Proceedings of 2016 IEEE 13th International Conference on Networking, Sensing, and Control
Mexico City, Mexico, April 28-30, 2016

# An Evaluation of Big Data Analytics in Feature Selection for Long-lead Extreme Floods Forecasting

Yong Zhuang[1], Kui Yu[2], Dawei Wang[1], and Wei Ding[†1]
[1]Department of Computer Science
University of Massachussets Boston
Boston, Massachussets 02125-3393
[2]School of Information Technology and Mathematical Science
University of South Australia
Adelaide, 5095, SA, Australia
Email: [1]$\{yong, dawei.wang, ding\}@cs.umb.edu$, [2]$kui.yu@unisa.edu.au$

***Abstract*—A type of extreme disastrous floods are associated with a sequence of prior heavy precipitation events occurring frequently from over several days to several weeks. Transitional methods for precipitation clusters prediction usually rely on the measurement and analyses of meteorological variables. However while a short-term prediction of certain location depends only on variables in near spatial and temporal neighborhood, predictions with long lead time must consider variables in a long time window and large spatial neighborhoods, this means an enormous amount of potentially influencing variables and only a subset of them strongly relate to prediction. Processing a deluge of variables and discovering strongly relevant features pose a significant challenge for big data analytics.**

**Finding influencing variables calls for automated methods of strongly relevant feature selection, which is what online streaming feature selection provides. In particular, online streaming feature selection, which deals with the stream of features sequentially added while the total data observations are fixed, aims to select a subset of strongly relevant features from the original feature set. In this paper, we apply four state-of-the-art online streaming feature selection methods for building long-lead extreme floods forecasting models, which identify optimal size of strongly relevant meteorological variables and confine learning the prediction model on the relevant feature set instead of the original entire feature set. The prediction models are evaluated and compared systematically on the historical precipitation and associated meteorological data collected in the State of Iowa.**

***Index Terms*—Online Streaming Feature Selection, Online Group Feature Selection, Precipitation Prediction.**

## I. Introduction

Extreme floods are the one of the most destructive hazards on Earth. Despite local efforts and national encouragement to mitigate flood hazards and regulate development in flood-prone areas, flood damages have increased in the United States in the past decades. Through analysing the figures from observation networks (rain gauges) and radar, National Weather Service (NWS) can provide a medium term (1-5 days ahead) flood warning currently[1]. However, long-lead (5-15 days ahead) prediction of extreme floods [7], which is great important to society for providing support of emergency response, still has relative low accuracy [8]. For instance, during the Colorado flooding in 2013, Because the amount of moisture in the atmosphere over Denver was at a record high for any September day on the morning of the $11^{th}$, computer models were not consistent on the exact location of the heavy rain [15].

Because a type of extreme floods are associated with a sequence of prior heavy precipitation events occurring frequently from over several days to several weeks and one of the most common and well studied approach for real-world predictive problems is classification. So long-lead forecasting of extreme floods can be formulated as a classification problem by identifying the precursors to heavy precipitation event clusters. Nevertheless, with the prediction lead time increasing, the potentially influencing meteorological variables in longer time window and larger spatial neighborhoods should be considered. For long-lead heavy precipitation prediction, this means an enormous amount of potentially influencing variables and only a subset of them strongly relate to prediction. How to deal with enormous variables and discovering strongly relevant ones are major challenges for big data analytics.

Finding strongly relevant features is what feature selection provides. Specially, Feature Selection aims to select a subset of relevant features from the original feature set for constructing forecasting model, in order to achieve simplification of models for easier interpretation, time efficiency, and enhanced generalization by reducing over-fitting. Traditionally online feature selection deals with the data observations sequentially added while the total dimensionality is fixed. In recent years, online streaming feature selection [1] has been attracted much attention. In contrary, it deals with sequentially added dimensions in feature space while the number of data instances is fixed. Many big data applications call for online streaming feature selection to consume sequentially added dimensions over time, especially extremely high feature space in big data analytics. For instance, Di et al. used OSFS for the heavy precipitation prediction [12]; Wang et al. implemented fast-OSFS algorithm for the flood forecasting [16]. However, there is no existing studies that apply different online streaming feature selection

†Corresponding author.
[1]http://water.weather.gov/ahps/

978-1-4673-9975-3/16/$31.00 

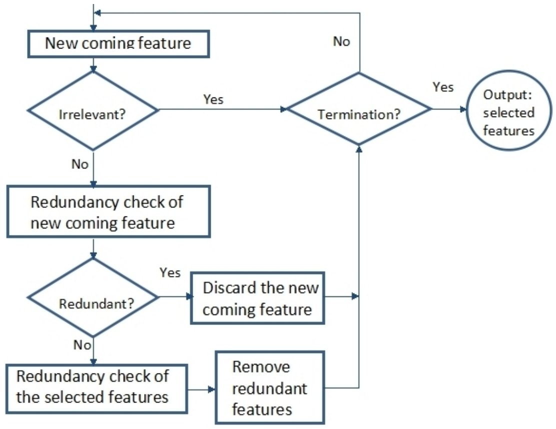


**Fig. 1:** The flow chart of online streaming feature selection.

methods on meteorological data and systematically study their performance on long-lead extreme precipitation forecasting. In this paper, we use four state-of-the-art online streaming feature selection methods, which have been successfully applied in computer forecast models, for the long-lead extreme flood prediction problem and compare the results systematically. Our contributions are:

- We formulate flood forecasting as a machine learning problem and construct a feature space which consists of potential relevant spatial and temporal meteorological features.
- We apply four state-of-the-art online streaming feature selection methods: alpha-investing, OSFS, SAOLA, and Group SAOLA, on the meteorological data to discover strongly relevant feature set respectively.
- Using four different relevant feature sets to build our long-lead extreme floods forecasting models, we systematically study historical precipitation and associated flood data in the State of Iowa. Based on the experiment results, the Group SAOLA algorithm is the most efficient in long-lead extreme floods forecasting problem.

## II. Method

Online streaming feature selection (Figure 1) is usually divided into two main types. one is processing one feature at a time. Its goal is to online select strongly relevant features $U$ from offered features $F$ for class $C$ by operating one feature at a time as well as in cases in which offered features of $F$ arrive sequentially in a stream. Here we are given a set of labeled inputs $D = \{F, C\}$ which is called the training set. And in general, $F$ is a set of $N$ sequential features ($F = \{f_1, ..., f_i, ...f_N\}$), such as the temperature and precipitation of different locations and days, $f_i$ is the $i^{th}$ feature in $F$. $U$ is a strongly relevant subset of $F$, and the form of the class label $C$ is usually a set of categorical variables from some finite set (e.g.,$C = \{c|c \in \{0, 1\}\}$), such as heavy precipitation and non-heavy precipitation. We donate $U_i$ as the currently selected feature set after processing the feature $f_i$, $K$ as any subset of selected feature set ($K \subseteq \{U_{i-1} \cup f_i\}$), $P(C|K)$ as the conditional probability distribution over class labels (given $K$), and $U'$ as candidate of current selected feature set. Then the approach of online streaming feature selection on streaming features added individually can be formulated as follows:

$$U_i = \underset{U'}{\operatorname{argmin}}\{|U'| : U' = \underset{K \subseteq \{U_{i-1} \cup f_i\}}{\operatorname{argmax}} P(C|K)\} \tag{1}$$

This corresponds to find the optimal set of relevant features $U_i$ for class $C$. Especially, when operating a new coming feature ($f_i$), the currently selected feature set $U_i$ will be updated dynamically. In this paper, we use Alpha-investing [2], OSFS [3,4], and SAOLA [5] to solve this problem.

The second main type of online streaming feature selection is processing grouped features sequentially. Here we are given features $F$ with prior group information ($F = \{G_1, ...G_j, ...G_M\}$), and the goal is to find the optimal set of relevant feature groups $U_G$ from offered features $F$, where $M$ refers to the number of groups, $G_j$ refers to the $i^{th}$ group features ($G_j = \{f_p, ...f_q\}$, $1 < p < q < N$), such as the temperature and precipitation of one-location-one-day, and $U_G$ refers to the a strongly relevant subset of $F$ with group information ($U_G = \{U_x | 1 \leq x \leq M, U_x \subseteq G_x\}$). We will denote the current selected feature groups after processing $G_j$ by $U_{G_j}$. So the online selection of dynamic groups can be formulated as follows:

$$\begin{aligned} U_{G_i} = & \underset{G' \subseteq \{U_{G_{i-1}} \cup G_i\}}{\operatorname{argmax}} P(C|G') \\ s.t. & \\ (a) & \forall f_i \in U_j, U_j \subset U_{G_i}, \\ & P(C|\{U_j - \{f_i\}, f_i\}) \neq P(C|\{U_j - \{f_i\}\}) \\ (b) & \forall U_j \subset U_{G_i}, \\ & P(C|\{U_{G_i} - U_j, U_j\}) \neq P(C|\{U_{G_i} - U_j\}). \end{aligned} \tag{2}$$

Here the objective corresponds to find the optimal set of feature groups $U_{G_i}$ for class $C$, Eq.(2 a) aims to find the minimal number relevant features in each group, and Eq.(2 b) aims to remove redundant features in currently selected set. In the paper, Eq.(2) can be solved by Group SAOLA [6].

### A. Alpha-investing

Alpha-investing method [2], which was proposed by Zhou et al. in 2006, is one of the well-known online streaming feature selection method. The idea is to dynamically update the relevant feature set by adding a new feature as addition into the current selected feature set if the new feature is correlated with the class feature. It use a dynamically threshold for adding a new feature, which is adjusted on the error reduction, against over-fitting. As a simple toy example of Alpha-investing, consider the process illustrated in Figure 2(A). We have nine streaming features as input and two outputs (selected features and irrelevant features). When analyzing the new coming feature (Feature 6), we will use linear regression and error reduction to calculate the correlation between temporary features, which are selected features (Feature 1, 5) and new coming feature (Feature 6), and class label (red cross). If the result greater than the given threshold, Feature 6 will be selected, otherwise it will be rejected. Alpha-investing has

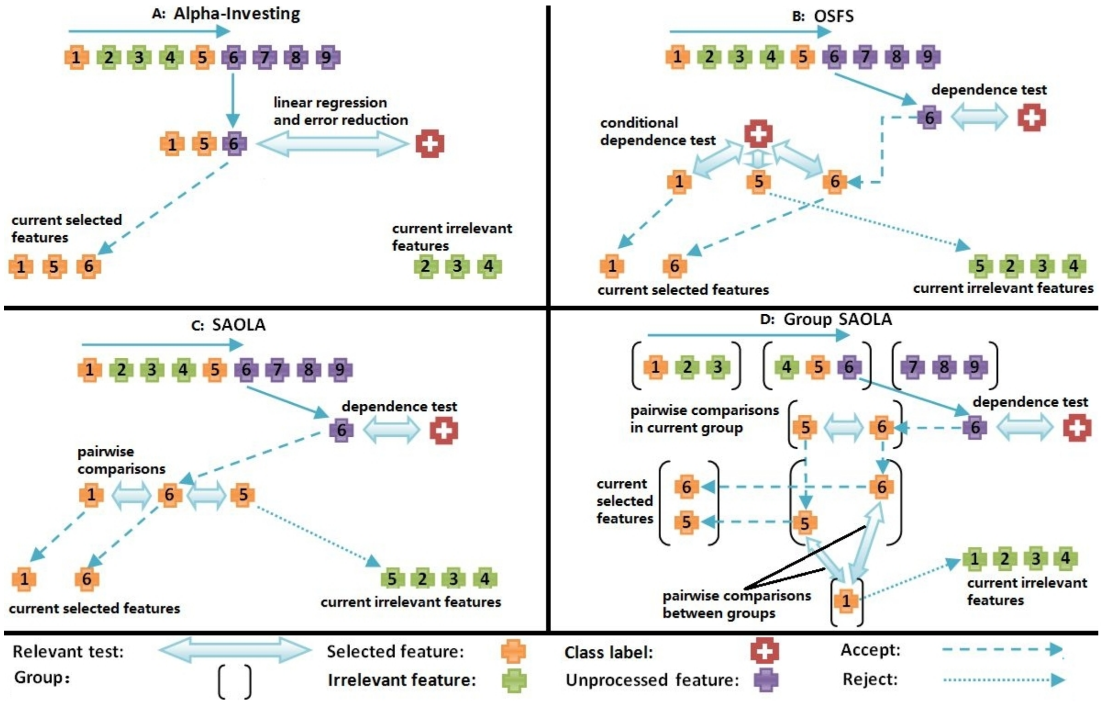


**Fig. 2:** Illustrating examples of the four relevant feature selecting processes that operate on Feature 6 of a set of nine features arriving one by one using four online streaming feature selection methods. "Current selected features" is the temporary relevant features which have been selected currently. "Current irrelevant features" is the irrelevant features which have been rejected by the relevant test.

been evaluated that it can efficiently work on the problems with high dimensionality [2]. However, it only calculates whether the new coming feature should be added into the current relevant feature set, and ignores removing redundant features from the selected feature set.

*B. OSFS*

Another well-known streaming feature selection method is Online Streaming Feature Selection (OSFS) [3,4], which Wu et al. proposed in 2010. Its idea is to find the best so far relevant feature set from the original feature set by two steps:

**Step 1.** Calculate whether the new coming feature is relevant to the class feature.

**Step 2.** Analyze whether there exists redundancy among the selected feature set currently once the new coming feature is added.

Such as the illustration in Figure 2(B), firstly, the new coming feature Feature 6 will be calculated the relevancy to the class label (red cross). If it related to class label, we will do conditional dependence test between temporary features (Feature 1, 5, 6), and class label for redundancy analysis. The redundant feature will be removed from selected feature set (Feature 5).

Compared with Alpha-investing, OSFS not only determines whether the new coming feature should be added into the current selected feature set, but also calculates if any feature can be removed from the current selected feature to keep the size of relevant feature set minimal.

*C. SAOLA*

Based on the OSFS method, Yu et al. proposed Scalable and Accurate OnLine Approach (SAOLA) by employing online pairwise comparisons between features in the currently selected feature set once a new coming feature is included [5]. As the process presented in Figure 2(C), after the dependence analysis between Feature 6 and class label, we will do pairwise comparisons in the temporary features (Feature 1, 5, 6) to find redundant features (Feature 5).

The benefits of SAOLA is when a data set includes extremely high dimensionality in big data analytics, SAOLA can significantly mitigates the expensively computational costs.

*D. Group SAOLA*

Unlike the methods above, Group SAOLA [6], which was proposed by Yu et al. in 2015, is a type of streaming feature selection method that particularly works on grouped streaming features (e.g., features which represent color, texture and other visual information). Utilizing the prior group information Group SAOLA can maximize each group's predictive performance for classification. Figure 2(D) gives the detail of this idea. Here the 9 input features are generated in 3 groups. For operating Feature 6, we firstly do dependence analysis on it, then we use pairwise comparisons on the selected features (Feature 5, 6) in the group which includes Feature 6 to make sure that we get the minimal relevant features of this group. Finally, we do pairwise comparisons between each selected groups for removing redundant features in selected feature set. For the effort of internal group pairwise comparisons and interactive group pairwise comparisons, Group SAOLA can greatly consume sequentially added features on problems with over a million potential features.

## III. Experiments

We apply alpha-investing, OSFS, SAOLA, and Group SAOLA on the historical meteorological data of Iowa for the long-lead extreme floods forecasting and evaluate the efficacy and accuracy systematically. This will be implemented as follows (Figure 3):

- Construct a feature set which consists of potential relevant spatial and temporal meteorological features.
- Define a binary class label for the extreme precipitation event depending on the precipitation information.
- Re-sample the potential relevant feature set to deal with class imbalance.
- Apply four online streaming feature selection methods to a select relevant feature sets from millions of spatial and temporal meteorological features.
- Building predictive models using the selected feature sets for long-lead extreme floods forecasting.

### *A. Data Preprocessing*

**Candidate Meteorological Variables Identification:** The potential relevant features for experiment are meteorological predictor variables with certain spatial and temporal information. We choose several variables from the NCEP-NCAR Reanalysis dataset [10], which are collected at different pressure surfaces and typically used by meteorologists for making forecasts, as meteorological predictor variables. Based on the theory of quasi-geostrophic and baroclinic [14], we specially choose 300hPa (U300) and 850hPa (U850) zonal winds(i.e. east-west) because they are a proxy for the location and strength of the jet stream which require wind shear (strong change in wind speed with height) to develop. And the information of the location of the jet stream exhibit persistence on scales much longer than individual storm events. Moreover, 300hPa (V300) and 850hPa (V850) meridional (i.e. North-South) winds are chosen because they are extremely important for the transport of heat and moisture from the tropics into the mid-latitudes. The geopotential height at 300hPa (Z300), 500hPa (Z500), and 1000hPa (Z1000) are chosen because the 500hPa field will contain information about Rossby wave propagation, which is a natural phenomenon in the atmosphere and oceans of planets that largely owe their properties to rotation, and the comparison with 300hPa and 500hPa fields allows us to infer where large-scale rising motion (and therefore precipitation) is likely to take place. On the other hand, the precipitation (PW) and 850hPa temperature (T850) fields are chosen because 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.

**Potential Relevant Feature Space Construction:** In order to build a feature space with the spatial and temporal information of the meteorological variables. We selected 9 variables from the NCEP-NCAR Reanalysis dataset [10] on constant pressure surfaces, which are typically used by meteorologists for making forecasts, as meteorological variables. Table I presents the information of the meteorological variables. Then we build the potential relevant feature space as following steps:

**Step 1.** Choose 5,328 locations, which are uniformly distributed between the equator and the North pole (37 latitudes and 144 longitudes).

**Step 2.** Sample every meteorological variable from 5,328 locations in the same day as the potential relevant features of one day.

**Step 3.** Repeat Step 2 until accumulating the potential relevant features of 10 continuous days before

**Class Label Creation:** Here, we are trying to predict an upcoming time period with the extreme precipitation event. We use 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 to create class label. We define any 14 days periods as extreme precipitation clusters and label it as a positive sample if the total amount of precipitations of the 14 days reaches a historical high level (i.e., above the $95\%$ percentile of the historical records). Otherwise, we label it as a negative sample. So our goal is to identify the positive samples in the evaluation set using the relevant feature set.

**Experimental Setup:** The dataset we used for experiment is the historical meteorological data collected in the State of Iowa, the United States from January 1st, 1948 to December 31st, 2010. It has totally 23,011 samples over 63 years and each sample has 479,520 features(9 variables $*$ 5,328 locations $*$ 10 days). This is a big data analytics problem, which fits for online streaming feature selection methods. In our experiments, we only pick the samples collected during the rainy season (April to October) every year, which might have correlation with precipitation events. The samples in (1948-2000) are used as training set to learn the forecasting model, and the other 10 years data are used as test set to evaluate the forecasting model.

**Re-sampling:** In our experiments, the extreme precipitation events rarely occurred in a year, so the total number of positive samples (extreme precipitation events) in the experimental data set is much less than the number of negative samples. If we directly use the imbalanced data for training our forecasting model, most of the negative prediction will be correct, and the accuracy will be high. However, we are interested in accurate classification of positive samples. So to deal with this class imbalance problem, we use the over-sampling [9] method and the under-sampling [11] method.

**Over-sampling:** We repeat all the positive samples until the number of the positive samples is approximately equal to the number of the negative samples. Then we combine them to create a new balanced data set for the following experiment.

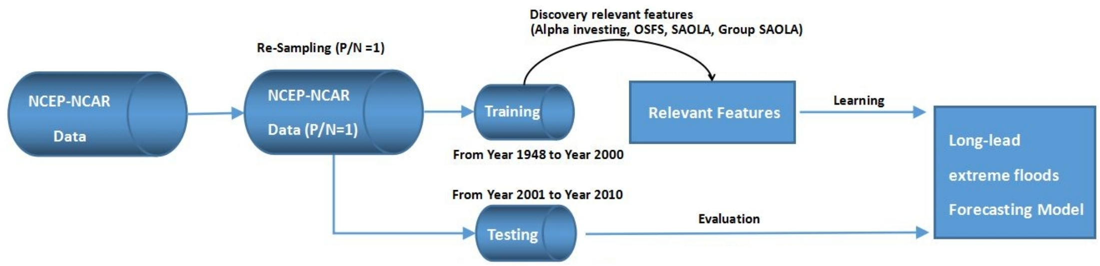


**Fig. 3:** The flow chart of our integrated data mining framework. The forecasting model is built through the learning on relevant feature set. P is the positive samples which means extreme precipitation event, otherwise, N is the negative samples.

**Under-sampling:** We count the number of positive samples. Then we randomly choose the same amount of negative samples from the experimental data set and combine them together to create a new balanced feature set for the following experiment.

### *B. Relevant feature set discovery*

We run four online streaming feature selection methods to discover the strongly relevant feature set. Especially in Group SAOLA, we assign every 9 meteorological features of one location one day as one group, then using this prior group information for relevant feature selection.

**Experiment 1:** Four online streaming feature selection methods + original data + KNN [13]. The aim of this experiment is to check the effect of four online streaming feature selection methods on imbalanced data.

**Experiment 2:** Four online streaming feature selection methods + over-sampled data + KNN. It aims to check the effect of four online streaming feature selection methods on the data balanced by over-sampling method.

**Experiment 3:** Four online streaming feature selection methods + under-sampled data + KNN. This experiment aims to check the effect of four online streaming feature selection methods on the data balanced by the under-sampling method. We do this experiment 10 times with randomly under sampled balanced data sets. Then we calculate the mean values of the static measures.

### *C. Experiments result*

Here we use Accuracy and F-measure for evaluation. Particular, Accuracy ($\frac{TP+FP}{TP+TN+FP+FN}$, where TP is true positive, TN is true negative, FP is false positive, FN is false negative) refers to the closeness of a predicted class label to a known class label. And F-measure ($\frac{2*TP}{2*TP+FP+FN}$) conveys the balance between the exactness and the completeness.

**Which experiment got the worst result?** Based on the experiment results in Table I, we can see that without class balance, all of four methods get high accuracy in Experiment 1 by the large number of correct predictions of negative samples.

**Which is the better class balance method for our experimental dataset, over-sampling or under-sampling?** Compared the result of Experiment 2 with Experiment 3, it can be seen that the under-sampling method works better than the over-sampling method. This is because in our experimental

| Experiments | Metrics | Alpha investing | OSFS | SAOLA | Group SAOLA |
|---|---|---|---|---|---|
| | The size of relevant feature set | 112 | 68 | 15 | 8 |
| Experiment 1 | Accuracy | 0.8235 | 0.827 | 0.8305 | 0.8435 |
| | F-measure | 0.1284 | 0.1128 | 0.1285 | 0.1425 |
| Experiment 2 | Accuracy | 0.4766 | 0.4789 | 0.4797 | 0.4976 |
| | F-measure | 0.239 | 0.2537 | 0.2594 | 0.2635 |
| Experiment 3 | Accuracy | 0.7028 | 0.7696 | 0.712 | 0.7189 |
| | F-measure | 0.7466 | 0.8039 | 0.7485 | 0.7589 |

**TABLE I:** The result of experiments, The classifier of 1-knn [13] is used to build the prediction model for both the validation and evaluation processes.

data, the positive samples are rare and any uncertainty with prediction errors from the positive class samples will be amplified in the over-sampling method.

**How does the redundancy analysis effect?**In the relevant feature selection process, for the new coming features, alpha investing only performs online relevance analysis, but OSFS also performs an online redundancy analysis. OSFS can get a more accurate relevant feature set. In the result of Experiment 3, although the size of relevant feature set of OSFS is only 68, which is nearly half of alpha investing's (112 features), all of measurements on OSFS are significantly better than the Alpha investing's. The effect of redundancy analysis can also be observed in the Figure 4 Alpha Investing. In our experiments, the variable of "T850" in day 1 is the first feature processed by all the online streaming feature selection methods. Alpha Investing are affected by this sequence and as a results, we can see many red crosses (T850 of day 1) occurs at the prime meridian. On the other hand, OSFS, SAOLA and Group SAOLA can improve this through the redundancy analysis process.

**Which is the best online streaming feature selection method for our experiment?** With the effort of online pairwise comparisons, the size of relevant feature set of SAOLA has sharlply decreased to 15. And using prior group information helps Group SAOLA get only 8 relevant features. However, the static result of them are still good. Moreover, in Group SAOLA, because we assigned every 9 meteorological

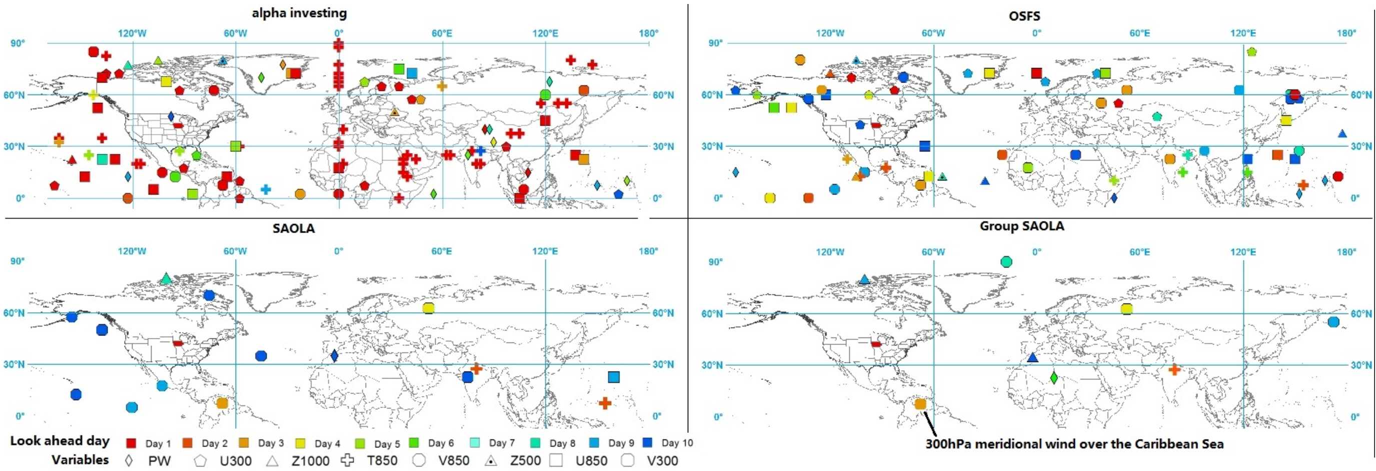


**Fig. 4:** The relevant feature set (according to the types of variables in the set). The red square in the map is the target area (Iowa) for flood labeling, and other symbols are the relevant meteorological variables. For example, the orange octagon in the plot of Group SAOLA means that during period from 2001 to 2010, the variable (300hPa meridional wind) over the Caribbean Sea has a significant effect on the upcoming extreme precipitation clusters in the state Iowa with a lead time of 3 days.

features of one-location-one-day as one group features, though the pairwise comparisons between groups, only one feature of one-location-one-day was chosen and the features of the other days of this location are removed (Figure 4 Group SAOLA). Also, we are encouraged to note that several relevant features, such as the 300hPa meridional wind over the Caribbean Sea, are physically meaningful.

## IV. Conclusion

In this paper, we apply four state-of-the-art online streaming feature selection methods for the long-lead extreme floods prediction problem and compare the results systematically. We use the historical precipitation and associated meteorological data collected in the State of Iowa to evaluate our prediction mode. In the experiments, because extreme precipitation event rarely occurred in a year, our experimental data is extremely imbalanced. We use over-sampling and under-sampling to deal with this problem respectively and get different balanced data sets, then we use these balanced data for the following experiment and compare the result. Based on the experiment result, under-sampling works better in our project. To deal with dimensionality reduction, we apply Alpha investing, OSFS, SAOLA, and Group SAOLA to discovery the relevant feature sets. Through the comparison of the experiment results of these 4 online streaming feature selection methods, we get that using OSFS can get the most accurate prediction, but SAOLA and Group SAOLA are more effective.

## References

[1] S-H.Ji, J-S.Choi, and B-H.Lee, A Computational Interactive Approach to Multi-agent Motion Planning, *International Journal of Control, Automation, and Systems*, Vol.5, No.3, 295-306, 2007

[2] Perkins, S. and Theiler, J., Online feature selection using grafting, *ICML*, 592-599, 2003.

[3] Zhou, J., Foster, D. P., Stine, R. A., and Ungar, L. H., Streamwise feature selection, *The Journal of Machine Learning Research*, Vol.7, 1861-1885, 2006.

[4] Wu, X., Yu, K., Wang, H., and Ding, W., Online streaming feature selection, *Proceedings of the 27th international conference on machine learning (ICML-10)*, 1159-1166, 2010.

[5] Wu, X., Yu, K., Ding, W., Wang, H., and Zhu, X., Online feature selection with streaming features, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol.35, Issue 5, 1178-1192, 2013.

[6] Yu, K., Wu, X., Ding, W., and Pei, J., Towards scalable and accurate online feature selection for big data, *Data Mining (ICDM), 2014 IEEE International Conference on*, 660-669, 2014

[7] Yu, K., Wu, X., Ding, W., and Pei, J., Scalable and Accurate Online Feature Selection for Big Data, *arXiv preprint arXiv:1511.09263*, 2015.

[8] de Roo, A. P., Gouweleeuw, B., Thielen, J., Bartholmes, J., Bongioanni niCerlini, P., Todini, E., ... and Pappenberger, F., Development of a European flood forecasting system, *International Journal of River Basin Management*, Vol.1, Issue 1, 49-59, 2003

[9] Cloke, H. L., and Pappenberger, F., Ensemble flood forecasting: a review, *Journal of Hydrology*, Vol.375, Issues 34, 613-626, 2009

[10] Ling, C. X., and Li, C., Data Mining for Direct Marketing: Problems and Solutions, *KDD*, Vol.98, 73-79, 1998

[11] Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., ... and Zhu, Y., The NCEP/NCAR 40-year reanalysis project, *Bulletin of the American meteorological Society*, Vol.77, Issue 3, 437-471, 1996.

[12] Kubat, M., and Matwin, S., Addressing the curse of imbalanced data sets: One sided sampling, *Proc. of the Int'l Conf. on Machine Learning*, 1997.

[13] Di, Y., Ding, W., Mu, Y., Small, D. L., Islam, S., and Chang, N. B., Developing machine learning tools for long-lead heavy precipitation prediction with multi-sensor data, *Networking, Sensing and Control (ICNSC), 2015 IEEE 12th International Conference on*, 63-68, 2015.

[14] Dudani, S. A., The distance-weighted k-nearest-neighbor rule, *Systems, Man and Cybernetics, IEEE Transactions on*, Vol.SMC-6, Issue 4, 325-327.1976.

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

[16] Gochis, D., Schumacher, R., Friedrich, K., Doesken, N., Kelsch, M., Sun, J., ... and Matrosov, S., The great Colorado flood of September 2013, *Bulletin of the American Meteorological Society*, Vol.96, Issue 9, 1461-1487, 2015.

[17] Wang, D., Ding, W., Yu, K., Wu, X., Chen, P., Small, D. L., and Islam, S., Towards long-lead forecasting of extreme flood events: a data mining framework for precipitation cluster precursors identification, *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1285-1293, 2013.