Galaxy: Towards Scalable and Interpretable Explanation on High-dimensional and Spatio-Temporal Correlated Climate Data

Yong Zhuang
Department of Computer Science
University of Massachusetts Boston
Boston, MA, USA
yong.zhuang001@umb.edu

David L. Small
Department of Civil
and Environmental Engineering
Tufts University
Boston, MA, USA
David.Small@tufts.edu

Xin Shu
Department of Computer Science
University of Massachusetts Boston
Boston, MA, USA
xin.shu001@umb.edu

Kui Yu
School of Computer and Information
Hefei University of Technology
Hefei, China
yukui@hfut.edu.cn

Shafiqul Islam
Department of Civil
and Environmental Engineering
Tufts University
Boston, MA, USA
Shafiqul.Islam@tufts.edu

Wei Ding
Department of Computer Science
University of Massachusetts Boston
Boston, MA, USA
Wei.Ding@umb.edu

***Abstract*—Interpretability has become a major criterion for designing predictive models in climate science. High interpretability can provide qualitative understanding between the meteorological variables and the climate phenomena which is helpful for climate scientists to learn causes of climate events. However, detecting the features which have efficient interpretability to observed events is challenging in spatio-temporal climate data because the key features may be overlooked by the redundancy due to the high degree of spatial and temporal correlations among the features, especially in high dimensionality. Furthermore, climate events occurred in different regions or different times may have different explanatory patterns, detecting explanations for overall climate phenomena is also difficult. Here we propose Galaxy, a new interpretable predictive model. Galaxy allows us to represent the explanatory patterns of subpopulations within an overall population of the target. Each explanatory pattern is defined by the smallest feature subset that the conditional distribution of target actually depends on, which we define as the minimal target explanation. Based on the detection of such explanatory patterns, Galaxy can detect the Galaxy space, the explanations for the overall target population, by sequentially detecting target explanation of every individual subpopulation of the target, and then forecast the target variable by their ensemble predictive power. We validate our approach by comparing Galaxy to several state-of-the-art baselines in a set of comparative experiments and then evaluate how Galaxy can be used to identify the explanatory space and give a referential explanation report in a real-world scenario on a given location in the United States.**

***Keywords*-Interpretable explanation, Long-lead rainfall forecasting, AdaBoost**

I. Introduction

With a rapid increase in the availability of spatio-temporal climate data and growing popularity of data mining techniques [1], qualitative understanding between the meteorological variables and the climate phenomena has become a major objective of current meteorology. This makes interpretability has a great need in the design of predictive models.

However, the climate data is always high-dimensional and spatio-temporal correlated, and so the relationships among the meteorological variables are very complicated - especially in spatial-temporal studies of numerous variables simultaneously [2], [3]. With time and space increasing, the number of elements potentially contributing to a meteorological event grows sharply. This makes identifying the causes of climate phenomena from large spatial-temporal scale meteorological features extremely difficult. For example, explaining phenomena 5 days ahead is typically less reliable than explaining phenomena for the next day. This occurs since small changes may likely influence observed weather events as time advances (butterfly effect), but such small changes can be easily overlooked due to the high degree of spatial and temporal correlations among the features as their magnitude decreases, so it is difficult to analyze how a multitude of tiny events will impact observed weather as time moves forward.

On the other hand, climate phenomena are the result of the interactions and operations of atmospheric physical effects on multiple spatial-temporal scales. This means the climate events occurred in different regions or different times may have different explanatory patterns. For example, compared with precipitation during the cold season, warm season precipitation generally occurs on smaller spatial-temporal scales with large gradients in precipitation amounts. We cannot use the same meteorological features' influence to explain all precipitation events. Thus identifying explanatory patterns in all perspective are of significant interest for understanding the causes of

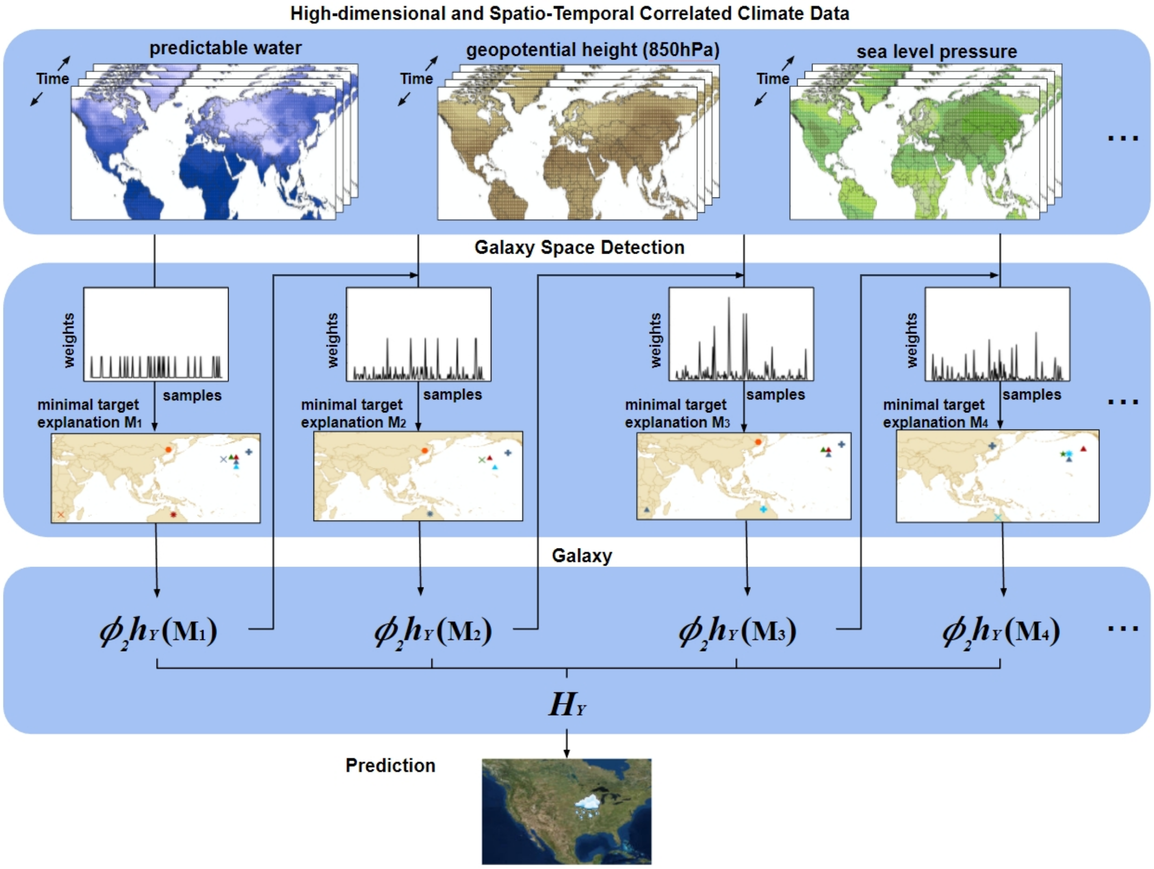

Fig. 1. An example of Galaxy for precipitation forecasting. Galaxy detects the Galaxy space, the explanations for overall target population, by sequentially detecting minimal target explanation of every individual subpopulation within the overall population, and then forecasting target by their ensemble predictive power.

climate phenomena.

In this paper, we propose a new interpretable predictive model Galaxy (Figure 1) which can detect explanatory patterns in all perspective of target climate phenomena from large spatial-temporal scale meteorological features and forecast by their ensemble predictive power. Our work is not only able to give interpretable explanations on high-dimensional spatio-temporal climate data, but also provide the preconditions for scientists for further studies. Thus, our main contributions are as follows:

- **Minimal Target Explanation:** We define the notion of minimal target explanation to represent the explanation that locally faithful to the target instances derived from the same subpopulation.
- **Galaxy Space:** We define the notion of Galaxy space to represent the explanation with global faithfulness of the overall population of the target feature.
- **Galaxy:** We design and implement the Galaxy algorithm, to discover the Galaxy space from mixture distributed feature data, and then forecast by its ensemble predictive power. In our empirical experiments, our algorithm outperforms state-of-the-art ensemble methods under dimensional feature space.
- **Interpretable on Real-world scenario:** We apply Galaxy to study historical precipitation data in the Des Moines river basin. Our empirical study includes 5,313,600 features over 67 years of data. We are able to understand precipitation forecasting in the area.

The rest of this paper is organized as follows. Section II reviews related work. Section III presents the Galaxy space, Galaxy detection algorithms, and as theoretical analysis of Galaxy. Section IV discusses our empirical studies on synthetic data and a real-world precipitation data set, and showcase the explanatory patterns. We conclude the paper in section V.

II. Related work

To provide high-quality explanations for observed weather events, we need to look for the features which are most likely to influence them. However, the high-dimensionality and high degree of spatio-temporal correlations are serious challenges of climate data. Although the existing dimension reduction methods are able to address the high dimensionality by reporting a subset of features which are strongly contribute on prediction, detecting the features most influent on observed weather events is still facing the following challenges.

- Scale amplification: Weather systems are very sensitive to changes in initial conditions. So many small perturbations in air motion could compound to result in large changes over longer time frames.

- Error magnification and analysis: Because the system is so sensitive, measurement error in monitoring devices can lead to errors in analysis.

Markov boundary based feature selection is the state-of-the-art dimension reduction technology using causal inference [4] [5] [6] [7], and a Markov boundary of the target feature can be tread as the knowledge needed to predict the behavior of the target. However, a unique Markov boundary ideally exists for targets in datasets under a strong faithfulness assumption [8] [9] [10], which is often violated in real-world data because of the occurrence of hidden variables, hypothesis test errors and some fake relevance of pure chances, so multiple Markov boundaries exist almost in all situations [11]. Which one can best explain observed weather events remains an open research problem.

On the other hand, the climate phenomena are the result of the interactions and operations of atmospheric physical effects on multiple spatial-temporal scales, the observed climate events are likely derived from mixture populations, which the climate events occurred in different regions or different times may derived from different subpopulations [12]. This makes one pattern may not be interpretable to all observed climate events. We are interested in how to detect the explanations for climate events of the mixture populations.

In this paper, we discuss a new predictive model with high interpretability to facilitate climate scientists to better understand causes of climate events.

III. Galaxy Space

A. Notation

For the remainder of this paper, we shall use:

- an italic lowercase letter to denote an instance (e.g. x).
- an italic capital Greek letter to denote a feature (e.g. X).
- a boldface, capital Greek letter to denote a feature set (e.g. $\mathbf{X} = (X_1, ..., X_k) \in \mathbb{R}^k$).
- a backslash to denote difference between feature sets. (e.g. $\mathbf{X} \backslash \{X_i\} = \{X_1, ..., X_{i-1}, X_{i+1}, ..., X_k\}$).
- $\mathbb{D}$ to denote a data set.
- Y to denote the target feature.
- P to denote a probability distribution.

B. Galaxy Space

The Galaxy space is designed for global faithfulness and efficient interpretability. Formally, suppose we have a dataset $\mathbb{D}$ that includes n instances. Each of the instance is in the form of $(\mathbf{X}, Y)$, where $\mathbf{X} = (X_1, ..., X_k) \in \mathbb{R}^k$ and $Y \in \mathbb{R}$, and the n instances of Y are derived from m subpopulations. Then the overall conditional distribution of Y can be represented as the mixture of subpopulation conditional distributions:

$$P(Y \mid \mathbf{X}) = \sum_{i=1}^{m} \phi_i P_i(Y \mid \mathbf{X}), \tag{1}$$

where ϕ is the mixture component weight, $P_i(Y \mid \mathbf{X})$ presents the conditional probability distribution of Y of the i^{th} subpopulation, and $P(Y \mid \mathbf{X})$ is the overall conditional population distribution of Y.

Definition 1. ***Target Explanation (TE):*** *A feature set $\mathbf{M} \subseteq \mathbf{X}$ is said to be a target explanation of $P(Y \mid \mathbf{X})$ if and only if:*

$$P(Y \mid \mathbf{X}) = P(Y \mid \mathbf{M}) \tag{2}$$

By Definition 1, a non-minimal target explanation can be trivially produced by adding redundant or irrelevant features into itself. Only minimal target explanations are of interest in this paper. If a feature subset $\mathbf{M} \subseteq \mathbf{X}$ is a minimal target explanation, it is more efficiently than $\mathbf{X}$ to interpret the instances derived from $P(Y \mid \mathbf{X})$.

Definition 2. ***Minimal Target Explanation (MTE):*** *A target explanation $\mathbf{M}$ is said to be a minimal target explanation if and only if no proper subset of $\mathbf{M}$ satisfies the definition of target explanation.*

Since a **MTE** of $P(Y \mid \mathbf{X})$ is the minimal explanation of $P(Y \mid \mathbf{X})$, it does not have any redundant or irrelevant feature. Then back to the mixture conditional distributions representation, we define partial target explanation as follows.

Definition 3. ***Partial Target Explanation (PTE):*** *If $P(Y \mid \mathbf{X})$ can be represented as mixture of subpopulation conditional distributions $\sum_{i=1}^{m} \phi_i P_i(Y \mid \mathbf{X})$, then we say a target explanation of $P_i(Y \mid \mathbf{X})$ is a partial target explanation of $P(Y \mid \mathbf{X})$.*

We also define a **MTE** of $P_i(Y \mid \mathbf{X})$ as a partial minimal target explanation **(PMTE)** of $P(Y \mid \mathbf{X})$. Then, we say **PMTE** is **locally faithful**, i.e. it is efficiently interpretable to the instances derived from a subpopulation of Y. Thus, for the overall population of Y, we define the Galaxy space of Y as follows.

Definition 4. ***Galaxy Space*** $\mathbb{G}$***:*** *If $P(Y \mid \mathbf{X})$ can be represented as mixture of subpopulation conditional distributions $\sum_{i=1}^{m} \phi_i P_i(Y \mid \mathbf{X})$, then we say $\prod_{i}^{m} \boldsymbol{M}_i$ is a Galaxy space $\mathbb{G}$ of Y if and only if every $\boldsymbol{M}_i$ corresponds a **MTE** of $P_i(Y \mid \mathbf{X})$.*

Based on the Definition 4, a Galaxy space $\mathbb{G}$ of Y is a set of **PMTE**s of $P(Y \mid \mathbf{X})$. Each **PMTE** provides a local partical minimal target explanation, and the complete set of **PMTE** is able to interpret the entire instances of Y, i.e. **globally faithful**. In order to look for a Galaxy space, we need to first detect **PMTE**.

C. Partial Minimal Target Explanation (PMTE) Detection

Detecting a **PMTE** of $P(Y \mid \mathbf{X})$ is to look for the “smallest” explanation that is locally faithful to a subpopulation of Y. Here we can further decompose it into two problems: (1) Detecting the “smallest” explanation **MTE** of a subpopulation of Y and (2) Identifying the instances from $\mathbb{D}$ which belong to this subpopulation.

To address the first problem, we utilize the approaches in causal inference. In the domain of causal discovery, a Bayesian network [8] is a standard tool for modeling the conditional dependencies of the features, and a Markov boundary of a

```
Algorithm 1: Partial Minimal Target Explanation (PMTE) Detection.
Input:
  • data set $\mathbb{D}$ for features $\mathbf{X}$; target feature $Y$; Markov boundary detection algorithm $f_Y$; learning algorithm $h_Y$;
    performance metric T;
Output:
  • $\mathbf{M}$, a partial minimal target explanation of $Y$.
  • $h_Y(\mathbf{M})$, a trained learning algorithm on $\mathbf{M}$.
begin
    $\mathbf{M}'_{init}$ = empty                      /* Initialize new Markov boundary with an empty set */
    $\mathbf{M}'$, $\mathbf{R} = f_Y(\mathbf{M}'_{init}, \mathbf{X})$    /* Detect $1_{st}$ Markov boundary $\mathbf{M}'$ and residual features $\mathbf{R}$ from
     $\mathbf{X}$ on $\mathbb{D}$ */
    $\mathbf{M} = \mathbf{M}'$
    Performance = $T_{h_Y(\mathbf{M}')}$
    for $\forall \mathbf{S} \subset \mathbf{M}'$ do
        $\mathbf{R}_{new}$ = $\mathbf{R}$
        $\mathbf{M}'_{init} = \mathbf{M}' \backslash \mathbf{S}$                 /* Initialize new Markov boundary as $\mathbf{M}' \backslash \mathbf{S}$ */
        repeat
            $\mathbf{M}'_{new}$, $\mathbf{R}_{new}$ = $f_Y(\mathbf{M}'_{init}, \mathbf{R}_{new})$ /* Replacing $\mathbf{S}$ by exploring its equivalent features
             from $\mathbf{R}_{new}$ */
            if $T_{h_Y(\mathbf{M}'_{new})} >$ Performance then
                $\mathbf{M} = \mathbf{M}'_{new}$
                Performance = $T_{h_Y(\mathbf{M}'_{new})}$
        until $\mathbf{R}_{new}$ is empty
    Return $\mathbf{M}$, $h_Y(\mathbf{M})$
```

target feature corresponds to a local causal neighborhood of it and consists of all its direct causes, effects, and causes of the direct effects. This means that knowledge of the values of the Markov boundary features should render all other features superfluous for predicting Y [13]. In faithful joint distributions of $(\mathbf{X}, Y)$, there exists a unique Markov boundary of Y [14]. However, in real-world data, the faithfulness condition may be violated by hidden variables, hypothesis test errors and some fake relevance of pure chances. This makes the Markov boundaries of the target variable not unique [11]. In order to define a unique minimal target explanation, we first state the definition of optimal predictor and link it with the concept of target explanation, then we detect the minimal target explanation using optimal predictor on the Markov boundaries of the target feature.

Definition 5. ***Optimal Predictor*** *[11]: Given a data set $\mathbb{D}$, a learning algorithm h_Y, and a performance metric T to assess the learner's model, a feature subset $\mathbf{M} \subseteq \mathbf{X}$ is an optimal predictor of Y if it maximizes the performance metric T for predicting Y using learner h_Y in the data set $\mathbb{D}$.*

The following theorem states the link between the optimal predictor and the target explanation.

Theorem 1. *If a conditional probability distribution $P(Y \mid \mathbf{X})$ can be estimated accurately by maximizing a performance metric T on a learning algorithm h_Y, then $\mathbf{M} \subseteq \mathbf{X}$ is a target explanation of $P(Y \mid \mathbf{X})$ if and only if it is an optimal predictor of $P(Y \mid \mathbf{X})$.*

Proof of Theorem 1:
1. Prove a **TE** of $P(Y \mid \mathbf{X})$ is an optimal predictor of $P(Y \mid \mathbf{X})$: If $\mathbf{M} \subseteq \mathbf{X}$ is a target explanation of $P(Y \mid \mathbf{X})$, then $P(Y \mid \mathbf{X}) = P(Y \mid \mathbf{M})$ and this implies that T will be maximized on learning algorithm h_Y, therefore, $\mathbf{M}$ is an optimal predictor of $P(Y \mid \mathbf{X})$.

2. Prove an optimal predictor of $P(Y \mid \mathbf{X})$ is a **TE** of $P(Y \mid \mathbf{X})$: Suppose $\mathbf{M} \subseteq \mathbf{X}$ is an optimal predictor of $P(Y \mid \mathbf{X})$ but it is not a target explanation, so $P(Y \mid \mathbf{X}) \neq P(Y \mid \mathbf{M})$, and this implies $T_{h_Y(\mathbf{M})} > T_{h_Y(\mathbf{X})}$. By Definition 1, $\mathbf{X}$ is always a target explanation, thus it is also an optimal predictor of $P(Y \mid \mathbf{X})$. Therefore, the following should hold: $T_{h_Y(\mathbf{M})} = T_{h_Y(\mathbf{X})}$. This is contradiction. Therefore, $\mathbf{M}$ is a target explanation. ■

By the Definition 1 and Theorem 1, we get Corollary 3.7 to address the second problem of Identifying the instances from $\mathbb{D}$ which belong to this subpopulation.

Corollary 1. *If conditional probability distribution can be estimated accurately by maximizing a performance metric T on a learning algorithm h_Y, then the instances predicted correctly by h_Y are derived from the same distribution.*

Now we can use Theorem 1 as the criterion for detecting the **MTE** of $P_i(Y \mid \mathbf{X})$.

$$\begin{aligned} \mathbf{M}_i \quad &= \operatorname*{argmax}_{\mathbf{M}' \in f_Y(\mathbf{X})} \ T_{h_Y(\mathbf{M}')} \\ \text{s.t.} \quad & P_i(Y \mid \mathbf{X}) = h_Y(\mathbf{M}_i) \end{aligned} \tag{3}$$

Algorithm 2: The Galaxy algorithm, to discover the Galaxy space from mixture distributed feature data and forecast via its ensemble predictive power.

```
Input:
  • data set $\mathbb{D}$ includes $n$ instances; target feature $Y$; Markov boundary detection algorithm $f_Y$; learning algorithm $h_Y$;
    performance metric T;
Output:
  • Galaxy space $\mathbb{G}$.
  • Galaxy $H_Y$, an trained ensemble algorithm.
begin
    $\mathbb{G} = $ empty                                   /* Initialize $\mathbb{G}$ with an empty set */
    $\mathbf{W} = \prod_{j=1}^{n} w_j$, where $w_j = \frac{1}{n}$      /* Initialize the instances' weights $\mathbf{W}$ using uniform
     distribution */
    $\mathbb{I}(h_Y(x_j), y_j)$ /* Predicting error of the instance $(x_j, y_j)$, where $\mathbb{I} = 0$ if the prediction
     is correct, otherwise 1 */
    $i = 1$
    repeat
        $\mathbf{M}_i$, $h_Y(\mathbf{M}_i)$ = PMTE Detection($\mathbb{D}$, $Y$, $f_Y$, $h_Y$, $T$)
        $\epsilon = \frac{\sum_{j=1}^{n} w_j \mathbb{I}(h_Y(x_j), y_j)}{\sum_{j=1}^{n} w_j}$      /* computer the weighted misclassification rate $\epsilon$. */
        $\phi_i = \log\left(\frac{1-\epsilon}{\epsilon}\right)$                  /* computer the mixture component weight $\phi$. */
        for $w_j \in \mathbf{W}$ do
            $w_j = w_j \exp(\epsilon \mathbb{I}(h_Y(x_j), y_j))$          /* strengthen the misclassified instances. */
        $i = i + 1$
    until $\epsilon < \delta$
    Return $\mathbb{G} = \prod_i \mathbf{M}_i$, $H_Y = \sum_i \phi_i h_Y(\mathbf{M_i})$
```

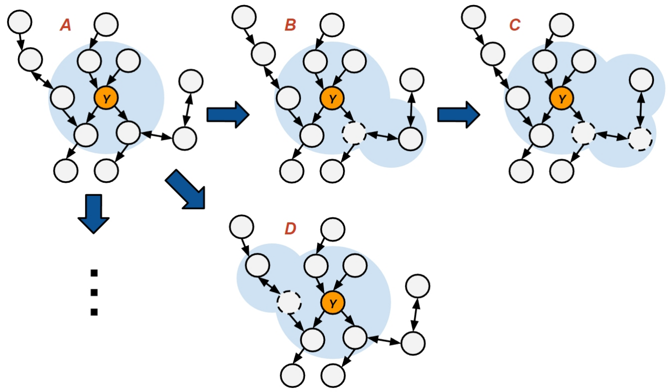

Fig. 2. Multiple Markov boundaries detection via equivalent information exploration. A, B, C, D are Markov boundaries of Y in a Bayesian network. A is the Markov boundary detected from the original feature space. B, C, D are Markov boundaries generated by replacing part of features in A by equivalent features explored from the residuals.

Here $f_Y : \mathbb{R}^k \to \mathbb{R}^d$ is a Markov boundaries detection algorithm, $d \leq k$, $h_Y : \mathbb{R}^d \to \mathbb{R}$ is a learning algorithm for predicting Y, and T is a performance metric to assess h_Y (the bigger the value, the better the performance). For f_Y, we utilize the idea mentioned in [11], which firstly detects a Markov boundary $\mathbf{M}$ from $\mathbf{X}$, and then tries to replace part of $\mathbf{M}$ by its equivalent feature sets explored from the residual features. Finally a **PMTE** of $P(Y \mid \mathbf{M})$, which is the **MTE** of $P_i(Y \mid \mathbf{M})$, can be detected by choosing the optimal Markov boundary $\mathbf{M}_i$ which maximizes the performance metric T for predicting Y using learner h_Y. The detail is illustrated in Fig 2 and Algorithm 1.

D. Galaxy Space Detection

The probability distribution of Y in the overall population is represented as a mixture distribution, then detecting the Galaxy space of Y is actually looking for the explanations which globally faithful to the mixture distribution of Y. It can be implemented by detecting every subpopulation's **MTE** in the mixture distribution of Y via Theorem 1. Since every **PMTE** in the Galaxy space is locally faithful to a subpopulation of Y, the overall explanation of Y can be represented as the ensemble of the Galaxy space. Thus, we give the definition of Galaxy predictor and then link it with the concept of Galaxy space.

Definition 6. ***Galaxy Predictor****: Given a data set $\mathbb{D}$, we say a family of feature subsets $\prod_i^m \mathbf{M}_i$, where $\mathbf{M}_i \subseteq \mathbf{X}$, is a Galaxy predictor of Y if it maximizes the performance metric T for predicting Y using an ensemble learning algorithm H_Y.*

We call the ensemble learning algorithm H_Y on Galaxy predictor as Galaxy. The following theorem provides the link between the Galaxy predictor and the Galaxy space.

TABLE I

Average F-measure on different dimensional synthetic datasets.

Dimensionality Classifier	450	550	650	750	850	950	1050	1150	1250	1350
Random Forest	0.812	0.777	0.761	0.671	0.707	0.672	**0.776**	0.753	0.662	0.764
AdaBoost	0.794	0.786	0.755	0.643	0.701	0.587	0.773	0.733	0.591	0.734
Gradient Boosting	0.791	0.842	0.792	**0.716**	0.691	0.612	0.781	0.744	0.652	0.758
Multilayer Perceptron	0.819	0.845	0.832	0.703	0.706	0.711	0.726	0.742	0.643	0.802
Galaxy(our method)	**0.82**	**0.847**	**0.841**	0.714	**0.723**	**0.803**	0.772	**0.804**	**0.683**	**0.824**

Theorem 2. *A family of feature subsets $\prod_i^m \mathbf{M}_i$, where $\mathbf{M}_i \subseteq \mathbf{X}$, is a Galaxy space of Y if and only if it is an Galaxy predictor of Y.*

Proof of Theorem 2:

1. Prove a Galaxy space of Y is a Galaxy predictor of Y:

If a family of feature subsets $\prod_i^m \mathbf{M}_i$, where $\mathbf{M}_i \subseteq \mathbf{X}$, is a Galaxy space of Y, then by Definition 4, every $\mathbf{M}_i$ corresponds a **PMTE** of $P_i(Y \mid \mathbf{X})$ in $\sum_{i=1}^m \phi_i P_i(Y \mid \mathbf{X})$. This implies that $\mathbf{M}_i$ is an optimal predictor of $P_i(Y \mid \mathbf{X})$ by maximizing the performance metric T on a learning algorithm h_Y. Therefore, $\prod_i^m \mathbf{M}_i$ is a Galaxy predictor and Galaxy is presented as $H_Y = \sum_{i=1}^m \phi_i h_Y(\mathbf{M_i})$.

2. Prove a Galaxy predictor of Y is a Galaxy space of Y:

Suppose $\prod_i^m \mathbf{M}_i$ is a Galaxy predictor of Y, but it is not a Galaxy space of Y and Galaxy is presented as $H_Y = \sum_{i=1}^m \phi_i h_Y(\mathbf{M_i})$, so there is at least one $\mathbf{M}_j \in \prod_i^m \mathbf{M}_i$ is not a **PMTE**. This implies that $\mathbf{M}_j$ is not an optimal predictor. so there exist an optimal predictor to make the performance of Galaxy better. This is contradict to that $\prod_i^m \mathbf{M}_i$ is a Galaxy predictor. Therefore $\prod_i^m \mathbf{M}_i$ is a Galaxy space of Y. ■

Based on Theorem 2 and Definition 6, we can detect a Galaxy space of Y via Galaxy, a variant of AdaBoost, as follows.

- **Step 1.** Detect a **PMTE** via Theorem 1 on weighted instances in $\mathbb{D}$ and then calculate the misclassification rate ϵ.
- **Step 2.** Compute mixture component weight.
$$\phi_i = \log\left(\frac{1-\epsilon}{\epsilon}\right). \tag{4}$$
- **Step 3.** Strengthen the misclassified instances by reweighting the misclassified instances.
$$w_j = w_j e^{\epsilon \mathbb{I}(h_Y(x_j), y_j)}. \tag{5}$$
where w_j is the weight of the instance (x_j, y_j), $\mathbb{I}$ is the predicting error of (x_j, y_j), where $\mathbb{I} = 0$ if the prediction is correct, otherwise 1.
- Repeat steps 1 - 3 until the misclassification rate ϵ lower than a threshold δ.

The Galaxy algorithm, to discover the Galaxy space from mixture distributed feature data, is explained in pseudo-code in Algorithm 2.

IV. Experimental Evaluation

In this section, we present experiments to evaluate the effectiveness and utility of explanations of Galaxy on synthetic data with different dimensionalities and a real-world precipitation data set. In particular, we address the following questions:

- **Q1. Effectiveness on highly correlated data:** How effective can Galaxy work on highly correlated data?
- **Q2. Interpretation of target explanations:** Are the target explanations detected by Galaxy on real-world data interpretable?

We implemented Galaxy in Python; all experiments were carried out on a 3.0 GHz Intel(R) Xeon(R) E5-2687 Linux server, 1007 GB RAM, running Ubuntu 16.04.2 LTS.

A. Synthetic data generation

In order to simulate highly correlated data that represent data collected in real-world climate applications, we generate a d-dimensional synthetic data set using classification data generator in Python package scikit-learn [15] with high redundancy and noise.

B. Experiment settings

We demonstrate the effectiveness of Galaxy by comparing its F-measure($\frac{2\times\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}$) against a bench of candidate methods: Random Forest, AdaBoost, Gradient Boosting, and multilayer perceptron. Simulated data were generated with feature counts d ranging from 450 to 1,350 in increments of 100 features. Each of the comparison methods was run against each of the subsets of the overall dataset after feature reduction. All computations were performed on the same hardware and datasets.

C. Q1. Effectiveness on highly correlated data

We report the F-measure for each classifier on the different-dimensional datasets, averaged by 10-fold cross-validation, in Table I. We can see that Galaxy outperforms others most often (8 wins, 2 losses). These results indicate that Galaxy achieve the satisfying predictive power while detecting the Galaxy space.

D. The Precipitation Data Set

The real-world dataset we used for the experiments is a subset of the NCEP/NCAR Reanalysis dataset [16] and includes 9 meteorological variables collected at different vertical levels in the atmosphere (Table II). All the variables are chosen by our domain scientists collaborators based on their physical relevance for precipitation analysis. By convention, atmospheric pressure (in units of hectopascals or hPa) is used as the vertical coordinate with the 1000hPa surface located

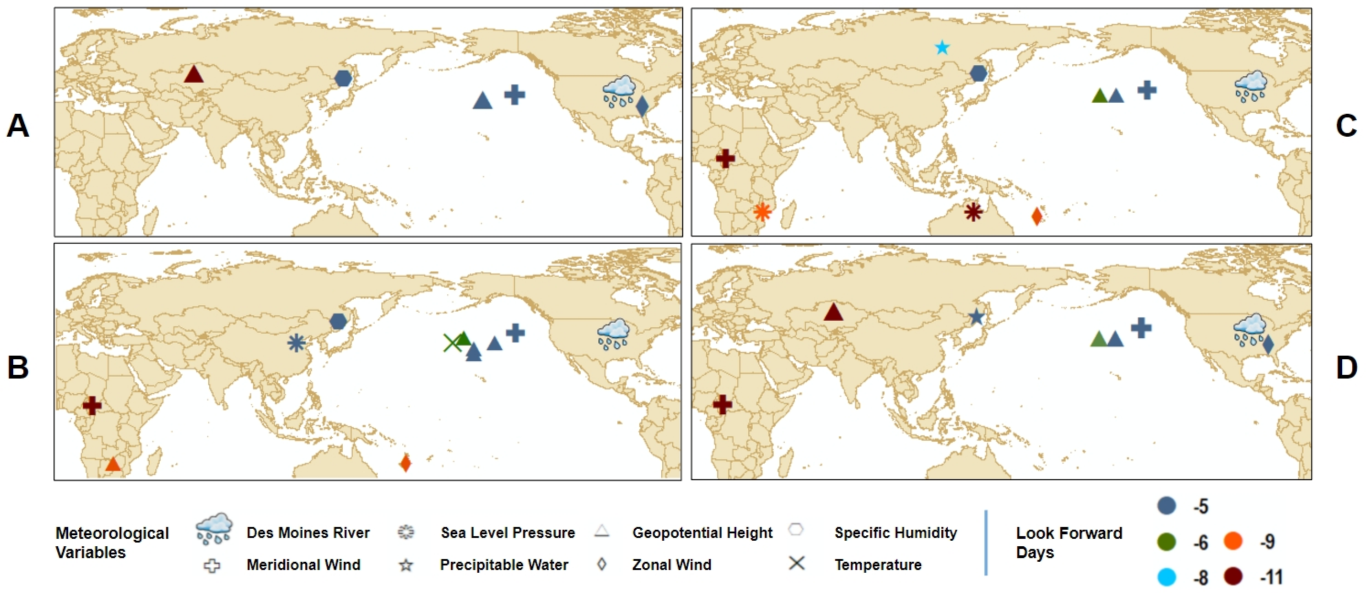

Fig. 3. Four PMTEs of the precipitation at Des Moines river basin detected by Galaxy.

TABLE II
METEOROLOGICAL VARIABLES.

Name	Level(hPa)
Zonal Wind	200, 500, 850
Meridional Wind	200, 500, 850
Geopotential Height	200, 500, 850
Temperature	200, 500, 850
Relative Humidity	700, 925
Specific Humidity	850
Pressure Vertical Velocity	700
Sea Level Pressure	-
Precipitable Water	-

near the surface and 200hPa near the top of the troposphere. According to the theory of quasi-geostrophic and baroclinic [17], we specially choose 200hPa, 500hPa, and 850hPa zonal winds(i.e. east-west) because they are a proxy for the location and strength of the jet stream which requires wind shear (strong change in wind speed with height) to develop. And the information of the location of the jet stream exhibits persistence on scales much longer than individual storm events. Moreover, 200hPa, 500hPa, and 850hPa 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 200hPa, 500hPa, and 850hPa 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 200hPa and 850hPa fields allows us to infer where large-scale rising motion (and therefore precipitation) is likely to take place. On the other hand, the temperature at 200hPa, 500hPa, and 850hPa 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. And specific humidity at 850hPa, relative humidity at 700hPa and 925hPa are chosen because the amount of water in the upper troposphere was thought to be negligible. The pressure vertical velocity at 700hPa, precipitable water (total water vapor integrated from the surface to the top of the atmosphere) and sea level pressure (atmospheric pressure at surface corrected to sea level) are also important in producing precipitation.

The total number of variables in all levels is 18. All these meteorological variables are sampled at the spatial domain of $0°E$ to $375.5°E$ and $90°N$ to $20°S$ with a resolution of $2.5° \times 2.5°$ (totally 5,904 locations) and a daily temporal resolution. We pick the samples collected during the rainy season (March to November) during the years 1951-2017. The target feature is the historical spatial average precipitation (the mean of daily precipitation totals from 23 stations) of the Des Moines River basin in Iowa from the same time period.

In the experiments, We set the lead time as 5 days, “look ahead period as 10 days. For example, to explain rainfall situations at today (Day_0) in the study area, we will look for the explanatory features in the time period from Day_{-14}(14 days ago) to Day_{-5}(five day ago). The precipitation data set presents two particularly difficult characteristics:

- **Extremely high dimensionality:** Each sample has $5,313,600$ features (18 variables × 5,904 locations × 10 days)
- **High intra-dataset correlation:** Meteorological variables presented at different levels, locations, and days are highly correlated. Different meteorological variables may also correlated.

E. Q2. Interpretation of target explanations

We run Galaxy on this extremely high dimensionality data set and finally got 15 **PMTE**s, which the minimum size is 4, the maximum size is 13. The top four weighted **PMTE**s are illustrated in Figure 3.

The **PMTE**s in Figure 3 (A) and Figure 3 (D) identifies a geopotential height anomaly over Eastern Europe at 11 days before. This is consistent with a deepening trough over Ural Mountains. Troughs deepening over the Urals are often triggers of wave trains across Asia (the so-called Silk Road pattern) that eventually end up propagating across the Pacific.

The **PMTE** in Figure 3 (B) includes meridional wind, specific humidity, and upper level (200hPa) geopotential height fields near the east coast of Asia. Many studies have identified cold surges along the Asian coast as important precursors to surface weather over the United States [18], [19]. The surge of cold air and deepening trough typically result in a strengthened jet stream and generate a Rossby wave that propagates across the North Pacific and breaks along the west coast of North America. The mechanism implied by the **PMTE** is that the North-South winds are transporting dry (and presumably cold) air from the north into the middle latitudes around Japan. This pattern results in the deepening of an upper-level trough (negative anomaly in the 200hPa Geopotential Height field) and strengthening of the mid-latitude jet stream. The strengthening of the jet stream in the **PMTE** (i.e. an upper level zonal (u) wind field being chose) was not observed, but it is implied. The upper-level geopotential height anomalies along the west coast of North America on day -5 imply a large, pre-existing upper-level ridge along the west coast of North America [20], [21] that is also very consistent with expectations of strong precipitation over the central U.S. This pattern suggests a "forcing" for a wave train setting up along the east coast and a pre-existing ridge along the west coast.

V. Conclusions

We propose a new interpretable predictive model Galaxy that can represent efficient explanations of subpopulations within an overall population of the target feature, and forecast target by their ensemble predictive power. We provide a theoretical framework and implementation details of Galaxy. Our empirical study on the synthetic and real data demonstrate the superb performance of Galaxy on predication and interpretation.

Acknowledgment

Funding for this research was provided in part by the National Science Foundation grant 1743010, Oracle Education Foundation grant to the College of Science and Mathematics at the University of Massachusetts Boston and the National Science Foundation of China under Grant 61876206.

References

[1] T. Wang, P. Chen, and B. Li, "Predicting the quality of short narratives from social media," *arXiv preprint arXiv:1707.02499*, 2017.

[2] J. Kawale, S. Chatterjee, D. Ormsby, K. Steinhaeuser, S. Liess, and V. Kumar, "Testing the significance of spatio-temporal teleconnection patterns," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 642–650.

[3] K. Hua and D. A. Simovici, "Long-lead term precipitation forecasting by hierarchical clustering-based bayesian structural vector autoregression," in *Networking, Sensing, and Control (ICNSC), 2016 IEEE 13th International Conference on*. IEEE, 2016, pp. 1–6.

[4] Y. Zhuang, K. Yu, D. Wang, and W. Ding, "An evaluation of big data analytics in feature selection for long-lead extreme floods forecasting," in *Networking, Sensing, and Control (ICNSC), 2016 IEEE 13th International Conference on*. IEEE, 2016, pp. 1–6.

[5] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proceedings of the 20th international conference on machine learning (ICML-03)*, 2003, pp. 856–863.

[6] C. P. d. Campos and Q. Ji, "Efficient structure learning of bayesian networks using constraints," *Journal of Machine Learning Research*, vol. 12, no. Mar, pp. 663–689, 2011.

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

[8] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

[9] C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos, "Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation," *Journal of Machine Learning Research*, vol. 11, no. Jan, pp. 171–234, 2010.

[10] J. D. Nielsen, T. Kočka, and J. M. Peña, "On local optima in learning bayesian networks," in *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'03. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003, pp. 435–442. [Online]. Available: http://dl.acm.org/citation.cfm?id=2100584.2100637

[11] A. Statnikov, N. I. Lytkin, J. Lemeire, and C. F. Aliferis, "Algorithms for discovery of multiple markov boundaries," *Journal of Machine Learning Research*, vol. 14, no. Feb, pp. 499–566, 2013.

[12] J. A. Smith, G. Villarini, and M. L. Baeck, "Mixture distributions and the hydroclimatology of extreme rainfall and flooding in the eastern united states," *Journal of Hydrometeorology*, vol. 12, no. 2, pp. 294–309, 2011.

[13] C. F. Aliferis, I. Tsamardinos, and A. Statnikov, "Hiton: a novel markov blanket algorithm for optimal variable selection," in *AMIA Annual Symposium Proceedings*, vol. 2003. American Medical Informatics Association, 2003, p. 21.

[14] I. Tsamardinos and C. F. Aliferis, "Towards principled feature selection: relevancy, filters and wrappers." in *AISTATS*, 2003.

[15] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[16] 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.

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

[18] R. M. Dole, "Persistent anomalies of the extratropical northern hemisphere wintertime circulation: Structure," *Monthly weather review*, vol. 114, no. 1, pp. 178–207, 1986.

[19] R. M. Dole and N. D. Gordon, "Persistent anomalies of the extratropical northern hemisphere wintertime circulation: Geographical distribution and regional persistence characteristics," *Monthly Weather Review*, vol. 111, no. 8, pp. 1567–1586, 1983.

[20] P. A. Harr and J. M. Dea, "Downstream development associated with the extratropical transition of tropical cyclones over the western north pacific," *Monthly Weather Review*, vol. 137, no. 4, pp. 1295–1319, 2009.

[21] D. Small, E. Atallah, and J. R. Gyakum, "An objectively determined blocking index and its northern hemisphere climatology," *Journal of Climate*, vol. 27, no. 8, pp. 2948–2970, 2013.