Stat 486 Final Project Report

1st Matthew Jensen
Brigham Young University
Stat 486
mj638@byu.edu

2nd Aaron Brown
Brigham Young University
Stat 486
abrown52@byu.edu

3rd Matthew Blackley
Brigham Young University
Stat 486
mtb77@byu.edu

I. Introduction

For our final project, we conducted an analysis on snow accumulation patterns spanning several decades in Utah. Specifically, our study utilized data initially gathered in the year 1979, encompassing observations through 2023. This dataset was sourced from the Natural Resources Conservation Service (NRCS, USDA) Public Database (refer to Exploratory Data Analysis and Collection). Our aim was to discern trends in snow accumulation over time and ascertain whether there are significant indications of declining snow levels across the state, and if so, determine an effective predictive model to aid in predicting for the expected snow accumulation totals near the end of the season, which will prove useful for local communities in estimating the environmental conditions for the upcoming year. In doing so, we engineered new features, constructed and compared an assortment of machine learning models to predict snow levels for each April (the month typically experiencing annual peak levels), implemented a neural network and attempted to predict April's snowpack through deep learning, and applied SHAP analysis to determine feature importance.

II. Exploratory Data Analysis and Collection

As previously stated, our dataset was collected from the NRCS Public Database. The site contained a table for each operational Snowpack Telemetry (SNOTEL) station in the state of Utah. Each station in the table included a text file of their respective observations from their initiatory year to either last year (2023) or the year prior to their deinstallation. Fig. 1 displays the main table used to collect the snowpack observations from each SNOTEL site.

To collect each text file from the table, we utilized Python's Selenium package to iterate through each link and web scrape the data. Each text file shared a similar structure to the following:

The datasets were organized such that each row represented a year of snow collection, with individual observations included for January, February, March, April, May, and June. The observations were measurements (inches) of the current snow level at the beginning of the specified month. In addition, the water equivalence (WE, inches) was estimated and collected during each month and included in the text data. However, for certain SNOTEL sites, the only data collected was on the observed WE.

Monthly Snow Data

View	Site_Name	Station	Ntwk	Elev	Lat	Lon	installed	state	County
Current Historic	Agua Canyon	12M26S	SNTL	8900	37.52	-112.27	1995	UT	Kane
Current Historic	Alta Central	11J61	SNOW	8800	40.59	-111.64	1984	UT	Salt Lake
Current Historic	Atwater		SNTL	8753	40.59	-111.64	2022	UT	Salt Lake
Current Historic	Bear River RS	10J17S	SNTL	8777	40.89	-110.83	2011	UT	Summit
Current Historic	Beaver Dams	11K13S	SNTL	7990	39.14	-111.56	1980	UT	Sanpete
Current Historic	Beaver Divide	11J46S	SNTL	8280	40.61	-111.10	1979	UT	Summit
Current Historic	Ben Lomond Peak	11H08S	SNTL	7689	41.38	-111.94	1979	UT	Weber
Current Historic	Ben Lomond Trail	11H30S	SNTL	5972	41.38	-111.92	1980	UT	Weber
Current Historic	Bevans Cabin	12J02S	SNTL	6520	40.46	-112.25	2013	UT	Tooele
Current Historic	Big Flat	12L07S	SNTL	10349	38.30	-112.36	1979	UT	Beaver
Current Historic	Birch Crossing	12M16	SNOW	8045	37.74	-112.84	1965	UT	Iron
Current Historic	Black Flat-U.M. Ck	11L04S	SNTL	9414	38.68	-111.60	1980	UT	Sevier
Current Historic	Blacks Fork Jct	10J22S	SNTL	8869	40.96	-110.58	2010	UT	Summit
Current Historic	Bobs Hollow	12L16S	SNTL	9319	38.95	-112.15	2017	UT	Millard
Current Historic	Box Creek	12L04S	SNTL	9853	38.51	-112.02	1979	UT	Sevier
Current Historic	Box Springs	12L05S	SNTL	9228	38.50	-112.01	2011	UT	Piute
Current Historic	Brian Head	12M14S	SNTL	10039	37.68	-112.86	2012	UT	Iron
Current Historic	Brighton	11J57S	SNTL	8766	40.60	-111.58	1986	UT	Salt Lake
Current Historic	Brown Duck	10J30S	SNTL	10574	40.58	-110.59	1979	UT	Duchesne

Fig. 1. NRCS Database

```
#-------------------------------------------------------------------------------
#
# Agua Canyon (907)
# Utah  SNOTEL Site - 8900 ft
# Reporting Frequency: Monthly; Date Range: Period of Record
#
# As of: Apr 13, 2024, 9:45:00 AM GMT-08:00
#
Water Year,Jan,Jan,Jan,Feb,Feb,Feb,Mar,Mar,Mar,Apr,Apr,Apr,May,May,May,Jun,Jun,Jun
,Snow Water Equivalent Collection Date Start of Month Values,Snow Depth (in) Start of Month Values,Snow Water Equivalent (in)
Start of Month Values,Snow Water Equivalent Collection Date Start of Month Values,Snow Depth (in) Start of Month Values,Snow Water
Equivalent (in) Start of Month Values,Snow Water Equivalent Collection Date Start of Month Values,Snow Depth (in) Start of Month
Values,Snow Water Equivalent (in) Start of Month Values,Snow Water Equivalent Collection Date Start of Month Values,Snow Depth
(in) Start of Month Values,Snow Water Equivalent (in) Start of Month Values,Snow Water Equivalent Collection Date Start of Month
Values,Snow Depth (in) Start of Month Values,Snow Water Equivalent (in) Start of Month Values,Snow Water Equivalent Collection
Date Start of Month Values,Snow Depth (in) Start of Month Values,Snow Water Equivalent (in) Start of Month Values
1995,,,4.7,,,10.5,,,12.5,,,15.5,,,5.8,,,0.0
1996,,,0.9,,,2.8,,,5.2,,,2.4,,,0.0,,,0.0
1997,,,3.4,,,7.3,,,8.5,,,2.0,,,0.0,,,0.0
1998,,,2.4,,,4.2,,,8.3,,,9.5,,,4.0,,,0.0
1999,,12,3.4,,14,4.4,,15,5.2,,2,0.0,,0,0.0,,0,0.0
2000,,1,0.1,,1,0.4,,20,4.8,,17,6.2,,0,0.0,,0,0.0
2001,,12,2.4,,32,6.7,,41,10.7,,22,9.0,,0,0.0,,0,0.0
2002,,11,1.8,,14,2.4,,9,2.5,,0,0.0,,0,0.0,,0,0.0
2003,,9,1.4,,5,1.7,,26,4.6,,8,2.5,,0,0.0,,0,0.0
2004,,20,4.3,,23,5.3,,50,10.9,,8,2.2,,0,0.0,,0,0.0
2005,,35,7.4,,57,15.9,,64,19.9,,65,22.4,,30,12.8,,0,0.0
2006,,7,1.4,,16,3.5,,15,4.8,,28,8.8,,0,0.0,,0,0.0
2007,,12,2.7,,17,3.8,,24,5.4,,0,0.0,,0,0.0,,0,0.0
2008,,18,3.8,,45,10.6,,43,13.3,,27,10.2,,0,0.0,,0,0.0
2009,,27,7.2,,22,7.2,,27,9.1,,0,0.0,,0,0.0,,0,0.0
2010,,21,4.6,,45,10.7,,43,12.5,,39,13.8,,1,0.2,,0,0.0
2011,,40,9.5,,31,9.5,,35,13.9,,38,14.3,,1,0.1,,0,0.0
2012,,13,2.9,,19,4.2,,27,6.1,,0,0.0,,0,0.0,,0,0.0
2013,,17,3.9,,21,5.6,,23,6.1,,4,1.5,,0,0.0,,0,0.0
2014,,19,4.7,,26,5.1,,13,4.2,,0,0.0,,0,0.0,,0,0.0
2015,,20,2.1,,19,4.5,,29,6.4,,5,1.8,,0,0.0,,0,0.0
2016,,14,3.2,,34,7.1,,15,5.5,,3,0.9,,1,0.2,,0,0.0
2017,,17,4.0,,37,9.7,,35,11.6,,8,2.2,,0,0.0,,0,0.0
2018,,0,0.0,,5,1.3,,8,1.6,,0,0.0,,0,0.0,,0,0.0
2019,,13,2.3,,21,5.1,,43,12.1,,38,14.9,,1,0.5,,0,0.0
2020,,26,6.6,,14,6.3,,20,7.0,,28,8.5,,0,0.0,,0,0.0
2021,,6,0.8,,15,3.2,,15,4.0,,10,4.1,,0,0.0,,0,0.0
2022,,27,5.8,,20,5.5,,24,6.7,,10,3.6,,0,0.0,,0,0.0
2023,,17,4.5,,34,9.9,,49,14.1,,62,21.4,,11,5.2,,0,0.0
```

Fig. 2. SNOTEL Text File

After web scraping each text file, we converted the text into datasets and compiled them together into a main dataset. The EDA-specific feature engineering involved creating a new variable "Decade" which represented the decade when the observation was recorded. As previously stated, the datasets of certain SNOTEL sites only included the respective WE levels for each month, and so we estimated the snow levels by creating a "Snow/WE" coefficient, which was computed by averaging the fraction of snowfall over respective WE levels for each month, January through April, for each row in the dataset where snowfall and WE were both included. We multiplied the observed WE levels by this coefficient to

estimate the snow levels for each site which failed to include snow level observations.

Site_Name	Elev	Lat	Lon	Installed	County	Water Year	Jan	Jan (WE)	Feb	Feb (WE)	Mar	Mar (WE)	Apr	Apr (WE)	May	May (WE)	Jun	Jun (WE)	Decade
Agua Canyon	8,900	37.52	-112.27	1995	Kane	1995	15.3731	4.7	34.3442	10.5	40.8859	12.5	50.6986	15.5	18.9731	5.8	0	0	1990
Agua Canyon	8,900	37.52	-112.27	1995	Kane	1996	2.9438	0.9	9.1584	2.8	17.0085	5.2	7.8501	2.4	0	0	0	0	1990
Agua Canyon	8,900	37.52	-112.27	1995	Kane	1997	11.121	3.4	23.8774	7.3	27.8024	8.5	6.5417	2	0	0	0	0	1990
Agua Canyon	8,900	37.52	-112.27	1995	Kane	1998	7.8501	2.4	13.7377	4.2	27.1483	8.3	31.0733	9.5	13.0835	4	0	0	1990
Agua Canyon	8,900	37.52	-112.27	1995	Kane	1999	12	3.4	14	4.4	15	5.2	2	0	0	0	0	0	1990

Fig. 3. Site and Snow Main Dataset

Figure 3 displays a portion of the cleaned dataset we used in our EDA.

EDA:

After cleaning and compiling our dataset, we proceeded to explore the data.

Our main goal in our EDA was to determine whether the time the observations were collected (more specifically, the observed year) had a significant effect on the snow levels measured by month. In addition, our goal in our EDA was to distinguish a particular month of each year which best represented the peak snow accumulation level on average. In doing so, we could predict for that particular month for a given year to estimate what the end results for that winter will be.

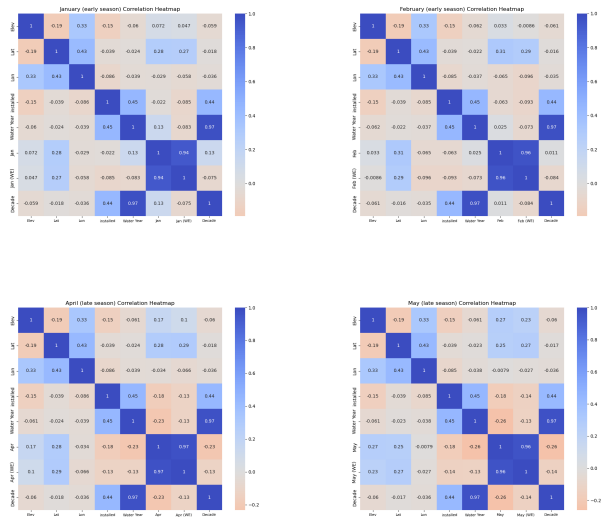

Fig. 4. Heatmaps for Jan, Feb, Apr, May

The correlation heatmaps in Figure 4 depict each monthly snowpack factor and its correlation with other variables. While there are positive correlations between snowpack/WE factors and time related variables for Jan and Feb, there appears to be a general trend of decreasing correlation between the monthly snowpack/water equivalent factors with decade, water year, and year installed. Of course, the correlations indicated from these plots are moderate at best, and will thus have to be explored further before we form conclusions about trends in snowpack over time.

The plot shown in Figure 5 displays the distribution of snowpack measurements for each month, ordered by decade

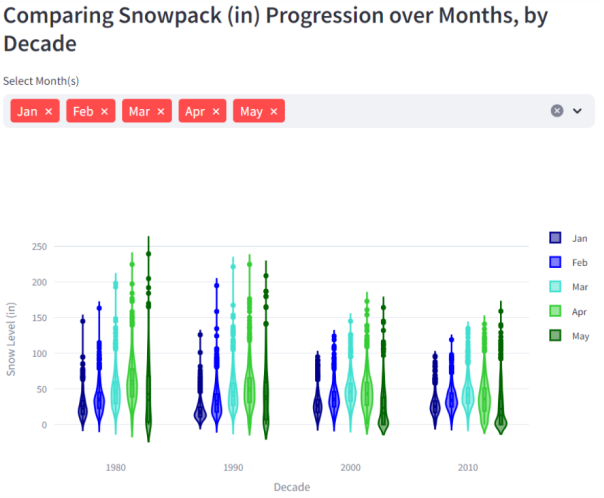

Fig. 5. Snow Patterns Over Months by Decade

(1980s through 2010s). While the months of Jan and Feb may see slightly higher averages in later decades, there is evidence of a general decline in snowpack for the majority of months, particularly in the late season (March through May). More than anything, our exploratory analysis of this data provides evidence of a "shortening season", since earlier months like Jan and Feb do not see much of a change in snowpack, while later months see a noticeable decline.

Given these trends, we find it crucial to develop a method of predicting levels of snowpack for coming years; that way, local communities may be informed on the expected results for a specific winter and prepare for the coming months accordingly.

Based on these plots, we also conclude that the month of April best represents the annual "climax" or resulting build-up of snowpack for the SNOTEL sites in Utah since for each decade, it is the distribution with either the highest or near-highest average snowpack.

After identifying April as the variable of focus for prediction, we tested the effect of "Water Year" on April's snow measurements just to verify that the year of the data collection has a statistically significant impact on April's observed values. To do so, we developed a simple linear regression model with "Apr" as a response and all other variables, other than April's WE levels, as features.

TABLE I
APRIL MEASUREMENTS OLS REGRESSION SUMMARY

	Coef	t	P> |t|	0.025	0.975
Intrcpt	0.3839	4.708	0.000	0.224	0.544
...	...	...	...	...	...
Water Year	-0.1828	-9.743	0.000	-0.220	-0.146

The results are statistically significant and lead us to reject the null hypothesis of no effect. Thus, we conclude that the collection year has a significant as well as negative effect on April's snow level measurements. This further confirms our need to predict for future snow levels to better inform

local communities of required courses of action for water preservation during seasons of low snowpack. However, in testing these features, it is important to be aware of both the geographical and time aspects of the data when considering correlation between observations. An implication for future research is to explore other possible modeling methods to better represent the data, including longitudinal regression to account for within-site correlation.

III. Methods

After deciding on April snow measurements as the variable of focus for predicting snow accumulation levels for SNOTEL sites in Utah, we moved onto feature engineering and comparing machine learning models and their predictive abilities.

Feature Engineering:

We developed a couple of new features and tested them with an assortment of machine learning models.

FE 1: The first one we tested was a lagged variable which took the snow and WE observations from the previous five years and calculated their averages; this was used for all months, January through June. While this new feature captured previous observations for the later season (April through June) and slightly improved our metrics for the tree-based models, it did not lead to significant improvements relative to our other feature engineering method.

FE 2: Our second new feature was not lagged and was tested separately from the first new feature. This feature was a ratio of the snow measurements from January, February, and March to the Elevation of the SNOTEL site (measured in feet). After reinterpreting these established features with a new variable, our results improved significantly across many of our models.

The following list of machine learning models were tested (it is indicated which new features were used with each model).

Methods:

- **Random Forest, FE 1**
  - **Description:** RF regression model using first new feature.
  - **Hyperparameters Tested:**
    * `n_estimators`
    * `max_depth`
  - **Results:** MSE: 121.11, RMSE: 11.00, MAE: 7.43, R-Squared: 0.86, Training Accuracy: 0.99
- **Boosted Trees (XGBoost), FE 2**
  - **Description:** Boosted Trees regression model (XGBoost) using the first new feature.
  - **Hyperparameters Tested:**
    * `n_estimators`
    * `max_depth`
  - **Results:** MSE: 74.95, RMSE: 8.66, MAE: 6.33, R-Squared: 0.91, Training Accuracy: .99
- **Random Forest Regressor, FE 2**
  - **Description:** Random Forest regression model using the second created feature.
  - **Hyperparameters Tested:**
    * `n_estimators`
    * `max_depth`
  - **Results:** MSE: 74.86, RMSE: 8.65, MAE: 6.36 R-Squared: 0.91, Training Accuracy: .98
- **Lasso Regression**
  - **Description:** Linear Regression model using the second created feature and a L1 penalty term.
  - **Hyperparameters Tested:**
    * `max_iter`
    * `alpha`
    * `fit_intercept`
  - **Results:** MSE: 125.58, RMSE: 11.21, MAE: 8.23, R-squared: 0.85, Training Accuracy: .86
- **Support Vector Machine Regressor, FE 2**
  - **Description:** A SVM regression model using the second created feature.
  - **Hyperparameters Tested:**
    * `kernel`
    * `c`
    * `degree`
  - **Results:** MSE: 128.44, RMSE: 11.33, MAE: 8.17, R-squared: 0.84, Training Accuracy: .82
- **KNN Regressor, FE 2**
  - **Description:** A KNN regression model using the second created feature.
  - **Hyperparameters Tested:**
    * `n_neighbors`
    * `c`
    * `degree`
  - **Results:** MSE: 118.57, RMSE: 10.8, MAE: 7.89, R-squared: 0.85, Training Accuracy: 1
- **Neural Network, FE 2**
  - **Description:** A Neural Network model using the second created feature.
  - **Hyperparameters Tested:**
    * `batch_size`
    * `activation functions`
    * `num_neurons`
    * `neuron_density`
    * `learning_rate`
  - **Results:** MSE: 160.51, RMSE: 12.7, MAE: 9.4, R-squared: .82

IV. Discussion on Model Selection

The model selection process involved testing a variety of machine learning models to determine which ones performed best for predicting April snow levels in Utah. We considered models such as Random Forest Regressor, XGBoost, Lasso

Regression, Support Vector Machine Regressor, KNN Regressor, and Neural Networks.

At this stage of our analysis, we used a combination of cross-validation and plugging in random values for hyperparameters to tune our models. We evaluated the models based on metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared. Based on these metrics, we generally saw that our tree-based models, such as Random Forest and XGBoost, consistently outperformed our Lasso, KNN, and SVM regression models. This is a sign that the data is non-linear and our tree-based models captured more complex relationships in the data better than our linear models. There was also worry that based on our metrics, some of our linear models, such as KNN, were overfitting. They did well on the training data but not on the testing data and had a larger RMSE and MSE, so we did not consider a linear model for our final model selection.

We also implemented a Neural Networks model to test how well a deep learning model would do on the data. This model was our most computationally intensive, had the worst metrics, and took the longest to run. Tuning the hyperparameters gave marginal improvements in metrics so we decided that a Neural Network model was not ideal for our dataset.

Based on metrics and computation time, it became obvious to us that tree-based models were outperforming our linear and deep learning models, so we decided to use a Random Forest Model as our final model.

V. Best Model

Across all the models we tested, our best-performing model was the Random Forest Regressor that incorporated FE 2 (the snowpack to elevation ratio feature). We decided to use this specific model as the foundation for our final model and found that optimal hyperparameter tuning only involved searching through values for `n_estimators` and `max_depth`. Because we did not initially cross-validate this Random Forest model, we tuned its hyperparameters using cross-validation. The best performing parameters were n_estimators: 500, max_depth: 20. The performance metrics for this model were MSE: 74.86, RMSE: 8.65, MAE: 6.36 R-Squared: 0.91, Training Accuracy: .98.

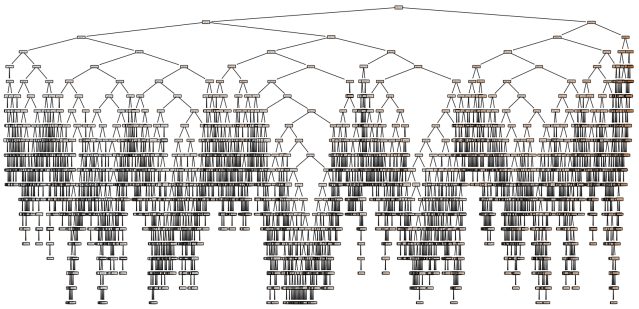

Fig. 6. Visualization of Final Model (Random Forest, FE 2)

In addition, because the categorical features "Site_Name" and "County" included 131 and 29 levels respectively, we decided to remove these features to perform our SHAP analysis in a computationally efficient way, thus avoiding the issues involved with high dimensional data. The performance metrics of the model after removing these features are the ones previously mentioned in this section.

In determining the feature importances of our RF model, we chose to run an analysis on the SHAP values of each feature. To start, we compared the features of the model based on their respective Random Forest Global Feature Importances:

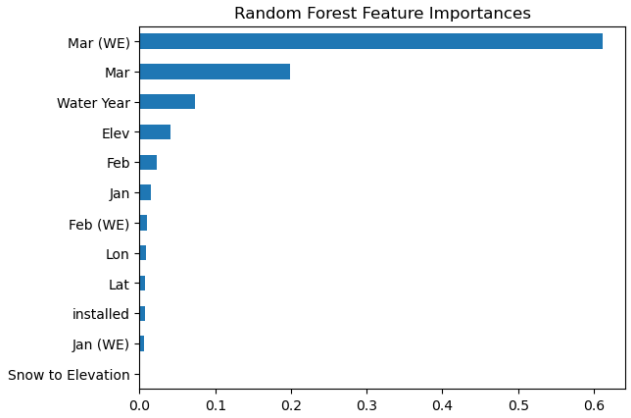

Fig. 7. RF Global Feature Importances

The RF Global Feature Importances are computed according to the mean node impurity decrease caused by splitting on each feature throughout the model (refer to Figure 7). Thus, we find that the WE measurements in March, followed by March's snowpack measurements and the collection year, contribute the most to the predictive ability of our model (according to average decreases in node impurity across all decision trees). More specifically, the WE measurements from March contribute significantly more to prediction than other features, with a contribution proportion of about 0.6. However, March's snowpack measurements contribute by a considerable proportion as well.

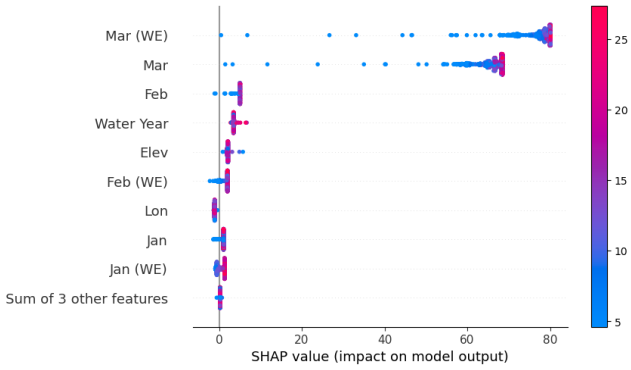

Fig. 8. SHAP Values Global Importances (Beeswarm Plot)

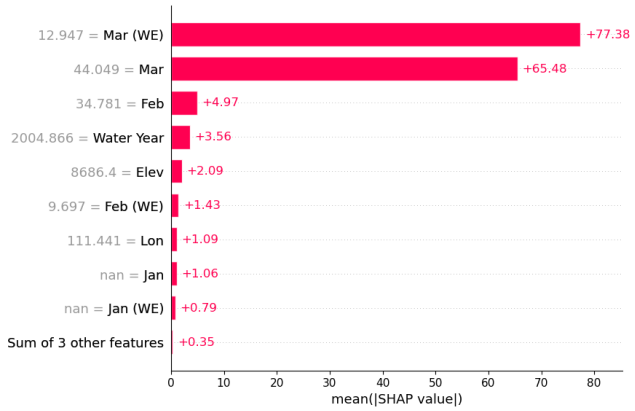

Fig. 9. Shap Values Global Importances

In analyzing the SHAP values of the test dataset, we find the ordering of features to be slightly different when it comes to global importance. The significance of calculating SHAP values for the features of the model is that we are comparing each feature and its contribution to the model relative to how much the other features affect the April snowpack measurements. Both the beeswarm plot and the global importances bar chart shown in Figure 8 and Figure 9 are represented by the average of the absolute values of SHAP (mean of |SHAP Values|) across all observations of the sample. From these calculated global importances, we find that the WE measurements in March and March's observed snowpack levels remain as the two most contributing factors in April's observed snowpack levels. However, the main difference from the RF Global Importances is that the SHAP values indicate February has a slightly greater effect on the response over Water Year. In addition to global importances, we delved into the effect of the features on a local level (i.e. observing data points individually).

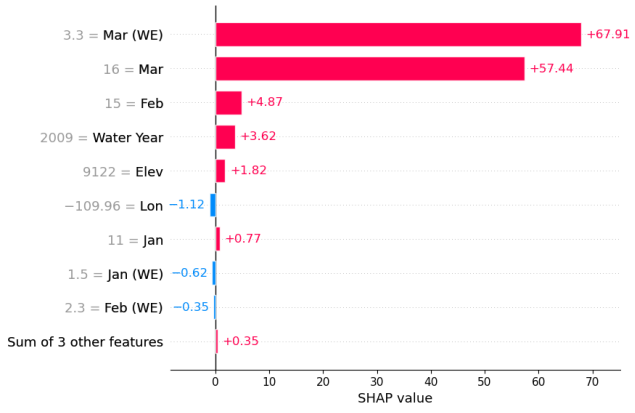

Fig. 10. Local Shap Values, Observation 13

The majority of observations from our test sample look similar when it comes to the SHAP values computed for each feature, but certain features with lower average SHAP values (the features with lower contributing rates or impacts on the response) occasionally show negative effects on April's snowpack measurements. In the case of the thirteenth ordered observation from our testing sample (see Figure 10), which has an observed April snowpack of 20.00 inches, we noted that the specific values of Longitude and the WE measurements for January and February in fact negatively swayed April's anticipated snowpack levels (relative to the baseline observation). In other words, these features, given their respective values for the specific observation, had a negative effect on the anticipated snowpack level for April (relative to the baseline observation). These effects vary between the observations in our sample, but the general trend indicates that the WE measurements for March, as well as the snowpack measurements for March, each contribute significantly (as well as positively) to April's snowpack.

VI. Conclusion

In conclusion, our analysis of snow accumulation patterns in Utah spanning several decades has provided valuable insights into the trends related to April snow levels. By leveraging data from the NRCS Public Database and employing machine learning models, we successfully discerned trends, gained insight, and predicted future April snow levels.

Our exploratory data analysis revealed that snowpack levels at the beginning of the year generally stay the same month to month, and later months exhibit more variability in their snow accumulation.

Through feature engineering and model comparison, we identified the Random Forest Regressor as the best-performing model for predicting April snow levels. This model, incorporating a snowpack-to-elevation ratio feature, demonstrated strong predictive ability with an R-squared value of 0.91 and RMSE of 8.65.

The feature importances analysis highlighted the importance of water equivalent and snowpack measurements in March, indicating that these variables are the most crucial in predicting April snow levels.

If we were to do this analysis again we would want to make more features, such as month to month difference in snow accumulation, and take advantage of categorical variables more thoroughly in our analysis. In addition, because we are dealing with observations collected by a list of SNOTEL sites across the state of Utah which vary in their geographical location and years of data collection, there may be some cause for concern with correlation in the model—more than anything, within-site correlation (i.e. correlation between observations collected from the same site). It would be helpful to take this aspect of the data into account for future studies and regression model testing. Lastly, it may be beneficial to look into predicting for April's snowpack levels earlier on in the season, such as with only using the data collected from January and February. According to our feature importance analysis, it may prove difficult to do so in the absence of data from March, but it could be important to observe how drastic the change is in our

predictive ability after taking those observations away from our model.

Overall, our study provides valuable insights into snow accumulation patterns in Utah and demonstrates the effectiveness of machine learning models in predicting snow levels. This information can be invaluable for local communities, water resource managers, and policymakers in planning for and adapting to changing snowpack conditions.