Towards Long-lead Forecasting of Extreme Flood Events: A Data Mining Framework for Precipitation Cluster Precursors Identification

Dawei Wang
Department of Computer Science
University of Massachussets Boston
dawei.wang@umb.edu

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

Kui Yu
School of Computer Science and Information Engineering
Hefei University of Technology, China
ykui713@gmail.com

Xindong Wu
Department of Computer Science
University of Vermont
xwu@cems.uvm.edu

Ping Chen
Computer and Mathematical Sciences Department
University of Houston-Downtown
chenp@uhd.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

The development of disastrous flood forecasting techniques able to provide warnings at a long lead-time (5-15 days) is of great importance to society. Extreme Flood is usually a consequence of a sequence of precipitation events occurring over from several days to several weeks. Though precise short-term forecasting the magnitude and extent of individual precipitation event is still beyond our reach, long-term forecasting of precipitation clusters can be attempted by identifying persistent atmospheric regimes that are conducive for the precipitation clusters. However, such forecasting will suffer from overwhelming number of relevant features and high imbalance of sample sets. In this paper, we propose an integrated data mining framework for identifying the precursors to precipitation event clusters and use this information to predict extended periods of extreme precipitation and subsequent floods. We synthesize a representative feature set that describes the atmosphere motion, and apply a streaming feature selection algorithm to online identify the precipitation precursors from the enormous feature space. A hierarchical re-sampling approach is embedded in the framework to deal with the imbalance problem.

An extensive empirical study is conducted on historical precipitation and associated flood data collected in the State of Iowa. Utilizing our framework a few physically meaningful precipitation cluster precursor sets are identified from millions of features. More than 90% of extreme precipitation events are captured by the proposed prediction model using precipitation cluster precursors with a lead time of more than 5 days.

Categories and Subject Descriptors

I.5.2 [**PATTERN RECOGNITION**]: Design Methodology; I.5.4 [**PATTERN RECOGNITION**]: Applications—*Weather Forecasting*

General Terms

Experimentation, Algorithm, Performance

Keywords

Flood Forecasting, Online Streaming Feature Selection, Spatial-temporal Data Mining

*Corresponding Author

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.

1. INTRODUCTION

Recent catastrophic floods in Australia, Brazil, Pakistan, Thailand and United States call for reliable flood forecasts and long-lead times so that we can better prepare and respond to disastrous events. With the advancement of observation network and computational power, we can now provide skilled short-term (1-5 days) weather forecasts. Long-

lead flood forecasting, on the other hand, works in a dramatically increased resolution of spatial grids, length of forecast intervals, number of variables and parameters related to physical processes, parameterizations, and interactions. Existing atmospheric models, relying on simple nonlinear deterministic systems, cannot deal with such a huge feature space to provide accurate long-range (5-15 days) predictability of weather [13]. Existing operational flood forecasting systems usually rely on precipitation inputs from observation networks (rain gauges) and radar. Because of the limitation on the predictability of individual weather events [12], such precipitation inputs limit the flood forecasting to only several days.

On the other hand,we have observed that extreme floods are consequences of long sequence of heavy precipitation events occurring over extended periods. Certain atmospheric regimes (e.g., blocking [16, 18]) can lead to sequence of precipitation events over periods of several days to several weeks. For example, in Pakistan, four very large precipitation events occurred during July, 2010, and the last of which triggered the record flood that started around July 28th (Figure 1a). During this period of heavy precipitation, there is a clear signature of blocking upstream over Russia (Figure 1b). The atmospheric regime behavior (i.e. blocking) may be more predictable than day-to-day precipitation events and the prediction of such regimes can lead to the possibility of long-lead (i.e. 5-15 days) extreme flood forecasting.

Data mining techniques have great potential to identify the precursors to the atmospheric regimes that typically lead to the occurrence of flood events as a consequence of series of precipitation events. There are two major challenges when transforming from the day-to-day forecasting to long-lead prediction:

- Curse of dimensionality. The feature space for predicting the atmospheric regimes is huge and complex. Atmospheric regimes usually cover large geographic areas and last up to months. The number of features contributing to the regimes, from the Cartesian product between the spatial and temporal domains, is enormous. Also, quasi-geostrophic theory [7] demonstrates that the development of storm events requires a coupling between upper and lower levels of the atmosphere, the atmospheric variables related to certain regimes vary both horizontally and vertically and have complex relationships between each other. Even the most efficient Monte Carlo methods will suffer from computational infeasibility for such high dimensional and complex data sets [14].
- Extremely imbalanced data. Precipitation clusters that can trigger extreme floods are rare. They happen once per year or once per several years in a certain region.

In this paper, we develop an integrated data mining framework to efficiently deal with complex, high-dimensional, imbalanced atmospheric data to forecast precipitation clusters. The key components are two-fold:

- Identify the precursors of precipitation clusters that are conducive for flooding.
- Predict extended periods of extreme precipitation and the resulting floods with a lead time of more than 5 days.

Specifically, we first construct a feature space to comprehensively represent relevant spatial and temporal atmospheric variables. To untangle the problem of identifying the precursors of precipitation clusters from extremely high-dimensional data , we preform online feature selection using the Fast Online Streaming Feature Selection (Fast-OSFS) algorithm to process one feature at one time. The algorithm dynamically selects strongly relevant and non-redundant features on the fly, and is ideal in dealing with huge feature spaces with high efficiency and effectiveness. To deal with the imbalance issue, we design a hierarchical re-sampling approach for the feature selection and prediction process. Finally, by building the "most-correlated" datasets we train classification models to predict precipitation clusters on the evaluation set.

In summary, the contributions of this paper include:

- We develop an integrated data mining framework to provide long-lead extreme flood forecasting through the identification of precursors of precipitation clusters, which bridges the gap between the climatology community and the data mining community by utilizing state of art knowledge and methods from both sides.
- We extend the lead time of flood forecasting from few days (up to 5 days) to a longer term (5-15 days).
- We build the prediction models using the most-correlated data.
- We apply our framework to study historical precipitation and associated flood data in the State of Iowa. The results validate the effectiveness of the framework in accurately forecasting precipitation clusters that are conducive for extreme flooding.

The rest of this paper is organized as follow. In Section 2 we review the related works on flood forecasting and data mining techniques. An overview of our forecasting framework is presented in Section 3. Section 4 describes the data preprocessing approaches. Section 5 lists the precursor identification process including the Fast-OSFS algorithm and the hierarchical re-sampling approach. We show our experimental results in section 6 and conclude the paper in section 7.

2. RELATED WORK

Recent years the ensemble numerical weather prediction systems (EPS) have drawn an increasingly attention from the hydrological community [2]. Many flood forecasting systems rely on precipitation inputs that usually come from observation networks (rain gauges) and radar [3, 2]. But for medium term forecasts (2-15 days ahead), EPS models must be used, especially when upstream river discharge data is not available [8]. Ensemble forecasts of precipitation are replacing single (deterministic) forecasts for extending streamflow predictions beyond 48 hours. However, improvements in the prediction of precipitation have lagged behind the much more significant improvements made by operational NWP models in forecasting many aspects of the large-scale circulation [1]. For example, [4] found using the Global Forecast System (GFS) the precipitation total was underestimated and that the spatial distribution of the rainfall was degraded by the resolution of the global model. Any uncertainty in the prediction of flood inundation and flood wave

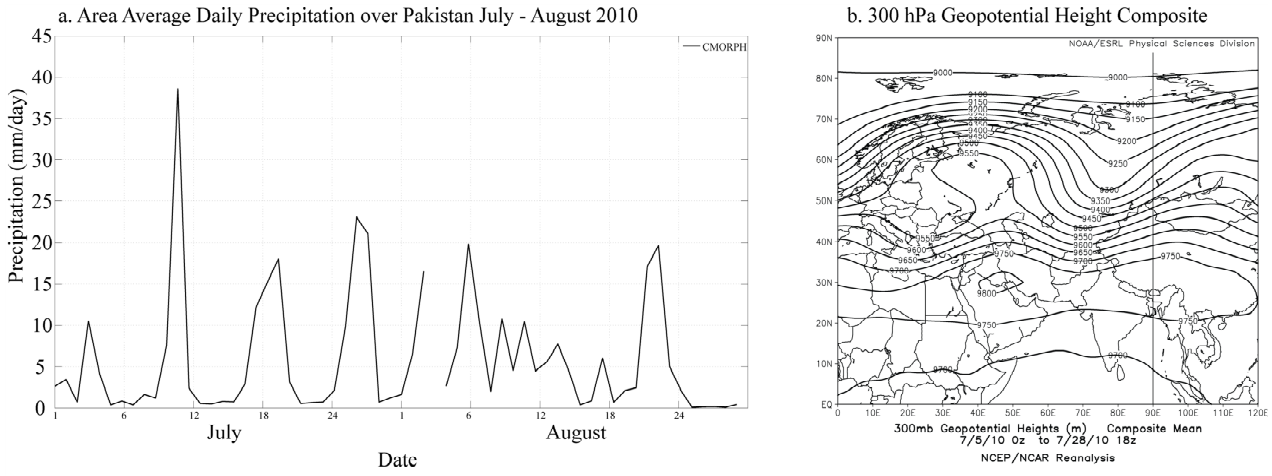

Figure 1: Panel a. is the daly average precipitation from the CMORPH satellite averaged over Pakistan, July and August 2010. Panel b. is the composite of the 300hPa geopotential height from July 5 to july 28, 2010 taken from NCEP/NCAR reanalysis. A very pronounced upper level ridge is clearly observed in Panel b. in the composite mean upstream of Pakistan that is closely flanked on both sides by upper level troughs that form what is often termed an "omega block" given its resemblance to the Greek letter ω.

propagation will be amplified by the large uncertainties in the prediction of rainfall and subsequent runoff generation [2, 15]

Feature selection that aims to select a desirable subset of existing features for predictive modeling has received considerable attention in statistics and machine learning in the past decades [6]. Recently, a new research approach has been proposed to solve feature selection problem where the feature space in data is extremely large, sometimes even infinite as not all features can be presented in the beginning. This is in contrast with most of the traditional feature selection methods which assume that all features are static and available to a learner before feature selection takes place. So far several algorithms have been proposed to battle this challenging problem, such as Grafting, Alpha-investing and Fast-OSFS.

- Perkins and Theiler [17] considered this problem as an online feature selection problem and proposed a grafting algorithm based on stagewise gradient descent. However, the grafting algorithm needs to determine the value of a regularization parameter in advance, and choosing a suitable regularization parameter requires information of a global feature set.

- Zhou et al. [24] approached this problem as streamwise feature selection and presented a simple yet fast algorithm called Alpha-investing. Alpha-investing sequentially considers new features for addition to a predictive model by modeling the candidate feature set as a dynamically generated stream. However, Alpha-investing cannot properly handle the original features without any prior information about the feature structure. In addition, Alpha-investing only considers adding new features but never evaluates the redundancy of selected features as time goes by.

- To overcome the deficiencies of those algorithms, Wu et al. [21] formulated the problem above as streaming feature selection, where features arrives one at a time and each new feature is required to be processed upon its arrival. Based on the idea of streaming features, Fast Online Streaming Feature Selection (Fast-OSFS) algorithm was proposed and can deal with extremely large or dynamic feature space such as the atmospheric dataset discussed later in this paper.

Class imbalance problem has been recognized to be existing in lots of application domains [20, 11]. Under-sampling, the method trying to balance class distribution through the random elimination of majority class examples, is very popular in dealing with such problems [5]. In the domain of test categorization, when training a binary classifier, all the samples in the training set that belong to the category is considered as relevant (positive) training data and all the samples belong to all the other categories as non-relevant (negative) training data. It is often the case that there is an overwhelming number of negative training data especially when there is a large collection of categories, which is typically an imbalanced data problem. To overcome this problem, in the work of [19] an under-sampling strategy is introduced by select a subset of *most-relevant non-relevant* data from the negative training set. The essential idea in this approach is to obtain more balanced positive and negative training data through under-sampling. In our work we adopt the about ideas by performing a hierarchical under-sampling process in both the feature selection process and the prediction modelling process.

3. OVERVIEW OF OUR APPROACH

In this section we give a overview of the proposed framework (Figure 2) for identifying the precipitation cluster precursors and forecasting periods of extreme precipitations that resulting floods. We address our two goals in the following steps:

- Identify meteorological predictor variables which contribute to atmospheric regimes that are conducive for

the precipitation clusters and construct a feature space to include the spatial and temporal information of the predictors.

- Apply under sampling and streaming feature selection techniques to select candidate precursors.
- Apply an advanced sampling approach to create the "most-correlated" datasets. The precipitation cluster precursors are identified through a validation process using the most-correlated datasets.
- Apply classification algorithms to train and evaluate the forecasting model using the precipitation cluster precursors and the evaluation set.

In step 3 only the features in the evaluation set are used to create the "most-correlated" datasets and leaving the class label in the evaluation set "untouched" (Section 4.3). We use the "most-correlated" datasets in the validation and forecasting process based on a hypothesis that our advance sampling process can improve the prediction model performance. The hypothesis is examined in the experiments. The identified precipitation cluster precursors are demonstrated and the forecasting model is evaluated using *Recall*, *Precision* and the *F-measure*. The classifier of adaboost with 1-knn as the weak learner is used to build the prediction model for both the validation and evaluation processes.

4. DATA PREPROCESSING

As a forecasting model, our data mining framework deals with two types of data: predictor variables (features) and criterion variable (class label). The predictor variables come from meteorology variables and the criterion variable is calculated using the historical precipitation data. In this section we introduce the data preprocessing approaches for both the two kind variables.

4.1 Class Label

The goal of our forecasting framework is to identify sequences of extreme precipitation events with a lead time of more than 5 days. In other words, we are trying to forecast an upcoming time period that has extreme heavy precipitations. In practise, we define any 21 days periods as extreme precipitation clusters if during which the total amount of precipitations reaches a historical high level (i.e., above the 95% percentile of the historical records). For example, we consider the day of July 1st a positive example if the total amount of precipitations from July 1st to July 21st is above 95% percentile of any sum of 21 days' precipitations in the historical records. We label such days as positive samples and our forecasting model aims to identify all the positive samples in the evaluation set using the precursors with lead times of more than 5 days.

4.2 Candidate Meteorological Variables Identification

The precursors we are looking for are meteorological predictor variables with certain spatial and temporal information. As meteorological variables, we have chosen several fields from the NCEP-NCAR Reanalysis dataset [9] on constant pressure surfaces that are typically used by meteorologists for making forecasts. The variables (Table 1) are chosen based on their fundamental importance in the basic physical processes involved in maintaining persistent large-scale flow regimes or in the production of precipitation. Specifically, we apply reasoning based on quasi-geostrophic (QG) theory and the theory of baroclinic [7]instability.

First, we chose the 300hPa zonal (i.e. east-west) winds, a proxy for the location and strength of the jet stream, which is important for several reasons. First, the jet stream serves as a waveguide directing the flow of Rossby waves as they propagate across the mid-latitudes. Rosby waves are important in the maintenance of persistent atmospheric regimes because they represent one mechanism through which energy propagates across the globe (over periods from days to weeks) and is transferred to the zonal mean flow (i.e. maintaining the westerly winds). The location of the jet stream is also important because storms require wind shear (strong change in wind speed with height) to develop, which is strongest near the core of the jet stream. By knowing the location of the jet stream, which is well known to exhibit persistence on scales much longer than individual storm events, we have information about the location where storms are likely to develop over multi-day periods. The geopotential height at 1000hPa and 500hPa are chosen because the 500hPa field will contain information about Rossby wave propagation and the two fields taken together allow us to infer where large-scale rising motion (and therefore precipitation) is likely to take place. The difference in the geopotential height between two constant pressure surfaces (i.e. the thickness) is proportional to the temperature of the atmospheric layer. The quasi-geostrophic (QG) vorticity is also proportional to the Laplacian of the geopotential height while the geostrophic component of the wind is proportional to the gradient of the geopotential height. The QG omega equation then relates vertical motion (needed for the production of precipitation) to the advection of thickness (i.e. temperature) and QG vorticity (at two levels) by the geostrophic wind, all of which can be inferred from the geopotential height at two levels. The 850hPa meridional (i.e. North-South) wind is chosen because it is extremely important for the transport of heat and moisture from the tropics into the mid-latitudes. We also include the precipitable water (i.e. total column water vapor) and 850hPa temperature so that we include explicit information about the transport of moisture and heat into the mid-latitudes. The moisture transport is needed to maintain the precipitation while the advection of temperature is crucial for strengthening (weakening) temperature gradients and the production (destruction) of fronts, which are important in producing vertical (i.e. rising) motion.

4.3 Feature Space Construction

The contribution of the meteorological predictors to certain precipitation clusters vary across space and time. For example, a predictor's value near the north pole may affect the atmospheric regimes over the Canada with a lead time of 2 days, but its effect to the atmospheric regimes over Mexico may have a lead time of 5 days. To counter this problem we build a feature space with the spatial and temporal information of the predictors to achieve a comprehensive coverage of the potential precursors. Particular, we sample every variable from 5,328 locations evenly distributed between the equator and the North pole (37 latitudes and 144 longitudes), and we do such sampling with a time span of 10 days to cover the period of 6 to 15 days lead time. For

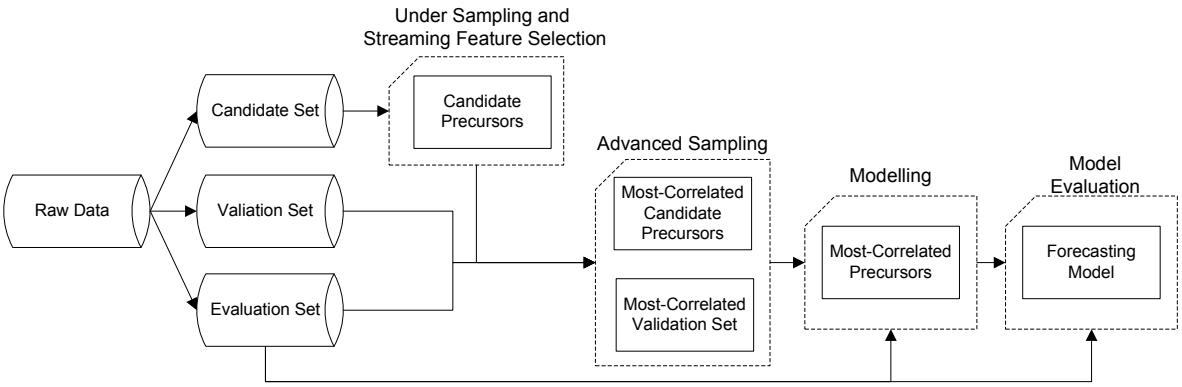

Figure 2: The flow chart of our integrated data mining framework. The forecasting model is built through the identification of precipitation cluster precursors.

Meteorological Variables	
Z300	300hPa Geopotential Height
Z500	500hPa Geopotential Height
Z1000	1000hPa Geopotential Height
U300	300hPa Zonal Wind
V300	300hPa Meridional Wind
U850	850hPa Zonal Wind
V850	850hPa Meridional Wind
T850	850hPa Temperature
PW	Precipitable water

Table 1: Candidate meteorological predictor variables that contribute to the atmospheric regimes leading to extreme precipitation clusters

example, we use the predictor variables sampled from July 1st to July 15th to predict whether there is an upcoming extreme precipitation clusters in July 21st. By doing so we construct an enormous spatial and temporal feature space of 479,520 features (9 predictor variables time 5328 locations time 10 days).

Finally, both the feature set and the class labels are divided to three parts: candidate set, validation set, and evaluation set (Figure 2) with the purpose of building, validating, evaluating the forecasting model, respectively.

5. PRECURSOR IDENTIFICATION

The task of identifying precipitation cluster precursors is completed through the feature selection and model validation processes using re-sampling and feature selection techniques. In this section we firstly give a series of formal notations and definitions related to the Fast Online Streaming Feature Selection (Fast-OSFS) algorithm. Then we introduce the streaming feature selection process by demonstrating the pseudo-code of Fast-OSFS and the hierarchical re-sampling processes.

5.1 Notations and Definitions

To characterize relevance between meteorological features and precipitation clusters, an input feature can be in one of three disjoint categories, namely, strongly relevant, weakly relevant or irrelevant [10]. Let $F = \{F_1, F_2, ..., F_n\}$ represent the full set of the meteorological features constructed in Section 4.3, C denotes the class attribute (sum of upcoming 21 days precipitations, Section 4.1) and $F - \{F_i\}$ represent the feature subset excluding F_i.

A data instance in our dataset can be described as an assignment of values $f = (f_1, f_2, ..., f_n)$ to the set of features F. Here we assume that all of the data instances are drawn from some probability distribution over the feature space. Formally, for each assignment of values f to F, we have a probability $P(F = f)$ (hereafter we use $P(F)$ for the probability).

Definition 1 (Conditional Independence) A feature $F_i \in F$ and the class attribute C are conditionally independent on a subset $S \subseteq F - \{F_i\}$, iff $P(C|F_i, S) = P(C|S)$

Definition 2 (Strong Relevance) A feature F_i is strongly relevant to C iff

$$\forall S \subseteq F - \{F_i\}, P(C|S) \neq P(C|S, F_i) \quad (1)$$

Definition 3 (Weak Relevance) A feature F_i is weakly relevant to C iff it is not strongly relevant, and

$$\exists S \subset F - \{F_i\}, P(C|S) \neq P(C|S, F_i) \quad (2)$$

Definition 4 (Irrelevance) A feature F_i is irrelevant to C iff it is neither strongly nor weakly relevant, and

$$\forall S \subseteq F - \{F_i\}, P(C|S, F_i) \neq P(C|S) \quad (3)$$

Weakly relevant features can be further divided into redundant features and non-redundant features [22].

Definition 5 (MB: Markov Blanket)

5.1 The Markov blanket of feature F_i: denoting $M_i \in F - F_i$ a subset of features, if for the given M_i the following property

$$\forall F_k \in F - \{M_i \cup F_i\}, P(F_i|M_i, F_k) = P(F_i|M_i) \quad (4)$$

holds, then M_i is a Markov blanket for F_i or ($MB(F_i)$ for short).

5.2 The Markov blanket of class label C: denoting $M_c \in F$ a subset of features, if for the given M_c the following property

$$\forall F_k \in F - \{M_C\}, P(C|M_i, F_k) = P(C|M_i) \quad (5)$$

holds, then M_C is a Markov blanket for C or ($MB(C)$ for short).

Definition 6 (Redundant Features) A feature F_i is redundant to the class attribute C, iff it is weakly relevant to C and has a Markov blanket, $MB(F_i)$, which is a subset of the Markov blanket of $MB(C)$.

5.2 Fast-OSFS

The pseudo-code of the Fast Online Streaming Feature Selection (Fast-OSFS) method is shown in Algorithm 1, where $Ind(C, X|S)$ denotes the conditional independence test between a feature X and the class attribute C given a subset S, $Dep(C, X|S)$ represents the conditional dependence test, and BCF stands for the set of best candidate features so far. Fast-OSFS employs a two-phase optimal subset discovery scheme: online relevance analysis (lines 5-8) and online redundancy analysis (lines 9-21). In the relevance analysis phase, Fast-OSFS discovers strongly and weakly relevant meteorological features and adds them into the set of best candidate precursors so far (BCF, Best Candidate Features). When a new meteorological feature arrives, Fast-OSFS assesses its relevance to the upcoming precipitation clusters (C) and decides to either discard the new feature or add it to BCF according to its relevance. Once a new feature is included into BCF, the redundancy analysis phase is triggered. If a subset exists within BCF to make any existing feature in BCF and the class attribute C conditionally independent, the previously selected candidate precursor Y ($Y \in BCF$) becomes redundant and is removed from BCF (line 18).

Algorithm 1: *The Hotspot Optimization Tool*

```
   Data:
   X, Y : features
   BCF : the best candidate feature set
1  BCF = {};
2  repeat
3  |  added = 0;
4  |  X ← get_new_feature()
5  |  /*online relevance analysis */
6  |  if Dep(C, X|∅) then
7  |  |  added=1;
8  |  end
9  |  /*Redundancy analysis 1:*/
10 |  if added then
11 |  |  if ∃S ⊂ BCF, Ind(C, X|S) then
12 |  |  |  go to Step 3 /*Discard X */
13 |  |  end
14 |  |  BCF = BCF ∪ X;
15 |  |  /*Redundancy analysis 2: */
16 |  |  for each feature Y ∈ BCF − X do
17 |  |  |  if ∃S ⊂ BCF, Ind(C, Y|S) then
18 |  |  |  |  BCF = BCF − Y;
19 |  |  |  end
20 |  |  end
21 |  end
22 until a predefined accuracy satisfied;
23 output BCF
```

5.3 Hierarchical Re-sampling

As motioned in Section 1, the extreme precipitation clusters we aimed are rare events. Particular, the studied datasets built with the label of such clusters (Section 4.1) are extremely imbalanced, with a positive sample and negative sample rate of 1:19. To deal with this problem we develop a hierarchical re-sampling approach that includes a under sampling process in the feature selection part and an advanced sampling process in the model validation part.

5.3.1 Under Sampling for Balance Dataset

As discussed in the work of [23], feature selection methods using two-sided metrics combine the positive and negative features so as to optimize the accuracy ($\frac{TP+FP}{TP+TN+FP+FN}$). In case of an imbalanced dataset with much more negative samples than the positive samples, two-sided metrics cannot ensure the optimal combination of positive and negative features according to F-measure ($\frac{2TP}{2TP+FP+FN}$). To counter this problem we apply an systematic under-sampling approach to achieve balanced datasets for the best performance of our feature selection algorithm.

Particularly, we count the number of extreme precipitation clusters (positive samples)in the candidate set, and randomly choose the same amount of negative samples from the set and combine them to create a new balanced feature set. The major drawback of the under sampling process is that it may discard potentially useful meteorological features that could be important for forecasting. In our work we perform the under-sampling N times and use the results to construct N balanced features sets. Then we run the streaming feature selection algorithm to identify candidate precursor sets from the balanced sets.

5.3.2 Advanced Sampling and Model Validation

The subsets of features generated by the Fast-OSFS algorithm constitute a series of candidate precursor sets. To identify the best set of precursors, prediction models are built and evaluated using the validation set. Instead of using all the samples in both the candidate precursor sets and the validation set, we hypothesize that using the "most-correlated" datasets generated through an advanced sampling approach can improve the prediction performance in flood forecasting.

Particularly, we are looking for datasets (both the candidate precursor set and the validation set) in which the features are most-correlated with the evaluation set. For example, assuming we have a dataset D (D can be a candidate set or the validation set) containing 10 years' data and a evaluation set T with 2 years' data. Both of them contain the same two candidate precursors: feature a and b. To avoid confusion we name the two precursors in D Da and Db, and the ones in the evaluation set Ea and Eb. We firstly divide D into 9 partitions with each part having the same length of samples as the evaluation set (i.e. 1st-2nd years, 2nd-3rd years,...). For each partition we calculate the correlation coefficients (Formula 5) between Da and Ta, Db and Tb, respectively. Then we sum the absolute values of the two correlation coefficients for each partition and sort the partitions using the sum values from high to low. The top t partitions will be selected to construct the most-correlated dataset. We call this process "advance sampling" instead of "under sampling" because in the most-correlated dataset some of the samples may be over sampled, i.e. dataset con-

Number of Features	Validation Sets	Evaluation Sets	Random Selected Features			Precursors Selected using Fast-OSFS		
			Recall	Precision	F-measure	Recall	Precision	F-measure
15	2001-2010	1999-2000	0.22	0.22	0.22	0.82	0.67	0.74
12	1999-2000,2003-2010	2001-2002	0.21	0.30	0.25	0.96	0.79	0.86
11	1999-2002,2005-2010	2003-2004	0.30	0.15	0.20	0.83	0.72	0.77
18	1999-2004,2007-2010	2005-2006	0.18	0.19	0.18	0.74	0.72	0.72
21	1999-2006,2009-2010	2007-2008	0.22	0.19	0.20	0.86	0.80	0.83
12	1999-2008	2009-2010	0.34	0.36	0.35	0.93	0.74	0.82
Average			0.23	0.24	0.23	0.86	0.74	0.79

Table 2: Model performance. For each pair of validation and evaluation sets, the best set of most-correlated precursors are selected using the validation sets and evaluated using the evaluation sets. Same number of features as the sets of most-correlated precursors are randomly selected from the same balanced candidate feature sets for comparison.

taining the partitions of 2nd-3rd years and 3rd-4th years over samples the 3rd year data.

$$r = \frac{\sum(X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum(X - \overline{X})^2}\sqrt{(Y - \overline{Y})^2}} \quad (6)$$

where X and Y are two features and $\overline{X}$ and $\overline{Y}$ are the averages of the features, respectively.

The most-correlated datasets are built for each candidate precursor sets and evaluated on the most-correlated validation set. The candidate precursor set with the best prediction performance is selected. We test the hypothesis of using most-correlated datasets for better performance in the experiments.

6. EXPERIMENTAL RESULTS

In this section we present the experimental results from our forecasting framework. We first introduce the historical data used in the experiments. Then, we show and discuss the advantages of using most-correlated dataset using the experimental results. The performance of the forecasting model is evaluated at the end.

6.1 Experimental Setup

The dataset used in our study has 23,011 observations over 63 years (from January 1st, 1948 to December 31st, 2010) and each observation is described by a set of 479,520 features (9 variables time 5,328 locations time 10 days). Historical spatial average precipitation data (the mean of daily precipitation totals from 22 stations divided by the standard deviation) of the state Iowa from the same time period is used to create the class label. The dataset of 1948-1998 (51 years) is used as the candidate set (Figure 2). The other 12 years data (1999-2010) are further divided into the validation set and the evaluation set in a rotated manner: 10 years for validation and the remained 2 years for evaluation.

For each pair of the validation and evaluation sets, we run the Fast-OSFS algorithm 10 ($N = 10$) times with random under sampled balanced candidate feature sets. We also conduct random feature selection on the same candidate feature sets for comparison. For each round, the most-correlated datasets are identified firstly using the advanced sampling process we proposed in section 4.3 with $t = 1$ for the most-correlated validation sets and $t = 5$ for the most-correlated candidate sets. Then the candidate sets (both from Fast-OSFS and random feature selection) having best performance during the validation processes are selected as the precursors and evaluated on the evaluation set (Table 2). As we are trying to predict a 21-days period of heavy precipitation, all of the evaluation results are adjusted by using a tolerance zone of one day. For example, if in the evaluation set June 1^{st} is a positive example, we consider the positive prediction of May 31^{st} or June 2^{nd} a "true positive".

6.2 Model Performance and Precursor Demonstration

The model evaluation results are shown in Table 2. On average the forecasting models built using the most-correlated precursors capture 86% (Average Recall=0.86) extreme precipitation clusters that are conducive for flooding in the evaluation sets with a precision of 74% (Average Precision=0.74). The overall performance of the models is evaluated with the average F-measure (0.79). The precursors selected using the Fast-OSFS algorithm significantly improve the prediction compare to the random selected features(F-measure: 0.79 compared to 0.23)

The identified 12 precursors that contribute to the extreme precipitation clusters in the state of Iowa between 2009-2010 are demonstrated using the map shown in Figure 3. For example, the blue square on the map shows that during the year of 2009-2010, the 850hPa zonal wind in that location has a significant effect on the upcoming extreme precipitation clusters in the state Iowa with a lead time of 7 days. All the precursors come from 4 out of the 9 meteorological predictor variables used in the model. Some of the unselected features may also have considerable influence on target precipitation clusters. The reason of discarding such influence features is because they are considered redundant by the Fast-OSFS algorithm giving the selected precursors. Also, we are encouraged to note that several identified precursors, like the precipitable water over the Gulf of Mexico, are physically meaningful.

6.3 Examing the Most-Correlated Datasets

We introduce the most-correlated datasets in our framework with two hypothesises:

- The model selected using the most-correlated validation set perform best in the evaluation set.
- The model built with the most-correlated precursors works best on the evaluation set.

The two hypothesises are tested using the experimental data. Particularly, we construct random datasets having same length

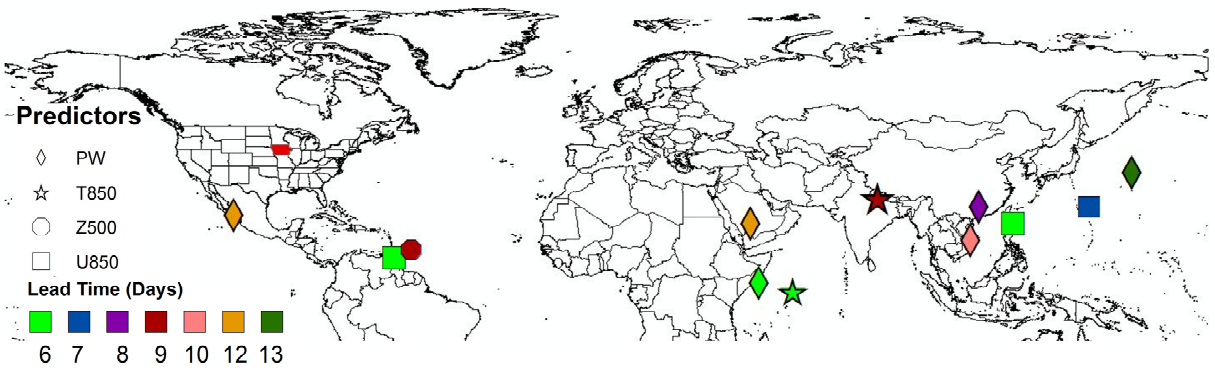

Figure 3: The map of the identified precursors that contribute to the extreme precipitation clusters in the state of Iowa (the red polygon in the map) between 2009-2010. PW, T850, Z500, and U800 stand for precipitable water, 850hPa temperature, 500hPa geopotential height, and 850hPa zonal wind, respectively. For example, the blue square on the map means the 850hPa zonal wind in that location has a significant effect on the upcoming extreme precipitation clusters in the state Iowa with a lead time of 7 days.

with the most-correlated sets and compare the model performance between them. In the tests, datasets from 1948-1998, 1999-2008, 2009-2010 are used as the candidate set, validation set, and evaluation set, respectively. Based on the three parts 10 most-correlated candidate sets, 1 most-correlated validation set and 1 most-correlated precursor set are built.

To test the first hypothesis, we randomly selected a dataset (the "random validation set") from the validation set (1999-2008) that having the same length and no over-lapping with the most-correlated validation set. Then we build 10 models using the 10 most-correlated candidate sets and evaluate them using the random validation set, the most-correlated validation set, and the evaluation set, respectively (Figure 4). Using the most-correlated validation set the best candidate set (NO.2) is successfully identified and the random validation set fails to achieve this. In the test the models' performance on the random validation sets are low. This is because the models are built using the most-correlated candidate sets which are customized for the evaluation set and there are very limited correlations between the evaluation set and the random validation set.

Secondely, we randomly sampled 9 datasets from the 1948-1997 dataset using the same features and the same length (10 years) as the most-correlated precursor set. We evaluated the models built using the most-correlated precursors and the 9 datasets on the evaluation set (Figure 5). The model built using the most-correlated precursors achieves a F-measure that is much higher than the other models.

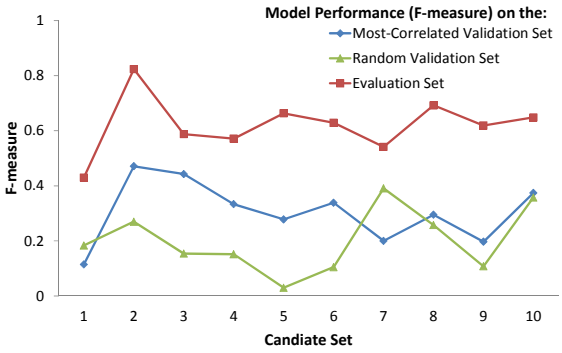

Figure 4: The models built using the 10 most-correlated candidate sets are evaluated on the random validation set, the most-correlated validation set, and the evaluation set, respectively. The best candidate set (set 2) is successfully identified using the most-correlated validation set.

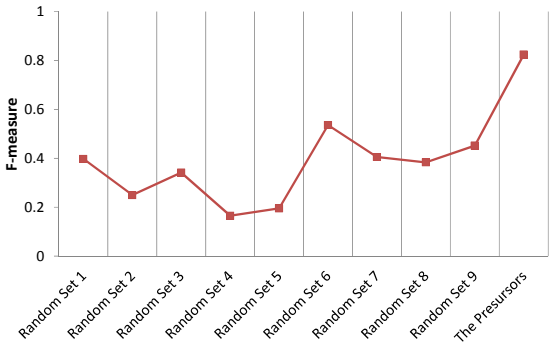

Figure 5: The models built using the 9 random sets and most-correlated precursor set are evaluated on the evaluation set. The precursors' model (rightmost) has the best performance.

7. CONCLUSION AND FUTURE RESEARCH

Improving the reliability and lead times of flood forecasts is critical for providing early earnings required to mobilize better preparedness for and response to disastrous flood events. In this paper, we discuss an integrated end-to-end data mining framework on precursor identification, dimensionality reduction, model validation and prediction to analyze the flood triggering precipitation clusters. In our future work, we want to explore the impact of the sequential order of the streaming features on precursor identification. Also, we want to explore an alternative class labeling process in the training data. Finally, we plan to extend our analy-

ses to other land. The framework is capable to be applied to other geographic areas because the candidate meteorological variables are collected globally. The project team is currently participates in flood warning planning organized by the World Bank and the Government of Pakistan.

8. REFERENCES

[1] J. P. Charba and F. G. Samplatsky. High-resolution gfs-based mos quantitative precipitation forecasts on a 4-km grid. *Monthly Weather Review*, 139(1):39–68, 2011.

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

[3] A. P. de Roo, B. Gouweleeuw, J. Thielen, J. Bartholmes, P. Bongioannini-Cerlini, E. Todini, P. D. Bates, M. Horritt, N. Hunter, K. Beven, et al. Development of a european flood forecasting system. *International Journal of River Basin Management*, 1(1):49–59, 2003.

[4] S. Dravitzki and J. McGregor. Predictability of heavy precipitation in the waikato river basin of new zealand. *Monthly Weather Review*, 139(7):2184–2197, 2011.

[5] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou. On the class imbalance problem. In *Natural Computation, 2008. ICNC'08. Fourth International Conference on*, volume 4, pages 192–201. IEEE, 2008.

[6] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[7] I. M. Held, R. T. Pierrehumbert, S. T. Garner, and K. L. Swanson. Surface quasi-geostrophic dynamics. *Journal of Fluid Mechanics*, 282:1–20, 1995.

[8] T. M. Hopson and P. J. Webster. A 1-10-day ensemble forecasting scheme for the major river basins of bangladesh: Forecasting severe floods of 2003-07*. *Journal of Hydrometeorology*, 11(3):618–641, 2010.

[9] 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*, 77(3):437–471, 1996.

[10] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324, 1997.

[11] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 30(1):25–36, 2006.

[12] E. N. Lorenz. Deterministic nonperiodic flow. *Journal of the atmospheric sciences*, 20(2):130–141, 1963.

[13] M. C. Morgan, D. D. Houghton, and L. M. Keller. The future of medium extended-range weather prediction: Challenges and a vision. *Bulletin of the American Meteorological Society*, 88:631, 2007.

[14] F. Pappenberger, K. J. Beven, N. Hunter, P. Bates, B. Gouweleeuw, J. Thielen, A. De Roo, et al. Cascading model uncertainty from medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions within the european flood forecasting system (effs). *Hydrology and Earth System Sciences Discussions*, 9(4):381–393, 2005.

[15] F. Pappenberger and R. Buizza. The skill of ecmwf precipitation and temperature predictions in the danube basin as forcings of hydrological models. *Weather and Forecasting*, 24(3):749–766, 2009.

[16] J. L. Pelly and B. J. Hoskins. A new perspective on blocking. *Journal of the atmospheric sciences*, 60(5):743–755, 2003.

[17] S. Perkins and J. Theiler. Online feature selection using grafting. In *International Conference on Machine Learning*. Citeseer, 2003.

[18] C. Schwierz, M. Croci-Maspoli, and H. Davies. Perspicacious indicators of atmospheric blocking. *Geophysical research letters*, 31(6):L06125, 2004.

[19] A. Singhal, M. Mitra, and C. Buckley. Learning routing queries in a query zone. In *ACM SIGIR Forum*, volume 31, pages 25–32. ACM, 1997.

[20] S. Visa and A. Ralescu. Issues in mining imbalanced data sets-a review paper. In *Proceedings of the sixteen midwest artificial intelligence and cognitive science conference*, pages 67–73. sn, 2005.

[21] X. Wu, K. Yu, W. Ding, H. Wang, and X. Zhu. Online feature selection with streaming features. *IEEE transactions on pattern analysis and machine intelligence*, 2012.

[22] L. Yu and H. Liu. Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*, 5:1205–1224, 2004.

[23] Z. Zheng, X. Wu, and R. Srihari. Feature selection for text categorization on imbalanced data. *ACM SIGKDD Explorations Newsletter*, 6(1):80–89, 2004.

[24] J. Zhou, D. P. Foster, R. A. Stine, and L. H. Ungar. Streamwise feature selection. *Departmental Papers (CIS)*, page 335, 2006.