere was used to identify the PWCs occurred at different locations during the same 40 years period of time. Then, we applied our support and confidence measurements to measure the relationship between the EPCs of Iowa and the PWCs of other locations. With this quantized measures, we then identified the potential initial states or locations of the Spatiotemporal Co-occurrence Patterns of EPCs in Iowa. Furthermore, the data mining techniques were adopted to learn the Spatiotemporal Co-occurrence Patterns from the atmospheric conditions of these identified locations.

newline These learned conditions are considered as the precursors of the occurrences of EPCs in Iowa so they can be used to build the model for predicting the EPC occurrences. Using the same approach and randomly selecting the locations from those locations which are not identified as the potential initial states, we built different models to compare with the one with initial states. The result showed that the model with the initial locations selected by our proposed support and confidence outperformed the other models built using other non-initial locations. The details of our experiments are illustrated in the following sections.

5.1 Data Preprocessing

The area average daily precipitation accumulation of Iowa between 1980 and 2010 was obtained for our experiments originally. Since our long term goal is to predict extreme flooding, the accumulated precipitations of winter seasons (from November to February) were removed to eliminate the precipitation of snowfalls. This way, the patterns of how the flooding caused by extreme rainfalls clusters can be truly caught by our proposed method.

Next, we evenly divided the northern hemisphere into $5,328$ geographic locations, latitude-wise and longitudes-wise. We then extracted historical atmospheric data of each location from the NCEP-NCAR Reanalysis dataset[**?**].

5.2 Identifying Extreme Precipitation Cluster

Based on Definition 1 and 2, the low-bound threshold θ was set to the 20% percentile value among the 30 year daily precipitation accumulation, the high-bound threshold α was set to the 90% percentile value, and the threshold for minimal length of a PC, π was set to 7 days. With this configuration, we identified 77 EPCs in Iowa during 1980 and 2010. Using the percentile values as thresholds for Identifying EPCs, we were able to find the abrupt increase in rainfalls at Iowa during a certain period of time.

The daily precipitable water of the northern hemisphere were obtained for identifying the PWCs occurred at the $5,328$ geographic locations. The same approach used for searching EPCs was adopted in searching for the PWCs. Therefore, the thresholds were also needed to begin the

search. We set π to 7 days, same as for EPCs. The percentile value concept was also used for obtaining θ and α. However, we experimented with three different sets of percentile values obtained from different scopes of precipitable water data.
The first set of thresholds are the percentile values among every locations in northern hemisphere. We called this set as global threshold because these thresholds repents the percentile of entire precipitable water

5.3 Co-Occurrence Locations

5.4 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] M. D. King, W. P. Menzel, Y. J. Kaufman, D. Tanré, B.-C. Gao, S. Platnick, S. A. Ackerman, L. A. Remer, R. Pincus, and P. A. Hubanks. Cloud and aerosol properties, precipitable water, and profiles of temperature and water vapor from modis. *Geoscience and Remote Sensing, IEEE Transactions on*, 41(2):442–458, 2003.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.