Spatio-Temporal Asynchronous Co-Occurrence Pattern for Big Climate Data towards Long-Lead Flood Prediction

Chung-Hsien Yu*, Dong Luo*, Wei Ding*, Joseph Cohen*, David Small† and Shafiqul Islam†
*Department of Computer Science
University of Massachusetts Boston, Boston, MA 02125
Email: {csyu, dluo, ding, joecohen}@cs.umb.edu
†Department of Civil and Environmental Engineering
Tufts University, Medford, MA 02155
Email: {David.Small, Shafiqul.Islam}@tufts.edu

***Abstract*—Recent research efforts aim at utilizing Big Climate Data to predict floods 5 to 15 days in advance. Current simulation models forecasting heavy precipitation, a major factor related with flood occurrences, are computationally expensive and limited by their error amplification. In this paper, we introduce Spatio-Temporal Asynchronous Co-Occurrence Pattern to associate heavy precipitation with dense precipitable water and explore long-lead flood prediction from machine learning perspective. Our model predicts one location's flooding risk by connecting the heavy precipitation with its preceding precipitable water through a association mining method for asynchronous co-occurrence location discovery and a spatio-temporal ensemble learning method for predictive modeling. Our framework requires less computational cost and smaller train data while being compared to other existing approaches. In addition, our framework is designed to be scalable and allows distributed computing. Our real-world case study has achieved 87% accuracy on predicting the heavy precipitations which trigger severe floods at least 9 days in advance.**

***Keywords*-Spatiotemporal Patterns; predictive modeling; Flood Prediction**

I. Introduction

Atmospheric and climate data have been intensively collected through various means and keep increasing enormously over time. Simulating atmospheric circulation using ensemble regression models is the most common way of extreme weather forecasting [1]. However, this type of simulations are computationally expensive when processing the global atmospheric data to simultaneously build the models [2]. Besides, the simulation errors are amplified after running regression over a long lead-time [3]. Therefore, scientists are still struggling to utilize this unstructured Big Climate Data [4] to predict severe floods with acceptable accuracy 5 to 15 days ahead.

The extreme precipitation accumulation has been studied by [5], [6] and is assumed to be a major trigger of floods. Since heavy precipitation come from dense precipitable water retained in the atmosphere [7], we propose a new concept, **Spatio-Temporal Asynchronous Co-Occurrence Pattern (STACOP)**, to extend the causal factors of floods from local precipitation to global precipitable water. Based on this concept, we design a temporal association mining method to identify the asynchronous co-occurrence between one location's heavy precipitation and dense precipitable water of other locations. To the best of our knowledge, we are the first team to propose a spatio-temporal modeling framework connecting flood, heavy precipitation, and dense precipitable water for long-lead flood prediction.

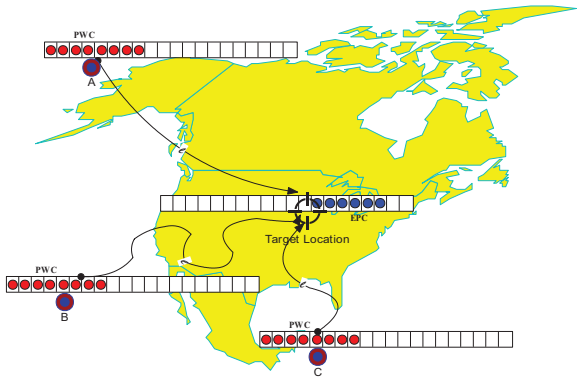

Figure 1. A Spatio-Temporal Asynchronous Co-Occurrence Pattern is an association connecting global dense precipitable water (PWC) to a target location's formation of heavy precipitation (EPC). This formation progresses over space and time and takes a transformation time of length l. The bars represent the time windows within the same periods.

We first define **Extreme Precipitation Cluster (EPC)** as a time window to represent the heavy precipitation accumulated within this window at a location. To ensure the EPCs identified by our search algorithm having the highest potential of triggering floods at one location, we utilize this location's historical severe flooding events to approximate this location's water-holding capacity [8] with three thresholds. We further define **Precipitable Water Cluster (PWC)** as a time window to represent dense precipitable water measured within this window at a location. Our key assumption is that any EPC is contributed by the PWCs occurring at certain co-occurrence locations some time earlier [5]. Using Figure 1 as an example, if there are PWCs at location A, B, and C today

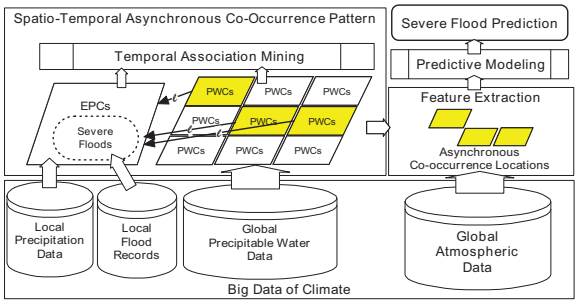

Figure 2. The proposed framework for long-lead flood prediction.

and these PWCs are moving toward the target location under certain circumstances, one week later, heavy rainfall starts to drop at this target location for days and to form an EPC due to the approaches of these PWCs. This scenario is considered as an STACOP of this target location's flood occurrences. A, B, and C are called the Asynchronous Co-occurrence Locations of this STACOP. The asynchronous co-occurrence locations are identified by our temporal association mining method which extracts the most associated spatial features before the predictive modeling. Therefore, we reduce the training data size and remove the irrelevant and redundant factors from the entire spatio-temporal domain. We further adopt a temporal cascading voting approach to assemble the Decision Tree models learned from different lead-time and build the final predictive model. An outlook of our framework is shown in Figure 2.

We have conducted a case study to predict Iowa's sever flooding events using thirty years of historical atmospheric and daily precipitation data of Iowa. Our association mining method identifies the asynchronous co-occurrence locations associated with the EPC occurrences in Iowa. The results obtained from the models using the features extracted from asynchronous co-occurrence locations have the advantages on computational cost and prediction performance over the models built on the entire spatial space. Overall, our contributions are:

- We are the first to utilize the historical severe flooding events to approximate the water-holding capacity for identifying the EPCs with the highest potential of triggering floods.
- We introduce **Spatio-Temporal Asynchronous Co-Occurrence Pattern** to associate heavy precipitation to dense precipitable water and then develop a temporal association mining method to identify the spatial associations between one target location and its asynchronous co-occurrence locations. Using features extracted from these asynchronous co-occurrence locations, we dramatically reduce the training data size for modeling.
- Our framework fits the Map-Reduce-style parallelization and is scalable.
- Our real-world case study has shown that our ensemble model requires less computational cost and has achieved 87% accuracy while predicting severe floods at least 9 days in advance.

The rest of this paper is organized as follows. In Section II-A, we define EPC and discuss how to approximate the water-holding capacity. Next, we define PCW and introduce STACOP to associate EPCs and PWCs with a temporal association mining method to identify the asynchronous co-occurrence locations in Section II-B. We then propose a spatio-temporal predictive model for flood prediction in Section II-C. In Section III, we evaluate our proposed framework through a case study using the real-world data sets. The related works are then discussed in Section IV. Finally, our study is concluded in Section V.

II. The Proposed Framework

To connect extreme precipitation and dense precipitable water for long-lead prediction on the flood triggered by heavy rainfalls, our framework includes three major tasks, heavy precipitation discovery, asynchronous co-occurrence location identification, and predictive modeling.

A. Extreme Precipitation Cluster Discovery

Flooding is often produced by the abnormal increase in precipitation within certain period of time. We consider the heavy precipitation within a time window as an **Extreme Precipitation Cluster**.

Definition 1: An **Extreme Precipitation Cluster (EPC)** is a time series data consisting of n precipitation measurements, $p_1, p_2, \ldots, p_n$, collected at a certain location, where $n \geq \pi$, $\forall p_i \geq \theta_p : i = 1, 2, \ldots, n$, and $\frac{\sum_{i=1}^{n} p_i}{n} > \alpha_p$.

By our design, an EPC can be used to represent the true nature of a temporal blocking of different time window lengths. The α_p and θ_p can be referred to as the volume thresholds and π can be referred to as the accumulation speed threshold. Different locations have different water-holding capacities for precipitations [8]. In other words, when the precipitation level exceed one location's water-holding capacity, flood is very likely to occur at this location. Therefore, the accumulation speed and volume of precipitation are two main indicators of flood occurrences. We let α and θ be the percentile between 0% to 100% and α_p be the α^{th} percentile value and θ_p be the θ^{th} percentile value among a collection of precipitation measurements. To approximate one location's water-holding capacity, we propose using this location's historical severe flooding events to validate α, θ and π. If a severe flood is covered by an EPC, this flood's starting time is within or soon after the time window of this EPC. We then define **Positive Extreme Precipitation Cluster (P-EPC)** as an EPC which covers a severe flooding event. From each setting of α, θ, and π, we can identify a set of EPCs including a subset of P-EPCs. We exam all the settings and find those settings with which

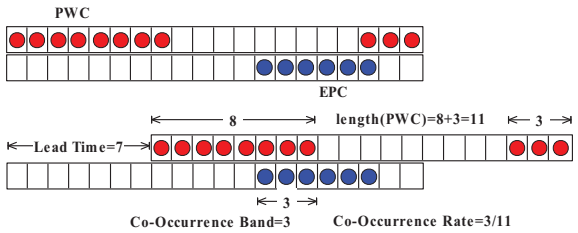

Figure 3. Co-Occurrence Band and Co-Occurrence Rate of a location are calculated based on the overlap between one time-shifted PWC at this location and one EPC at the target location.

our search algorithm identifies a set of P-EPCs covering every historical severe flooding. We choose the setting with the highest P-EPC rate, the total number of P-EPCs divided by the total number of EPCs, as the best configuration to identify the EPCs with high possibility of causing floods.

B. Spatio-Temporal Asynchronous Co-Occurrence Pattern

Heavy precipitation comes from dense precipitable water [7]. The precipitable water is the total water vapor contains in an atmospheric column bottomed at the ground surface and will start to turn into precipitation under certain conditions, such as the changes of temperature [9]. Similar to EPC, we define **Precipitable Water Cluster (PWC)** to represent dense precipitable water.

Definition 2: A **Precipitable Water Cluster (PWC)** is a time series data consisting of n precipitable water measurements, $w_1, w_2, \ldots, w_n$, at a certain location, where $n \geq \pi$, $\forall w_i \geq \theta_w : i = 1, 2, \ldots, n$, and $\frac{\sum_{i=1}^{n} w_i}{n} > \alpha_w$.

Considering the transformation from precipitable water to precipitation progressing over space and time (Figure 1), we define **Spatio-Temporal Asynchronous Co-Occurrence Pattern (STACOP)** as follows.

Definition 3: A **Spatio-Temporal Asynchronous Co-Occurrence Pattern (STACOP)** of a location G is a transformation pattern indicating the fact that the PWCs occurring at the asynchronous co-occurrence locations on time t always lead to the occurrences of EPCs on time $t+l$ at G under certain circumstances, where l is the **lead-time** of this STACOP.

To identify asynchronous co-occurrence locations, we propose a temporal association mining method with two new measures, Co-Occurrence Band (COB) and Co-Occurrence Rate (COR), used to evaluate the relationship between a target location's EPCs and a source location's PWC. We first define an $overlap()$ function to calculate the temporal association between two clusters.

Definition 4: Given an EPC $P = \{t_{a+1}, t_{a+2}, \ldots, t_{a+q}\}$ and a PWC $W = \{t_{b+1}, t_{b+2}, \ldots, t_{b+r}\}$, where a and b are the start time, and q and r are the lengths of P and W, respectively. Given a lead-time l, we have $\acute{P} = \{t_{a+1-l}, t_{a+2-l}, \ldots, t_{a+q-l}\}$, where $\acute{P}$ is obtained by shifting P by l forward. An overlap function, denoted as $overlap(P, W, l)$, returns $length(\acute{P} \cap W)$ as the temporal association between P and W.

This $overlap()$ function returns the overlapping length of a PWC and a time-shifted EPC and is used to indicate the possible contribution of this PWC to this EPC over time. Using Figure 3 as an example, after shifting the PWC by 7 forward, the overlapping between the EPC and the shifted PWC is 3. We consider this overlapping indicates the degree of contribution by the PWC to the EPC. Next, we define **Co-Occurrence Band** to measure the association of a source location with respect to a target location.

Definition 5: At a target location G, there are total j EPCs, $P_1, P_2, \ldots, P_j$. Meanwhile, at a source location S, there are total k PWCs, $W_1, W_2, \ldots, W_k$. The **Co-Occurrence Band (COB)** of location S with respect to location G with a lead-time l is calculated as:

$$COB(S, G, l) = \sum_{x=1}^{j} \sum_{y=1}^{k} overlap(P_x, W_y, l).$$

We consider the source locations having high COB with respect to a target location have high possibilities of being the asynchronous co-occurrence locations of this target location. However, certain locations having high COB is due to the long length of PWCs, not due to the transformation from PWCs to EPCs. Therefore, the locations with long PWCs consistently have dense precipitable water so the target location's EPCs always overlap with these locations' PWCs. We further introduce **Co-Occurrence Rate** to measure this situation.

Definition 6: The **Co-Occurrence Rate (COR)** of location S with k PWCs, $W_1, W_2, \ldots, W_k$, with respect to location G with a lead-time l is calculated as:

$$COR(S, G, l) = \frac{COB(S,G,l)}{\sum_{y=1}^{k} length(W_y)}.$$

A illustrative example of how to calculate COB and COR is given in Figure 3. A location with low COR means that the PWCs occurring at this location does not frequently contribute to the formations of the EPCs at a target location after a lead-time l, so this location is not considered as asynchronous co-occurrence location. Distributed computing can be applied to this asynchronous co-occurrence location discovery since the evaluation of each location is independent from each other.

C. Spatial-Temporal Predictive Modeling

With different lead-time, we obtain different sets of asynchronous co-occurrence locations to indicate the progression of EPCs over spatial space. Using the atmospheric factors collected at each set of asynchronous co-occurrence locations as the features, we train the models under different given lead-time. These models learn the conditions under which the PWCs occurring at the asynchronous co-occurrence locations are most likely to form P-EPCs which eventually trigger floods at the target location. Each model can be trained via distributed computing because

Table I
THE SETTINGS WITH THE HIGHEST P-EPC UNDER DIFFERENT π.

α	θ	π	P-EPCs	EPCs	Rate
87.8%	39.5%	3	280	910	30.7%
87.8%	39.5%	4	248	747	33.2%
84.7%	46.2%	5	270	794	**34.0%**
78.3%	46.7%	6	242	902	26.8%

the data used to train an individual model is independent from the others. We choose the Decision Tree algorithm to build these models because of its build-in feature selection and interpretable output. To consolidate these models into one ensemble predictive model, we propose a Temporal Cascading Voting Ensemble Modeling approach with the considerations of all contributive factors over spatial and temporal domains. If the majority of the Decision Tree models predict the same P-EPC occurrence at the target location, our ensemble model outputs a positive prediction of a P-EPC occurrence. A P-EPC prediction on time t means that there will be a severe flood occurring on or after time $t + \pi$, where π is the minimal length of an P-EPC. As a result, we can predict the severe floods caused by heavy precipitation at least $l+\pi$ in advance, where l is the shortest lead-time used in the model learning process.

III. CASE STUDY: LONG-LEAD FLOOD PREDICTION IN IOWA

A. Data Sources Description

The daily averages of the precipitation accumulations measured in Iowa between 1980 and 2010 are used to investigate the EPC concurrences. We also obtain Iowa's severe flooding events occurring between 1980 and 2010 from the Iowa Homeland Security and Emergency Management's website [10]. Next, we evenly divide, by 2.5 degree latitude-wise and 2.5 degree longitudes-wise, the northern hemisphere into $5,328$ geographic locations and then extract historical atmospheric data recorded at each location from the NCEP-NCAR Reanalysis dataset [11]. We select nine atmospheric factors, 1,000, 500, and 300 hectopascal (hPa) geopotential height, the temperature at 850 hPa geopotential height, the speed of zonal wind at 850 and 300 hPa geopotential height, the speed of meridional wind at 850 and 300 hPa geopotential height, and the total precipitable water in the column of the atmosphere, as the features of each location.

B. Identifying Extreme Precipitation Clusters

Three thresholds, θ, α, and π, are needed for EPC discovery. By varying θ and α from 0% to 100% and π from 3 to 14 days, we identify the EPCs in Iowa between 1980 and 2010. Using the records of the historical severe flooding events [10], we evaluate each configuration by comparing the dates of flooding events with the dates of the EPCs identified. If there is a flooding event occurring during or right after an EPC, this EPC is marked as a P-EPC and this flooding event is considered being covered by this P-EPC. For each given π, we select the maximal θ and the maximal α with which our searching algorithm finds the EPCs covering every flooding events as the best setting. We then find the best setting of θ and α under different π, which are listed in Table I. When $\pi \geq 7$, not all the flooding events are covered by the identified EPCs. The setting of $\theta = 46.2\%$, $\alpha = 84.7\%$, and $\pi = 5$ days is selected as the best configuration because it yields the highest P-EPC rate and is then used in the rest of our experiments.

C. Identifying Precipitable Water Clusters and Asynchronous Co-Occurrence Locations

Using the daily precipitable water of those $5,328$ locations, identify the PWCs of every location with the same setting of $\theta = 46.2\%$, $\alpha = 84.7\%$, and $\pi = 5$ days. However, we obtain percentile values from three different spatial scales of precipitable water data. The global percentile values are obtained from the entire precipitable water data because they represent the ranking of entire precipitable water of the northern hemisphere. The local percentile values are obtained from the precipitable water data of every individual location. As a result, each location has its own local percentile values. The third type of percentile values are called target percentile values. In our case, these values are obtained from the precipitable water data observed at those locations in Iowa. Using these three types of percentile values obtained by different spatial scales as thresholds, we identify three types of PWCs at each location.

We further identify the asynchronous co-occurrence locations with respect to Iowa by measuring the COB and COR of every location. To understand the impact of l, we identify seven sets of asynchronous co-occurrence locations with lead-time from 4 to 10 days. Using those three types of PWCs, we calculate 3×7 groups of COB and COR of every locations. From each group, we first select a subgroup of locations whose COR are higher than the average COR and then select the locations with top 500 COB of this subgroup as the asynchronous co-occurrence locations. As a result, we have identified 21 sets of asynchronous co-occurrence locations over three different spatial scales and seven different temporal distances.

D. Pattern Learning

We consider each EPC as one instance and let the start date of an EPC, t, represent an occurrence of this EPC. Next, we extract those nine atmospheric factors measured at time $t - l$ from one set of 500 asynchronous co-occurrence locations as this instance's features. Therefore, there are 9×500 features included in one instance. We generate 21 data sets from those 21 sets of the asynchronous co-occurrence locations. The number of instances in each data set is the number of EPCs identified in Iowa. Fro

Table II
THE PREDICTION ACCURACY OF THE MODELS USING ASYNCHRONOUS CO-OCCURRENCE LOCATION DATA SETS AND ALL-LOCATION DATA SETS UNDER DIFFERENT LEAD-TIME.

l	Global	Iowa	Local	All
4	0.724	0.735	**0.762**	0.698
5	**0.781**	0.709	0.735	0.716
6	0.724	0.739	0.720	**0.769**
7	0.754	0.739	**0.773**	0.735
8	0.732	0.720	0.720	**0.762**
9	0.732	0.698	**0.792**	0.764
10	0.762	**0.769**	**0.769**	0.716
Avg.	0.744	0.730	**0.753**	0.737

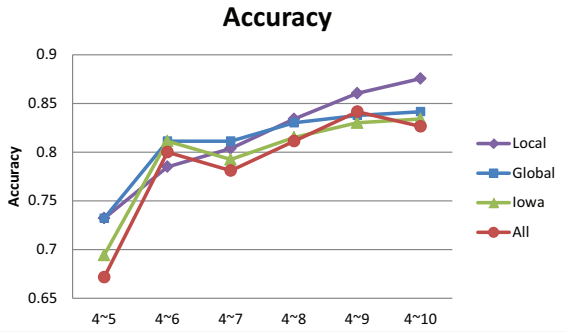

Figure 4. The accuracies of those four ensemble models built by temporal cascading voting approach.

comparison, we also generate seven data sets each of which has $9 \times 5,328$ features extracted from the $5,328$ locations and name these data sets the All-location data sets.

Using the sever flooding records [10], we identify those EPCs covering a flood event and label them as P-EPCs. We aim to learn the circumstances under which PWCs contribute to the P-EPCs. Moreover, each data set is split into a training set and a test set with 2:1 ratio. The Decision Tree model is used to train on a training set and then test on its corresponding test set. The results of the models of different spatial scales, and the results of the trained model using all-location data sets under different lead-time l are listed in Table II.

Our experiments are conducted on a High Performance Computing Cluster managed by the Research Computing Department at the University of Massachusetts Boston. The results have shown that using the features extracted from our proposed asynchronous co-occurrence locations not only yields higher or similar accuracies, but also dramatically reduces the execution time, compared using the features extracted form all locations. In addition, adopting the local percentile values performs the best among the others, because the locally identified PWCs have a better representation of abrupt increases of precipitable water at one single location than those globally identified PWCs.

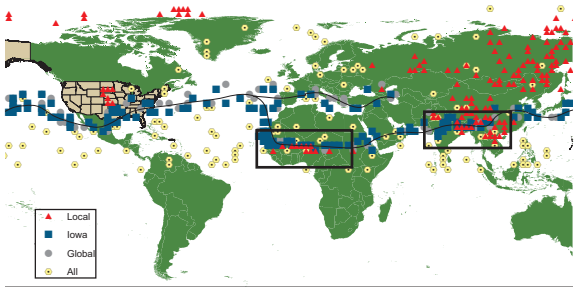

Figure 5. The visualization of four groups of locations selected by the Decision Tree models under three different spatial scales of Global, Local, and Iowa, and by the models using all locations.

E. Ensemble Modeling

We consolidate the four groups, Global, Local, Iowa, and All-location, of Decision Tree models (as shown in Table II) into four ensemble models through our proposed Temporal Cascading Voting Ensemble Modeling. In each ensemble model, every Decision Tree model has equal voting power and the majority votes is the final output of this ensemble model. We vary the number of models joined to the voting processes and choose these models based on their lead-time in chronological order starting from $l = 4$. The performances of these four ensemble models built on different series of lead-time are shown in Figure 4. We have achieved about 87.5% accuracy when predicting the P-EPCs with Local spatial scale at least 4 days in advance. Since $\pi = 5$, we are able to predict the severe floods caused by heavy precipitation at least $5+4$ days in advance with our ensemble model.

F. Selected Location Visualization

For further investigation, we collect all the features selected by the models under four different spatial scales: Global, Local, Iowa and All-location, so we obtain four collections of features. We trace back four groups of locations from which these four collections features are extracted. We then visualize these four groups of locations on the map, shown in Figure 5. With this visualization, we have uncovered the paths of moving PWCs, showing the evidence that the floods could be foreseen by monitoring the movements of PWCs along these paths. We have also observed that the locations near the southern border of the Sahara Desert and the Himalayas mountains are selected by all four type of models. This observation suggests that the atmospheric changes occurring at the largest desert and the highest mountain might have tight connections with Iowa's severe floods caused by heavy precipitation.

IV. RELATED WORK

Regression models are commonly used by most researchers to simulate atmospheric circulations for severe weather forecasting [1]. The limitation of these models is

caused by the amplified effect on simulation errors [3], [12]. As a result, these models have the acceptable prediction within five days range. In addition, most of the simulation models are very computational expensive.

Data mining techniques have been adopted to study extreme weather phenomena and to deeply understand the causal factors of these phenomena [6], [13], [14], [15], having the potential of delivering long-lead weather prediction with less data inputs. Unlike the simulation models, data mining approaches can be applied to predict the occurrences of precipitation blocking having high risks of triggering extreme flooding events [5], [6]. In [6], Wang et al. considered precipitation blocking as temporal cluster with a fixed time window size to accumulate precipitation. Unlike Wang's approach, our newly defined EPC is of various window sizes. We also incorporate the historical flooding events in finding the P-EPCs which represent the true nature of heavy precipitation with high potential of triggering floods. Furthermore, current simulation models and data mining models focus on predicting heavy precipitation without further indications what the flooding possibilities are. Instead, our proposed model is designed to predict the P-EPCs.

To achieve load-lead prediction, Wang et al. set a 15-day windows size for their clusters as the lead-time. Different from their approach, we propose STACOP to extend the lead-time by connecting heavy precipitation to PWCs as the causal factors. From feature selection point of view, other approaches include the entire feature space during the modeling process. Through our proposed STACOP, we preprocess the data to identify the asynchronous co-occurrence locations for feature extraction before building the models. As a result, we dramatically reduce the training date size and computational cost.

V. Conclusion

The goal of this study is to help the domain scientists developing early flood warning systems from machine learning perspective, which can accurately forecast extreme floods caused by heavy precipitation at least 9 days in advance at a global level. Our proposed framework is designed to be scalable and for distributed computing so it can also be applied to other spatio-temporal related big data analytics, such as droughts or hurricanes prediction, in our future work.

Acknowledgment

The authors would like to acknowledge Research Computing Department at the University of Massachusetts Boston for providing the support and the use of the supercomputing facilities. We also thank the TCL Research America for the funding.

References

[1] H. Cloke and F. Pappenberger, "Ensemble flood forecasting: a review," *Journal of Hydrology*, vol. 375, no. 3, pp. 613–626, 2009.

[2] 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*, vol. 128, no. 7, pp. 2077–2107, 2000.

[3] L. Alfieri, P. Salamon, F. Pappenberger, F. Wetterhall, and J. Thielen, "Operational early warning systems for water-related hazards in europe," *Environmental Science & Policy*, vol. 21, pp. 35–49, 2012.

[4] National Center for Atmospheric Research. (2015, Jan.) 2014 ncar annual report. [Online]. Available: http://nar.ucar.edu/2014/ncar/2014-ncar-annual-report

[5] X. Gao, C. A. Schlosser, P. Xie, E. Monier, and D. Entekhabi, "An analogue approach to identify extreme precipitation events: Evaluation and application to cmip5 climate models in the united states," MIT Joint Program, Tech. Rep., 2013.

[6] 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*, ser. KDD '13. New York, NY, USA: ACM, 2013, pp. 1285–1293. [Online]. Available: http://doi.acm.org/10.1145/2487575.2488220

[7] G. Lenderink and E. Van Meijgaard, "Increase in hourly precipitation extremes beyond expectations from temperature changes," *Nature Geoscience*, vol. 1, no. 8, pp. 511–514, 2008.

[8] K. E. Trenberth, "The impact of climate change and variability on heavy precipitation, floods, and droughts," *Encyclopedia of hydrological sciences*, vol. 17, 2012.

[9] 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*, vol. 41, no. 2, pp. 442–458, 2003.

[10] Iowa Homeland Security and Emergency Management, "Iowa disaster history," http://homelandsecurity.iowa.gov/disasters/iowa_disaster_history.html, accessed: 2015-3-21.

[11] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen *et al.*, "The ncep/ncar 40-year reanalysis project," *Bulletin of the American meteorological Society*, vol. 77, no. 3, pp. 437–471, 1996.

[12] J. Lubchenco and T. R. Karl, "Predicting and managing extreme weather events," *Physics Today*, vol. 65, no. 3, p. 31, 2012.

[13] 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*, vol. 1, no. 2, pp. 49–57, 2008.

[14] 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*. IEEE, 2009, pp. 630–635.

[15] 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*, vol. 4, no. 4, pp. 407–429, 2011.