ILLINOIS INSTITUTE OF TECHNOLOGY

CSP 571 Data Preparation & Analysis

Data Analysis on 11 Years of Data for Rain to Snow Transition Zone

Shriya Prasanna

sprasanna@hawk.iit.edu

A20521733

Girish Rajani-Bathija

grajanibathija@hawk.iit.edu

A20503736

Ranjan Mishra

rmishra11@hawk.iit.edu

A20521033

Raghukarn Sharma

rsharma15@hawk.iit.edu

A20511671

Prof. Jawahar Panchal

Submission Date: 30 April 2023

Contents

Abstract

The hydrometeorological dataset consisting of time-series data from the water year 2004 to 2014 for weather, soil, precipitation, and snow data was used to study what factors affect snow accumulation within the rain-to-snow transition zone in southwestern Idaho. There were a total of 24 datasets that had to be preprocessed and merged into one dataset.

To study this dataset and answer the research question, the correlation was observed using corrplot, scatterplots and PCA biplot. Features that were positively and negatively correlated to snow depth were analyzed. Different variations of this dataset were created to study how dealing with outliers, correlation and performing dimensionality reduction would impact our interpretation and analysis of the features. To test these variations, models such as multiple linear regression, ridge regression and lasso regression were used.

1 Overview

1.1 Problem Statement

In this project, we are looking at hydrometeorological data, which is used to study the atmospheric water, precipitation, snow depth, soil moisture, soil temperature, wind speed/direction, and other factors. This would help in flood control, water supply, power generation, and reducing the effects on agricultural processes during the downslope water delivery, which occurs at rain-to-snow transition zones.

The dataset used in this project consists of a climatically sensitive zone located at Johnston Draw (JD) watershed, a sub-watershed of the Reynolds Creek Critical Zone Observatory, Idaho. This comprehensive hydrometeorological dataset consists of time-series data for 11 water years (1st October - 30th September) from the water year 2004 to 2014 that spans the rain-to-snow transition zone in southwestern Idaho.

Within the scope of this project, the objective is to look for the key features that can contribute to snow accumulation.

1.2 Questions the Project Seeks to Address

1. How does the timing of downslope water delivery impact the hydrological and biological processes in climatically sensitive regions during the transition from rain to snow?
2. What factors affect the snow conditions (snow depth) at Johnston Draw (JD) watershed?
3. Can we get the inference of snow accumulation, and soil moisture/temperature?

1.3 Literature Review and Related Work

The primary research paper for this project by Godsey, S. Marks, D., et al., (2018) [1] explains that the rain-to-snow transition zone in mountain regions is a climatologically sensitive region, and detailed hydrometeorological data from this zone are limited. The ongoing climate change is causing the rain-to-snow transition zone to move to higher elevations, which is impacting hydrological and biological processes. To understand these changes, a complete hydrometeorological dataset for water years 2004 through 2014 has been presented for the Johnston Draw watershed in southwestern Idaho, USA. The dataset includes hourly measurements of various hydrometeorological variables across a 372 m elevation gradient, such as air temperature, relative humidity, snow depth, incoming shortwave radiation, precipitation, wind speed and direction, soil moisture, and soil temperature. These data can be used to develop and validate hydrological models for better representation and understanding of the complex processes in the rain-to-snow transition zone.

The research paper by Newman, A. J., Clark, M. P., Sampson, K., et al., (2015) [2] presents a comprehensive data set of daily forcing and hydrologic response data for 671 small- to medium-sized basins in the contiguous United States, spanning a wide range of hydroclimatic conditions. Area-averaged forcing data for the period 1980-2010 was generated for three basin spatial configurations, and daily streamflow data was compiled from the United States Geological Survey National Water Information System. The Snow-17 snow model and the Sacramento Soil Moisture Accounting Model were calibrated using the shuffled complex evolution global optimization routine. Model performance was benchmarked using the Nash-Sutcliffe efficiency score, highlighting some regional variations in model performance. Data points with extreme error were found to be more common in arid basins with limited snow.

The research paper by Wayand, N. E., Massmann, A., Butler, C., Keenan, E., Stimberis, J., and Lundquist, J. D., (2015) [3] introduces a quality-controlled observational dataset from Snoqualmie Pass, Washington, which includes atmospheric, snow, and soil data. The dataset provides continuous meteorological forcing data at hourly intervals for a 24-year historical period and at half-hourly intervals for a more recent period. Additional observations include 40 years of snowboard new snow accumulation, multiple measurements of total snow depth, and manual snow pits. The recent years' data also include sub-daily surface temperature, snowpack drainage, soil moisture and temperature profiles, and eddy co-variance derived turbulent heat flux. This dataset can be used to test hypotheses about energy balance, soil and snow processes in the rain-snow transition zone.

The research paper by Winstral, A., Marks, D., and Gurney, R., (2013) [4] presents a new algorithm for modeling wind-affected snow accumulation and melt patterns in non-forested mountain regions. The algorithm uses terrain structure, vegetation, and wind data to adjust precipitation data and simulate wind-affected snow distributions. The model was developed and applied in three catchments in southwest Idaho, USA, and was tested against previously published cross-validation results. Results showed that the wind-affected model accurately located large drift zones, snow-scoured slopes, and produced melt patterns consistent with observed streamflow. The algorithm's computational efficiency and modest data requirements make it ideal for large-scale operational applications.

2 Data Processing

2.1 Data Sources

2.1.1 Overview

The hydrometeorological dataset retrieved from US Department of Agriculture, has been collected from twelve (12) stations at the Johnston Draw (JD) watershed, a sub-watershed of the Reynolds Creek Critical Zone Observatory, Idaho to study the rain-snow transition zone. This area is generally known as a climatically sensitive zone, meaning that small changes in weather conditions can alter seasonal snow cover, the timing of melt, the delivery of liquid water to soil, and, ultimately, the ecosystems they sustain.

Each station started at different times, but all ended in 2014. The description of each station can be seen below.

Station	Elevation (m)	Aspect	Start Date	Duration (WY)	# of Observations
125b	1496	NE	10 Jan 2003	11	96432
125	1508	SE	10 Jan 2003	11	96432
Jdt1	1552	N	11 May 2005	9	78048
Jdt2b	1611	S	4 Mar 2011	4	31357
Jdt2	1613	N	5 Nov 2005	9	78048
Jdt3	1655	N	21 Sep 2005	9	79128
Jdt3b	1659	S	13 Dec 2010	4	33299
Jdt4b	1704	S	4 Mar 2011	4	31356
Jdt4	1706	N	2 Nov 2005	9	78120
Jdt5	1757	N	2 Nov 2005	9	78120
124b	1778	SE	11 Nov 2006	8	69384
124	1804	NE	1 Oct 2003	11	96432

Table 1 - Description of each station

2.1.2 Feature Description

- Meteorological Data Description:

Field Name	Type	Description
Date_time	Date/Time	Date followed by time
WY	Numeric	Water Year
Year	Numeric	Calendar Year
Month	Numeric	Month of Year
Day	Numeric	Day of Month
Hour	Numeric	Hour of Day
Minute	Numeric	Minute of Hour
ppt_a	Numeric	Wind Corrected Precipitation (mm)
perc_snow	Numeric	Fraction of precipitation that is snow (unitless ratio from 0 to 1)

Table 2 - Precipitation - 3 time-series files (Stations 125, 124, and 124b)

Field Name	**Type**	**Description**
Date_time	Date/Time	Date followed by time
WY	Numeric	Water Year
Year	Numeric	Calendar Year
Month	Numeric	Month of Year
Day	Numeric	Day of Month
Hour	Numeric	Hour of Day
Minute	Numeric	Minute of Hour
T_a	Numeric	Air Temperature ~3 m above ground surface (^oC)
RH	Numeric	Relative Humidity ~3 m above ground surface (0 - 1)
e_a	Numeric	Water Vapor Pressure ~3 m above ground surface (Pa)
T_d	Numeric	Dew Point Temperature ~3 m above ground surface (^oC)
S_i	Numeric	Incoming Solar Radiation (W m^{-2})
W_s	Numeric	Wind Speed ~3 m above ground surface (ms^{-1})
W_d	Numeric	Wind Direction ~3 m above ground surface (^{o} from N)

Table 3 - Weather - 11 time-series files (All Stations)

- Snow, and Soils Data Description:

Field Name	**Type**	**Description**
Date_time	Date/Time	Date followed by time
WY	Numeric	Water Year
Year	Numeric	Calendar Year
Month	Numeric	Month of Year
Day	Numeric	Day of Month
Hour	Numeric	Hour of Day
Minute	Numeric	Minute of Hour
T_g5	Numeric	Soil Temperature at 5cm depth (^oC)
T_g20	Numeric	Soil Temperature at 20cm depth (^oC)
T_g35	Numeric	Soil Temperature at 35cm depth (^oC)
T_g50	Numeric	Soil Temperature at 50cm depth (^oC)
T_g75	Numeric	Soil Temperature at 75cm depth (^oC)
T_g90	Numeric	Soil Temperature at 90cm depth (^oC)
T_g100	Numeric	Soil Temperature at 100cm depth (^oC)
S_m5	Numeric	Soil Moisture at 5cm depth (m3 m-3 - dimensionless)
S_m20	Numeric	Soil Moisture at 20cm depth (m3 m-3 - dimensionless)

S_m35	Numeric	Soil Moisture at 35cm depth (m3 m-3 - dimensionless)
S_m50	Numeric	Soil Moisture at 50cm depth (m3 m-3 - dimensionless)
S_m75	Numeric	Soil Moisture at 75cm depth (m3 m-3 - dimensionless)
S_m90	Numeric	Soil Moisture at 90cm depth (m3 m-3 - dimensionless)
S_m100	Numeric	Soil Moisture at 100cm depth (m3 m-3 - dimensionless)

Table.4: Soil - 9 time-series files (Stations 124ba, 124bs, jdt1, jdt2, jdt2b, jdt3, jdt3b, jdt4, jdt4b)

Field Name	Type	Description
Date_time	Date/Time	Date followed by time
WY	Numeric	Water Year
Year	Numeric	Calendar Year
Month	Numeric	Month of Year
Day	Numeric	Day of Month
Hour	Numeric	Hour of Day
Minute	Numeric	Minute of Hour
z_s_124	Numeric	Snow depth data for site 124
z_s_124b	Numeric	Snow depth data for site 124b
z_s_125	Numeric	Snow depth data for site 125
z_s_jdt1	Numeric	Snow depth data for site jdt1
z_s_jdt2	Numeric	Snow depth data for site jdt2
z_s_jdt3	Numeric	Snow depth data for site jdt3
z_s_jdt4	Numeric	Snow depth data for site jdt4
z_s_jdt5	Numeric	Snow depth data for site jdt5
z_s_jdt2b	Numeric	Snow depth data for site jdt2b
z_s_jdt3b	Numeric	Snow depth data for site jdt3b
z_s_jdt4b	Numeric	Snow depth data for site jdt4b

Table.5: Snow_Depth – 1 time-series file (Report snow depth for all stations)

Dataset	Files	Stations
Precipitation	3 time-series files	Stations (125, 124, and 124b)
Weather	11 time-series files	All 11 Stations
Soil	9 time-series files	Stations (124ba, 124bs, jdt1, jdt2, jdt2b, jdt3, jdt3b, jdt4, jdt4b)
Snow Depth	1 time-series file	All 11 stations

Table 6 - Summary of all datasets and stations

2.2 Data Preprocessing

After loading each dataset, the Date_time and Minute features were removed as these features were found to be redundant. The Date_time feature was a combination of year, month, and day which were already separate features within each dataset. The Minute feature contained only 0's and did not play a significant part in the inference of the data.

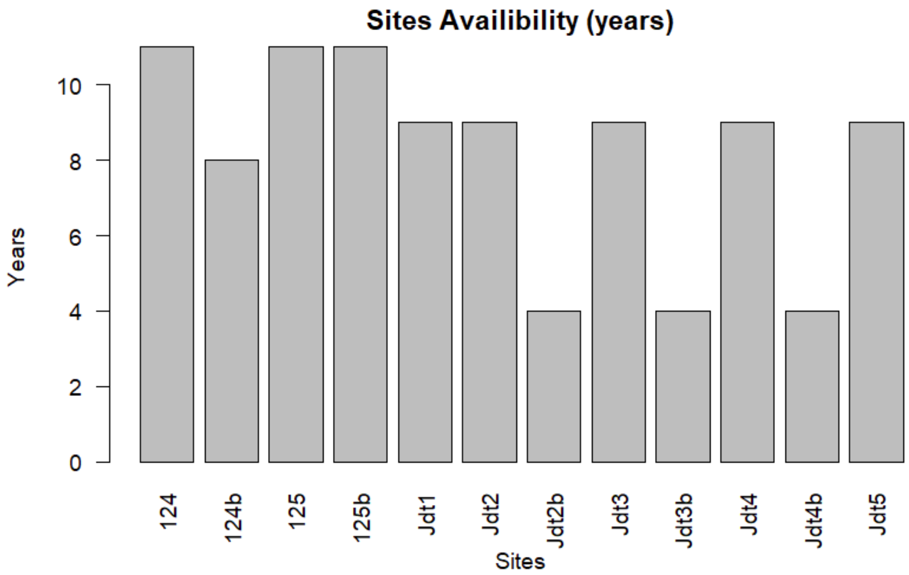


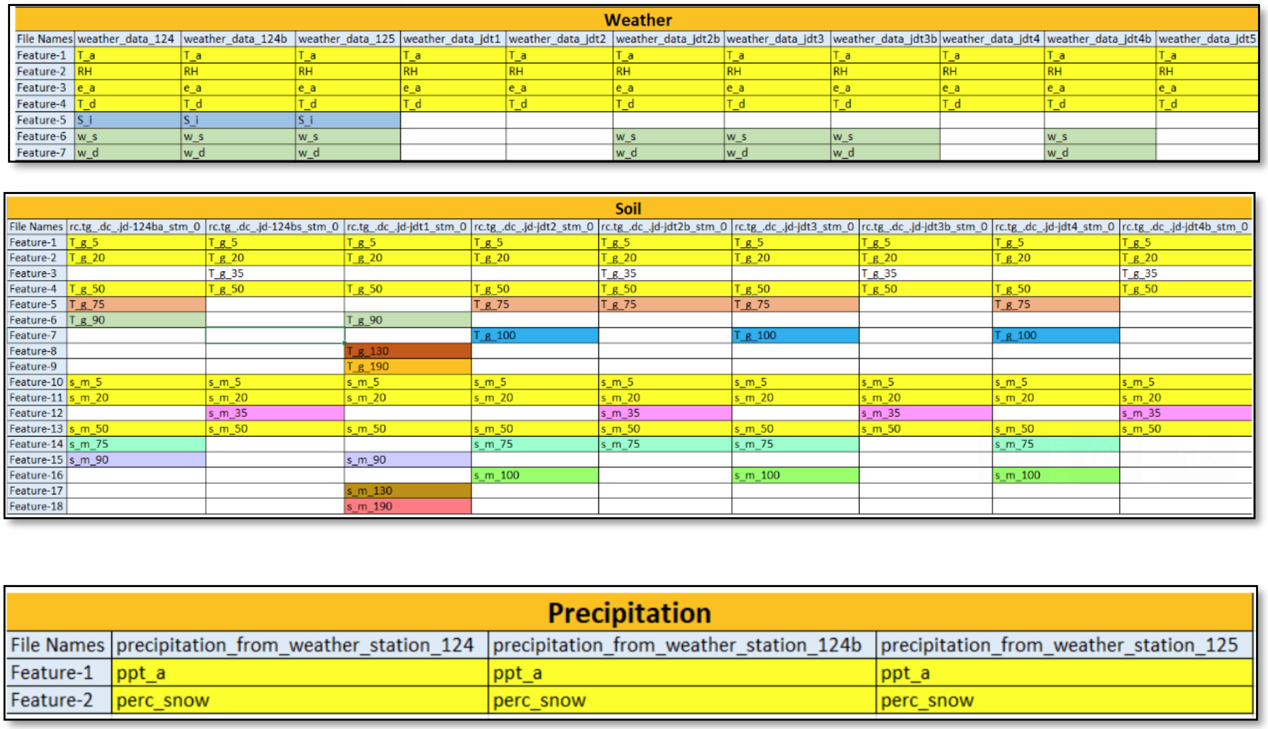
Figure 1 – Showing number of years each site captured meteorological data

Data Issues – Challenges encountered by this time series data.

- The data collected at uneven time intervals may contain missing data and features with varying time intervals.
- To deal with the above challenge, we may need to interpolate the data or resample it to a consistent time interval. However, this can introduce errors and biases in the analysis.
- The above graph shows, not all the sites were active/working for 11 years. Due to this, the data won't be consistent with time and can lead to further complexities like incorrect conclusions, biased results, inaccurate predictions, and wasted resources.

Having carefully studied the structure, quality, completeness, and consistency of the datasets, it was observed that there were some missing values.

- Missing data in some files represented by -9999 and NA.
- Some stations were missing a few features due to missing instrumentation installed at stations. For example, stations jdt1, jdt2, jdt4, and jdt5 did not record weather data for solar radiation (S_i), wind speed (w_s), and wind direction (w_d).
- The soil dataset contained soil temperature and soil moisture features at varying depths and as shown below, the features recorded throughout each station varied.



Weather

File Names	weather_data_124	weather_data_124b	weather_data_125	weather_data_jdt1	weather_data_jdt2	weather_data_jdt2b	weather_data_jdt3	weather_data_jdt3b	weather_data_jdt4	weather_data_jdt4b	weather_data_jdt5
Feature-1	T_a	T_a	T_a	T_a	T_a	T_a	T_a	T_a	T_a	T_a	T_a
Feature-2	RH	RH	RH	RH	RH	RH	RH	RH	RH	RH	RH
Feature-3	e_a	e_a	e_a	e_a	e_a	e_a	e_a	e_a	e_a	e_a	e_a
Feature-4	T_d	T_d	T_d	T_d	T_d	T_d	T_d	T_d	T_d	T_d	T_d
Feature-5	S_i	S_i	S_i								
Feature-6	w_s	w_s	w_s			w_s	w_s	w_s		w_s	
Feature-7	w_d	w_d	w_d			w_d	w_d	w_d		w_d	

Soil

File Names	rc.tg_.dc_.jd-124ba_stm_0	rc.tg_.dc_.jd-124bs_stm_0	rc.tg_.dc_.jd-jdt1_stm_0	rc.tg_.dc_.jd-jdt2_stm_0	rc.tg_.dc_.jd-jdt2b_stm_0	rc.tg_.dc_.jd-jdt3_stm_0	rc.tg_.dc_.jd-jdt3b_stm_0	rc.tg_.dc_.jd-jdt4_stm_0	rc.tg_.dc_.jd-jdt4b_stm_0
Feature-1	T_g_5	T_g_5	T_g_5	T_g_5	T_g_5	T_g_5	T_g_5	T_g_5	T_g_5
Feature-2	T_g_20	T_g_20	T_g_20	T_g_20	T_g_20	T_g_20	T_g_20	T_g_20	T_g_20
Feature-3		T_g_35			T_g_35		T_g_35		T_g_35
Feature-4	T_g_50	T_g_50	T_g_50	T_g_50	T_g_50	T_g_50	T_g_50	T_g_50	T_g_50
Feature-5	T_g_75			T_g_75	T_g_75	T_g_75		T_g_75	
Feature-6	T_g_90		T_g_90						
Feature-7				T_g_100		T_g_100		T_g_100	
Feature-8			T_g_130						
Feature-9			T_g_190						
Feature-10	s_m_5	s_m_5	s_m_5	s_m_5	s_m_5	s_m_5	s_m_5	s_m_5	s_m_5
Feature-11	s_m_20	s_m_20	s_m_20	s_m_20	s_m_20	s_m_20	s_m_20	s_m_20	s_m_20
Feature-12		s_m_35			s_m_35		s_m_35		s_m_35
Feature-13	s_m_50	s_m_50	s_m_50	s_m_50	s_m_50	s_m_50	s_m_50	s_m_50	s_m_50
Feature-14	s_m_75			s_m_75	s_m_75	s_m_75		s_m_75	
Feature-15	s_m_90		s_m_90						
Feature-16				s_m_100		s_m_100		s_m_100	
Feature-17			s_m_130						
Feature-18			s_m_190						

Precipitation

File Names	precipitation_from_weather_station_124	precipitation_from_weather_station_124b	precipitation_from_weather_station_125
Feature-1	ppt_a	ppt_a	ppt_a
Feature-2	perc_snow	perc_snow	perc_snow

Figure 2 (a) – (c) - Showing a summary of all features for each station in each dataset

- As shown above, the variation between different sites having different features results in missing values and makes it very complex to combine all the data uniformly.

Steps taken to identify, handle missing values, and merge datasets into one.

1. To deal with **missing values**, we first checked the files that contained missing values in the form of -9999 and replaced those with NA's. The -9999 missing values represented missing values due to the station starting at a later date (as shown in Figure 1 above, some stations started late and so recorded observations for fewer years).
2. Dropping observations with missing values was not a viable option, as this would significantly reduce the number of samples in our datasets. These missing values were instead replaced with either the median or mean, depending on the density distribution of the feature.
3. A histogram and density plot were created for the features with missing values and mean imputation was performed if the distribution was approximately normal and median imputation was performed if the distribution was skewed [5].

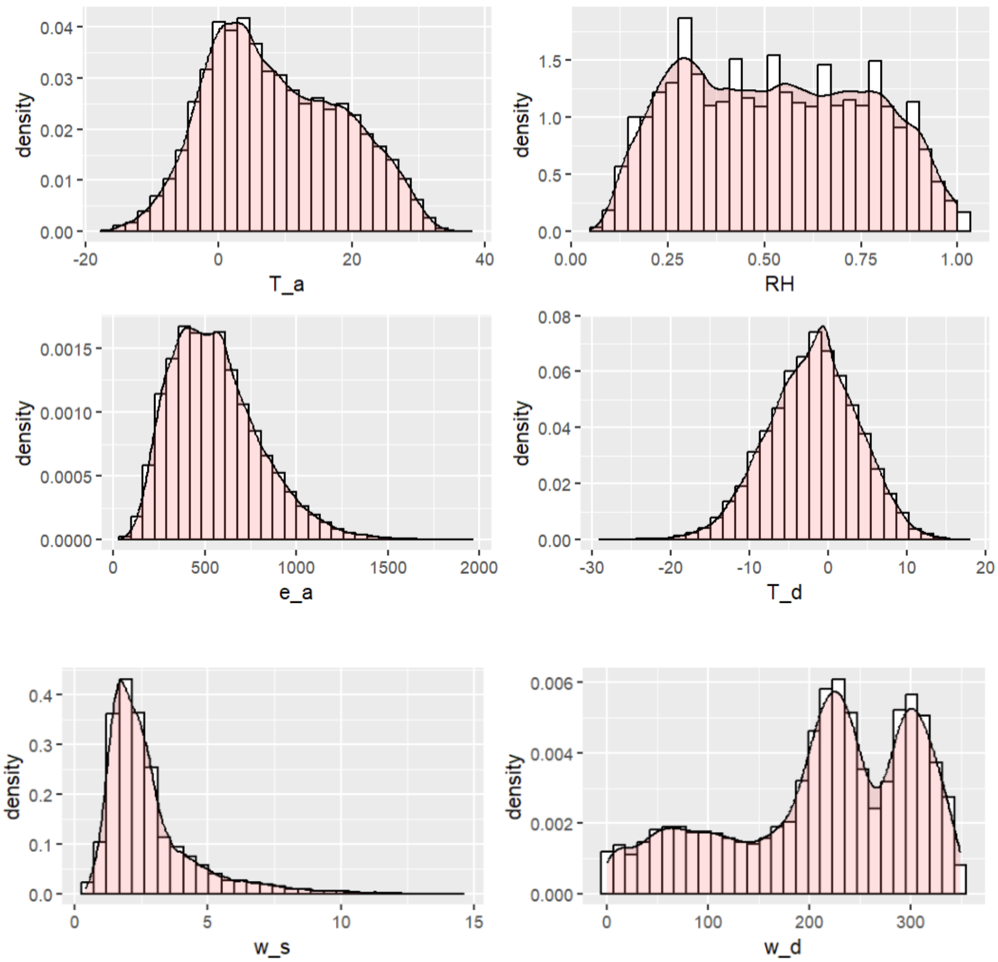


Figure 3 – Distributions for features T_a, RH, e_a, T_d, w_s, and w_d from Weather Data

4. The above sample distributions from weather data shows that:
   - The features Air Temperature (T_a), Relative Humidity (RH), Water Vapor Pressure (e_a), Wind Speed (w_s), and Wind Direction (w_d) had some level of skewness, so their missing values were replaced with the median.
   - However, Dew Point Temperature (T_d) had an approximately normal distribution so the missing values for this feature were replaced with mean.
   - Once all missing values for all 11 stations in weather data were dealt with separately for each file, then we took a mean of common features of all 11 files and then were merged into 1 file grouping by WY, Year, Month, Day, and Hour.
5. Similarly, for soil data, it was observed that:
   - All the soil temperature (T_g) and soil moisture (s_m) features contained missing values, and they all had a skewed distribution. Therefore, all missing values for all 9 stations in soil data were replaced with a median.
   - Merged the 9 files; since each file contains varying features, as shown in Figure 2 above, we identified common features among the sites and took an average of them to make more relevant one uniform feature and the remining features which

were not common were appended to the main data frame grouping by WY, Year, Month, Day, and Hour.

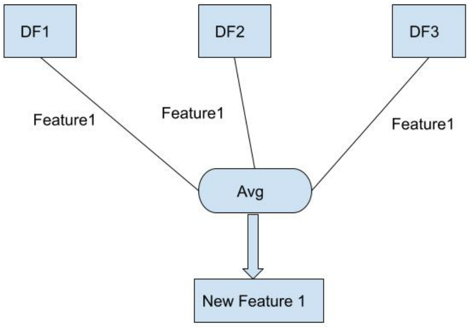


Figure 4 – Sample approach of common features and their processing

6. For Precipitation data, all three time-series precipitation files (stations 124, 124b and 125) had observations from WY 2004 – 2014 and therefore did not contain any missing values. All three files for precipitation were merged into 1 file grouping by WY, Year, Month, Day, and Hour.
7. Lastly, for snow data, the single file contained the snow depth (in centimeters) for each station. Figure 5 below illustrates the distribution of snow_depth for a few stations. It was observed that throughout most parts of the year, the reported snow depth was 0, and therefore, all missing values were replaced with 0. Following this, the feature snow depth for all stations was averaged and merged into 1 snow depth (z_s) feature. This feature will be used as the response(Y) during data modeling since our main research question is to understand how the different meteorological features impact snow depth.

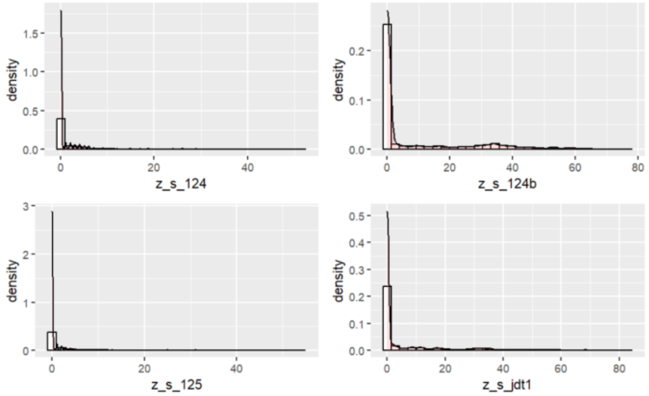


Figure 5 – Showing distributions for snow depth from stations 124, 124b, 125, and jdt1

8. To create our final dataset, all four datasets were merged into one.
   - We now have observations for each time-period from the water year 2004 to 2014 for weather, soil, snow, and precipitation in one file. However, some features such as the soil temperatures (T_g) and soil moisture began recording observations from the water year 2011.

- To deal with this, those missing values were replaced with 0's to show that these features were not recorded prior to the water year 2011.

Why Scaling and normalization were a drawback for performing inference?

Scaling and normalization were proposed as part of the data preprocessing stage; however, after carefully understanding the data, the best option was not to scale or normalize the features. Scaling/normalization is especially useful for machine learning models to better predict the features and have less bias.

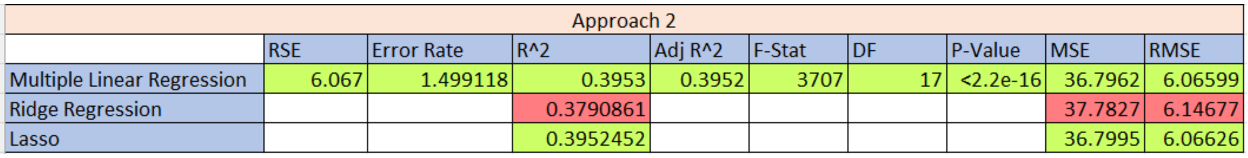
However, for this project, our main intention is not to build an accurate machine-learning model for prediction, instead, it is to make inferences and understand the relationships between the features. Since we are working with weather data, we have features such as temperature, Vapor pressure, relative humidity, precipitation, moisture, wind speed/ direction, snow depth, etc., and if we perform scaling, this will negatively impact our ability to understand and interpret the data.

An alternative dataset preprocessing approach (1): Creating multiple datasets by stations for different years may not be a good idea for several reasons:

1. **Loss of information**: By splitting the data into multiple datasets by station, year or based on another feature, you are losing the information contained in the full dataset. This may make it more difficult to identify patterns or relationships that exist across all stations and years. This could cause a loss of generality in our analysis.
2. **Reduced statistical power**: smaller datasets may have fewer observations, resulting in reduced statistical power. This can make it more difficult to detect meaningful relationships or differences between variables.
3. **Difficulty in comparing/combining results**: When you analyze data from multiple datasets, it can be difficult to compare results across the different datasets. For example, if you find a significant difference between two stations, you may not know if the same difference exists between the other stations. Also keeping track of these different analyses would require lot of manual interventions.
4. **Cannot answer one of the research questions:** How does the timing of downslope water delivery impact the hydrological and biological processes in climatically sensitive regions during the transition from rain to snow? Due to the complexity and reasons mentioned above of this time series data, this question was not answered.

An alternative dataset preprocessing approach (2): Creating a single merged dataset using only the common features across different stations.

1. Instead of taking all the features across all stations and imputing values for the missing features, only the common features were taken and merged into a single dataset.
2. Multiple linear regression, ridge regression, and lasso regression were performed.



Approach 2									
	RSE	Error Rate	R^2	Adj R^2	F-Stat	DF	P-Value	MSE	RMSE
Multiple Linear Regression	6.067	1.499118	0.3953	0.3952	3707	17	<2.2e-16	36.7962	6.06599
Ridge Regression			0.3790861					37.7827	6.14677
Lasso			0.3952452					36.7995	6.06626

Figure 6 – Showing model results on alternative dataset approach using only common features

- This approach contained only 17 out of 32 features due to only taking the common features.
- The above results of this dataset are worse than the original approach (which will be discussed later on) and therefore this approach was dropped.
- The models were only able to explain about 38-40% of the target variance.
- Taking a subset of features, although common, negatively affected our interpretation of the data as the dropped features were relevant for inference.

3 Exploratory Data Analysis

3.1 Bivariate Visualizations

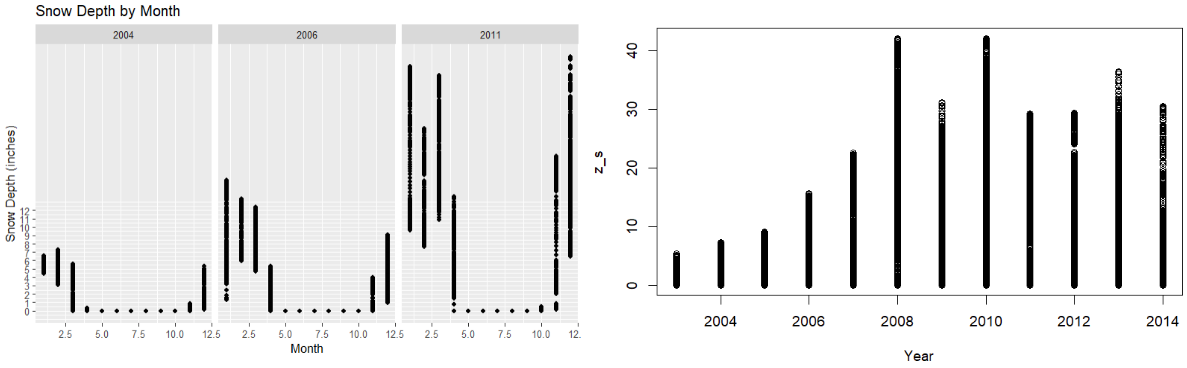


Figure 7 (a) – (b) – Showing snow accumulation per month for three years (a) and per year (b)

Since the snow depth (z_s) is the response feature, and we are trying to understand this feature, the snow depth pattern was observed each month for 3 years (2004, 2006, and 2011).

- During the months Jan, Feb, Mar, Nov, and Dec, the snow depth was higher than other months in that water year,
- This pattern is observed throughout all years for the Johnston Draw watershed in southwestern Idaho.

In addition to observing the peak months for snow depth (z_s), we also observed the peak years for this feature as shown above on the right. It was seen that as the years went on, the overall snow depth increased.

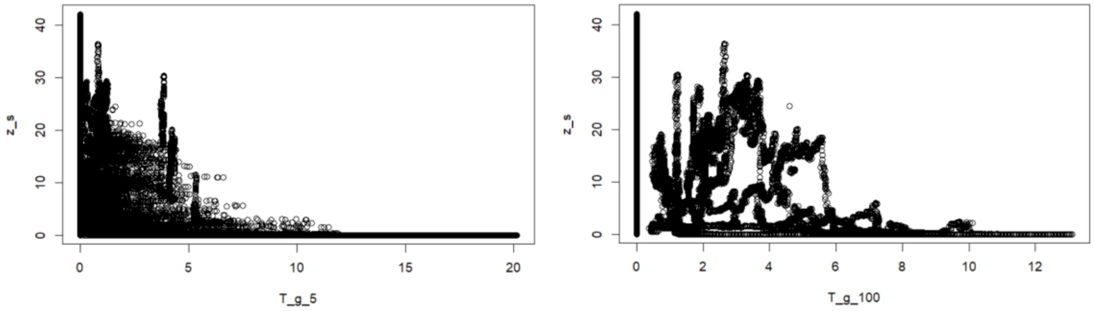


Figure 8 (a) – (b) - Showing snow depth and soil temperature negative correlation

An inverse correlation was observed between the snow depth (z_s) and soil temperatures at different depths (T_g) after preliminary analysis.

- This meant that as the soil temperature increased, the snow depth decreased.
- This correlation will be further explored during principal component analysis (PCA).

3.2 Correlation

After having preprocessed the data, the next step is to explore and understand the data. Firstly, a correlation matrix was created and visualized. The corrgram function was used to observe the correlation among features.

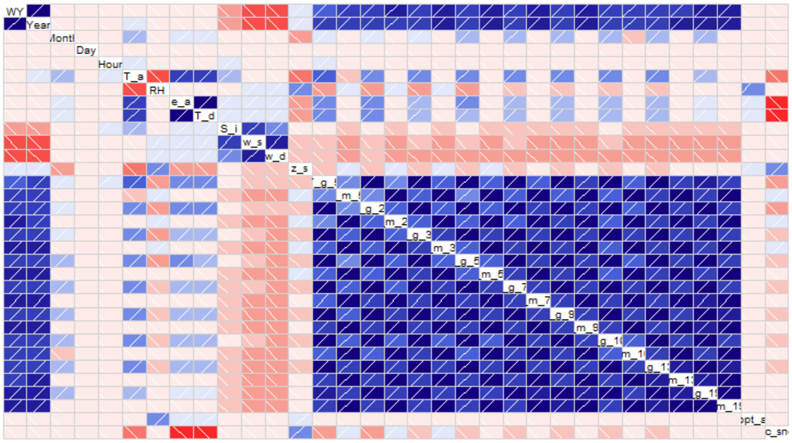


Figure 9 – Showing correlation visualization among features

- All the soil temperature features (T_g) are highly correlated with each other.
- All the soil moisture features (s_m) are highly correlated with each other.
- Highly correlated features may indicate redundancy or multicollinearity in the dataset, which can lead to overfitting and inaccuracies in predictive modeling.
- It may be necessary to remove or combine highly correlated features to improve the quality of the analysis.
- The multicollinearity within this dataset will be explored further during Principal Component Analysis (PCA).

Low-correlated features may measure different aspects of the underlying phenomenon or be unrelated to each other and may provide unique information for predictive modeling. Therefore, it is important to consider the research question and analytical goals when interpreting the correlation relationship among features.

First approach to handling correlated features:

To handle the fully correlated features:

- All the soil temperature (T_g) features were merged into one soil temperature feature.
- All the soil moisture (s_m) features were merged into one soil moisture feature.
- The models, which will be discussed later on, will be trained using the data with correlation as is, as well as after handling the highly correlated features to observe whether handling such correlation will impact the results.

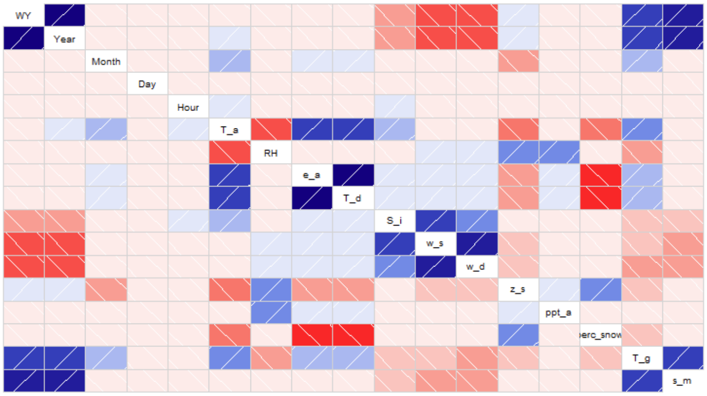


Figure 10 – Correlation visualization after handling T_g and s_m correlation

3.3 Detecting/Handling Outliers

The next step in the EDA process is to detect which features contain outliers and then handle them using the appropriate technique.

Challenges faced when detecting outliers:

- A Box Plot for each feature was plotted.
- Used Visual Inspection initially to visually examine and identify outliers.
- Due to the complexity of time series weather data, this approach was subjective, and it was difficult to visually identify anomalies.
- **Solution:** Apply Tukey’s method

How were outliers detected?

- All features were plotted using box plots to observe the presence of outliers.
- Applied the Tukey method to visualize the outliers.
- For performing Tukey method:

  a. We first calculated the lower and upper bounds for outlier detection using the Tukey method.
  b. We used the quantile function to calculate the 25th and 75th percentiles (Q1 and Q3) for each feature.
  c. The interquartile range (IQR) is then calculated as the difference between Q3 and Q1.
  d. The lower and upper bounds are calculated as Q1 - 1.5 * IQR and Q3 + 1.5 * IQR, respectively.
  e. Lastly, we identified outliers using the lower and upper bounds calculated above. It uses the which function to find the indices of the observations in the dataset where the feature is less than the lower bound or greater than the upper bound.

How were outliers dealt with?

The Winsorize function from the DescTools package was used to deal with outliers:

- Winsorizing is a data preparation method that swaps out extreme values for less extreme ones, minimizing the influence of outliers.
- This function was used to substitute the 5th and 95th percentiles for the lowest and highest 5% of values.
- Finally, a boxplot was plotted after this process to confirm that outliers were minimized.

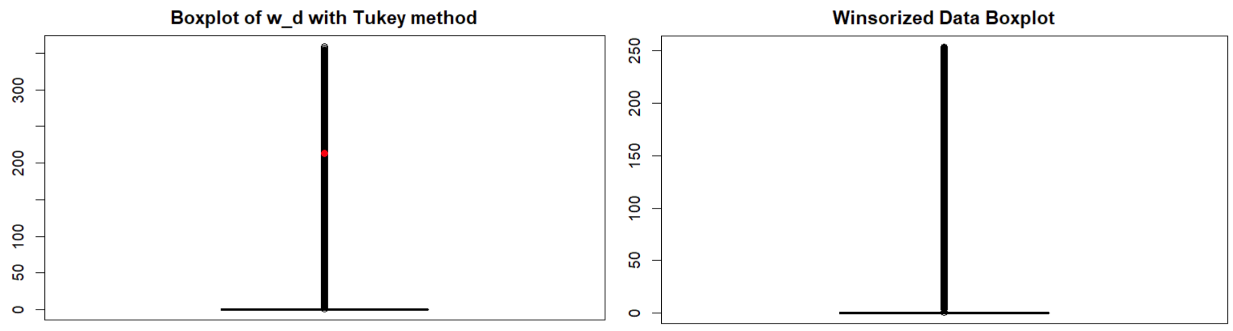


Figure 11 - Visualizing the process of detecting outlier using Tukey method and handling them using Winsorizing technique.

3.4 Principal Component Analysis (PCA)

Performing principal component analysis (PCA) on weather, soil, snow, and precipitation data involves reducing the dimensionality of the dataset by creating new variables that capture the most significant variation in the original data.

Process of performing PCA for dimensionality reduction:

1. Standardize the data to have a mean of zero and a standard deviation of one.
2. A PCA biplot was created to study the correlation of the features.

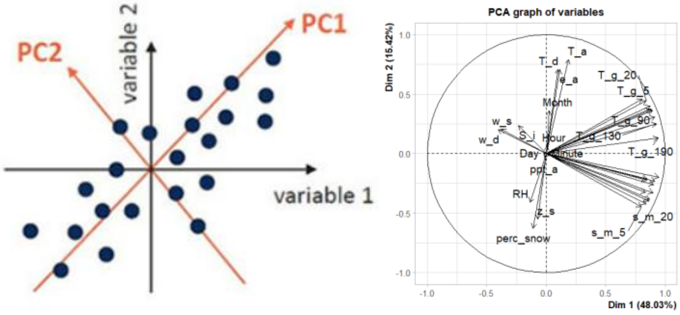


Figure 12 (a) – (b) - Showing PC1 and PC2 axis (a) and Biplot of dataset after PCA (b)

Link: Principal Component Analysis Using R (soa.org)

In the above left figure, the PC1 axis is the first principal direction along which the samples show the largest variation. The PC2 axis is the second most important direction, and it is orthogonal to the PC1 axis.

From the above PCA biplot:

- We observe that positively correlated variables are grouped together.
- Negatively correlated variables are located on opposite sides of the plot origin.
- The distance between variables and the origin measures the quality of the variables on the factor map.
- Variables that are away from the origin are well represented on the factor map.
- Variables that are close to circumference (like T_g_20, s_m_5) manifest the maximum representation of the principal components.
- Feature like RH and ppt_a show weak representation of the principal components.
- It was also observed that features such as relative humidity (RH), Wind Corrected Precipitation (ppt_a), and fraction of precipitation that is snow (perc_snow) are all positively correlated to snow depth (z_s).
- However, features such as Air Temperature (T_a), Dew Point Temperature (T_d), and Water Vapor Pressure (e_a) are negatively correlated to snow depth (z_s).
- Lastly, features such as Soil Temperatures (T_g) and Soil Moistures (s_m) are not likely to be correlated to snow depth (z_s) since they meet at 90 degree angle.

3. The covariance matrix was computed from the standardized data, and the eigenvectors and eigenvalues were extracted to identify the principal components of the data.
4. The principal components were then ranked in order of the amount of variance they explain, with the first component explaining the most variance and subsequent components explaining progressively less variance.
5. Following this, we observed the contributions of variables from PC1 with the top 22 contributors (left figure) and the top 10 contributors of variables from PC2 (right figure).

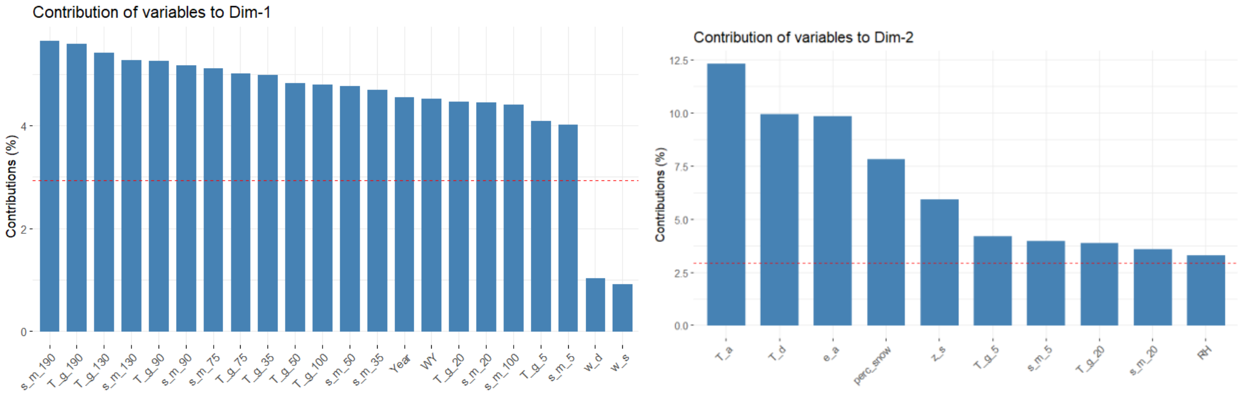


Figure 13 (a) – (b) – Contributions of variables from PC1 (a) and PC2 (b)

It was observed that the features after s_m_5 are not very relevant based on the above graph so we can restrict our feature subset to the top 20 most relevant features.

4 Model Training/Validation

The following models were developed to understand how various features play a key role in snow conditions at JD catchment.

- Multiple Linear Regression
- Ridge Regression
- Lasso Regression

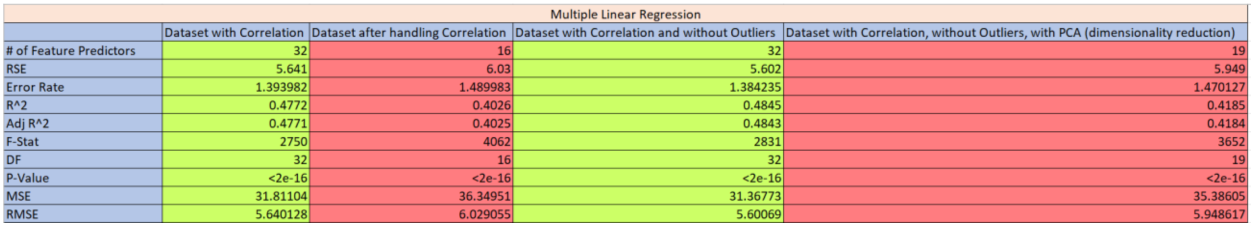
Each model was trained four (4) times:

- Original dataset with correlated features and with outliers (Model 1)
- Dataset after handling the T_g and s_m correlated features with outliers (Model 2)
- Original dataset with correlated features after handling outliers (Model 3)
- Original dataset after handling outliers and performing dimensionality reduction using PCA (Model 4)

This will give us an idea and a better understanding as to how handling outliers and highly correlated features, as well as performing dimensionality reduction, will impact the association between the predictors and the snow depth (z_s) response.

4.1 Multiple Linear Regression

Multiple Linear Regression was performed using the snow depth (z_s) as the response variable while using all other features as predictors. The purpose of this is to observe how significant the different features are and their association between those variables and snow depth.



Multiple Linear Regression				
	Dataset with Correlation	Dataset after handling Correlation	Dataset with Correlation and without Outliers	Dataset with Correlation, without Outliers, with PCA (dimensionality reduction)
# of Feature Predictors	32	16	32	19
RSE	5.641	6.03	5.602	5.949
Error Rate	1.393982	1.489983	1.384235	1.470127
R^2	0.4772	0.4026	0.4845	0.4185
Adj R^2	0.4771	0.4025	0.4843	0.4184
F-Stat	2750	4062	2831	3652
DF	32	16	32	19
P-Value	<2e-16	<2e-16	<2e-16	<2e-16
MSE	31.81104	36.34951	31.36773	35.38605
RMSE	5.640128	6.029055	5.60069	5.948617

Figure 14 - Showing results obtained from Multiple Linear Regression

As shown above, all four approaches result in very similar performance. It was observed that when dimensionality reduction was performed either by averaging the correlated features (model 2) or performing PCA (model 4), the performance dropped. This means that most of the features are relevant in understanding how snow depth is affected and, therefore, should not be removed.

Summary of observations:

- The p-value of the F-Stat is <2e-16 which is very significant.
- Almost all of the features were highly significant.
- The day of the month did not have a significant association with snow depth (z_s).
- The R^2 for the above models range from 0.40 to 0.48, showing that 40-48% of the target variance can be explained. (Higher R^2 is better)
- Small difference between R^2 and Adj R^2, which indicates that the model is less likely to overfit.
- Model 1 and Model 3, using all features, perform better than Model 2 and Model 4, which have reduced dimensions.

4.2 Ridge Regression

Ridge Regression was performed using the snow depth (z_s) as the response variable while using all other features as predictors with the goal of shrinking the coefficients close to 0.

- This model is used when there is multicollinearity present.
- Shrinkage penalty finds optimal lambda to reduce MSE.

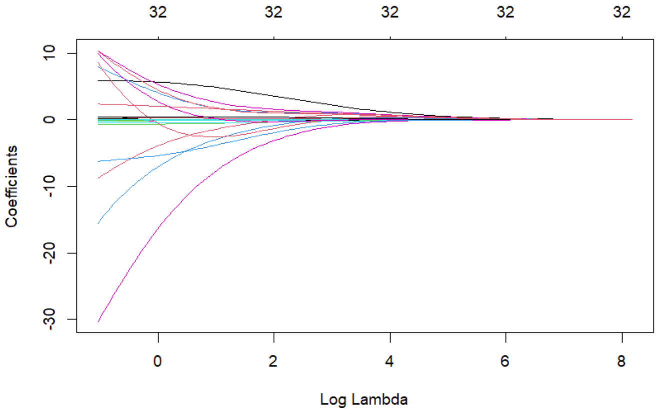
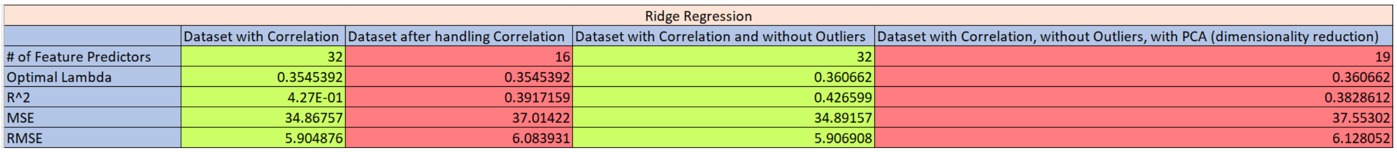
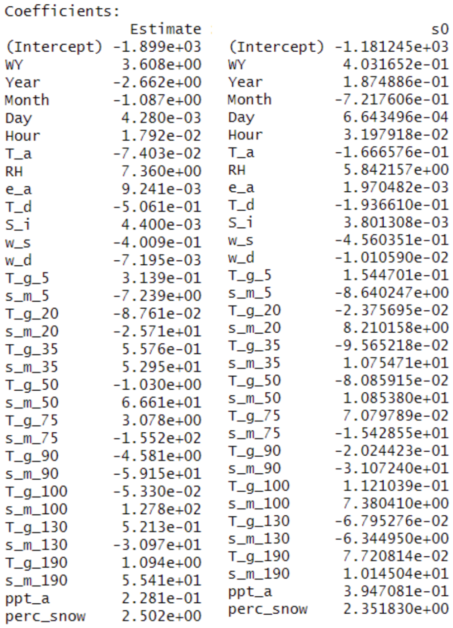


Figure 15 - Trace plot to visualize how the coefficients are shrunk by increasing lambda



Ridge Regression				
	Dataset with Correlation	Dataset after handling Correlation	Dataset with Correlation and without Outliers	Dataset with Correlation, without Outliers, with PCA (dimensionality reduction)
# of Feature Predictors	32	16	32	19
Optimal Lambda	0.3545392	0.3545392	0.360662	0.360662
R^2	4.27E-01	0.3917159	0.426599	0.3828612
MSE	34.86757	37.01422	34.89157	37.55302
RMSE	5.904876	6.083931	5.906908	6.128052

Figure 16 - Showing results obtained from Ridge Regression



```
Coefficients:
              Estimate
(Intercept) -1.899e+03   (Intercept) -1.181245e+03
WY           3.608e+00   WY           4.031652e-01
Year        -2.662e+00   Year         1.874886e-01
Month       -1.087e+00   Month       -7.217606e-01
Day          4.280e-03   Day          6.643496e-04
Hour         1.792e-02   Hour         3.197918e-02
T_a         -7.403e-02   T_a         -1.666576e-01
RH           7.360e+00   RH           5.842157e+00
e_a          9.241e-03   e_a          1.970482e-03
T_d         -5.061e-01   T_d         -1.936610e-01
S_i          4.400e-03   S_i          3.801308e-03
w_s         -4.009e-01   w_s         -4.560351e-01
w_d         -7.195e-03   w_d         -1.010590e-02
T_g_5        3.139e-01   T_g_5        1.544701e-01
s_m_5       -7.239e+00   s_m_5       -8.640247e+00
T_g_20      -8.761e-02   T_g_20      -2.375695e-02
s_m_20      -2.571e+01   s_m_20       8.210158e+00
T_g_35       5.576e-01   T_g_35      -9.565218e-02
s_m_35       5.295e+01   s_m_35       1.075471e+01
T_g_50      -1.030e+00   T_g_50      -8.085915e-02
s_m_50       6.661e+01   s_m_50       1.085380e+01
T_g_75       3.078e+00   T_g_75       7.079789e-02
s_m_75      -1.552e+02   s_m_75      -1.542855e+01
T_g_90      -4.581e+00   T_g_90      -2.024423e-01
s_m_90      -5.915e+01   s_m_90      -3.107240e+01
T_g_100     -5.330e-02   T_g_100      1.121039e-01
s_m_100      1.278e+02   s_m_100      7.380410e+00
T_g_130      5.213e-01   T_g_130     -6.795276e-02
s_m_130     -3.097e+01   s_m_130     -6.344950e+00
T_g_190      1.094e+00   T_g_190      7.720814e-02
s_m_190      5.541e+01   s_m_190      1.014504e+01
ppt_a        2.281e-01   ppt_a        3.947081e-01
perc_snow    2.502e+00   perc_snow    2.351830e+00
```

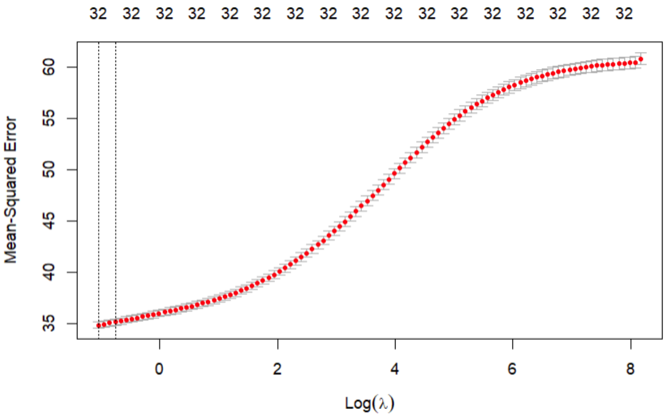


Figure 17 - Left: coefficients before Ridge, middle: coefficients after Ridge Regression, and right: showing how lower lambda value results in lower MSE

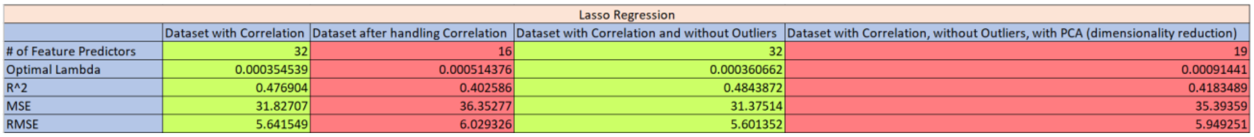
Summary of observations:

- The R^2 for the above models range from 0.39 to 0.42, showing that 39-42% of the target variance can be explained.
- Model 1 and Model 3, using all features, perform better due to lower MSE when compared to Model 2 and Model 4, which have reduced dimensions.
- It was observed from the figure above that ridge regression was successfully able to shrink the coefficients closer to 0.
- Ridge regression depends less on correlations and reduces the variance by shrinking the coefficients hence can give a more robust fit.

4.2 Lasso Regression

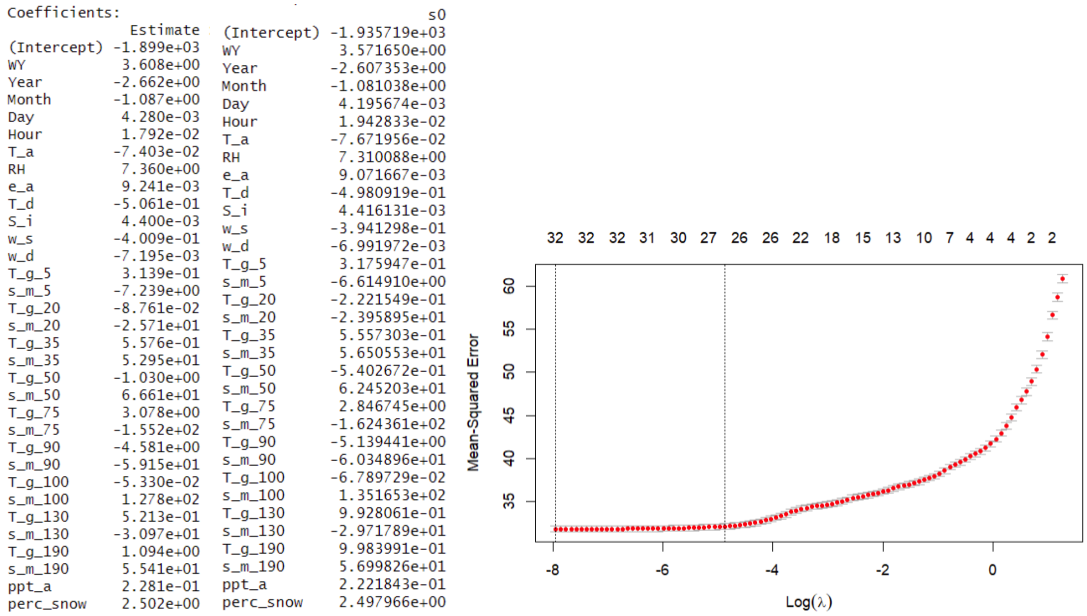
Lasso Regression was performed using the snow depth (z_s) as the response variable while using all other features as predictors with the goal of shrinkage and potential model selection.

- This model is also used when there is multicollinearity present.
- Shrinkage penalty finds optimal lambda to reduce MSE using k-fold cross validation.



Lasso Regression				
	Dataset with Correlation	Dataset after handling Correlation	Dataset with Correlation and without Outliers	Dataset with Correlation, without Outliers, with PCA (dimensionality reduction)
# of Feature Predictors	32	16	32	19
Optimal Lambda	0.000354539	0.000514376	0.000360662	0.00091441
R^2	0.476904	0.402586	0.4843872	0.4183489
MSE	31.82707	36.35277	31.37514	35.39359
RMSE	5.641549	6.029326	5.601352	5.949251

Figure 18 - Showing results obtained from Lasso Regression



```
Coefficients:
               Estimate
(Intercept) -1.899e+03
WY           3.608e+00
Year        -2.662e+00
Month       -1.087e+00
Day          4.280e-03
Hour         1.792e-02
T_a         -7.403e-02
RH           7.360e+00
e_a          9.241e-03
T_d         -5.061e-01
S_i          4.400e-03
w_s         -4.009e-01
w_d         -7.195e-03
T_g_5        3.139e-01
s_m_5       -7.239e+00
T_g_20      -8.761e-02
s_m_20      -2.571e+01
T_g_35       5.576e-01
s_m_35       5.295e+01
T_g_50      -1.030e+00
s_m_50       6.661e+01
T_g_75       3.078e+00
s_m_75      -1.552e+02
T_g_90      -4.581e+00
s_m_90      -5.915e+01
T_g_100     -5.330e-02
s_m_100      1.278e+02
T_g_130      5.213e-01
s_m_130     -3.097e+01
T_g_190      1.094e+00
s_m_190      5.541e+01
ppt_a        2.281e-01
perc_snow    2.502e+00
```

```
                        s0
(Intercept) -1.935719e+03
WY           3.571650e+00
Year        -2.607353e+00
Month       -1.081038e+00
Day          4.195674e-03
Hour         1.942833e-02
T_a         -7.671956e-02
RH           7.310088e+00
e_a          9.071667e-03
T_d         -4.980919e-01
S_i          4.416131e-03
w_s         -3.941298e-01
w_d         -6.991972e-03
T_g_5        3.175947e-01
s_m_5       -6.614910e+00
T_g_20      -2.221549e-01
s_m_20      -2.395895e+01
T_g_35       5.557303e-01
s_m_35       5.650553e+01
T_g_50      -5.402672e-01
s_m_50       6.245203e+01
T_g_75       2.846745e+00
s_m_75      -1.624361e+02
T_g_90      -5.139441e+00
s_m_90      -6.034896e+01
T_g_100     -6.789729e-02
s_m_100      1.351653e+02
T_g_130      9.928061e-01
s_m_130     -2.971789e+01
T_g_190      9.983991e-01
s_m_190      5.699826e+01
ppt_a        2.221843e-01
perc_snow    2.497966e+00
```

Figure 19 - Left: coefficients before Lasso, middle: coefficients after Lasso Regression, and right: showing how lower lambda value results in lower MSE

Summary of observations:

- The R^2 for the above models range from 0.40 to 0.48, showing that 40-48% of the target variance can be explained.
- Model 1 and Model 3, using all features, perform better due to lower MSE when compared to Model 2 and Model 4, which have reduced dimensions.
- From above, it was observed that some shrinkage occurred, however, no features got shrunk to 0 therefore, model selection did not occur.
- This tells us that all the features were influential to snow depth (z_s)

Conclusion

- It was found that any form of dimensionality/feature reduction negatively impacted our results.
- When performing multiple linear regression, almost all the features were highly significant.

What factors affect the snow conditions (snow depth) at Johnston Draw (JD) watershed?

- Features such as relative humidity (RH), Wind Corrected Precipitation (ppt_a), and fraction of precipitation that is snow (perc_snow) are all positively correlated to snow depth (z_s).
- Features such as Air Temperature (T_a), Dew Point Temperature (T_d), and Water Vapor Pressure (e_a) are negatively correlated to snow depth (z_s).
- Ridge regression performed the worst out of all 3 models in all variations of the dataset.
- We were successfully able to study the factors that positively and negatively affected snow depth at Johnston Draw (JD) watershed and were successfully able to make inferences based on our study.

Can we get the inference of snow accumulation, and soil moisture/temperature?

- Features such as Soil Temperatures (T_g) and Soil Moistures (s_m) are not likely to be correlated to snow depth (z_s) since they meet at 90 degree angle.
- One caveat was that since there were stations with missing observations for some features with different time periods, it made it difficult to get a holistic picture and understanding of the data.

We implemented more than what was initially planned in an organized way and implemented additional tasks/processes as our understanding of the project and dataset got better.

Future work

- In the future, we plan to expand our knowledge in this domain and study different meteorological datasets within different regions to get a better understanding of how different features impact snow accumulation within different regions.

References

[1] Godsey, S. E., Marks, D. G., Kormos, P. R., Seyfried, M. S., Enslin, C. L., McNamara, J. P., and Link, T. E.: Eleven years of mountain weather, snow, soil moisture and stream flow data from the rain-snow transition zone – the Johnston Draw catchment, Reynolds Creek Experimental Watershed and Critical Zone Observatory, USA. USDA Ag Data Commons, Idaho, USA, https://doi.org/10.5194/essd-10-1207-2018. 2018.

[2] Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T., and Duan, Q.: Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance, Hydrol. Earth Syst. Sci., 19, 209–223, https://doi.org/10.5194/hess-19-209-2015, 2015.

[3] Wayand, N. E., Massmann, A., Butler, C., Keenan, E., Stimberis, J., and Lundquist, J. D.: A Meteorological and Snow observational data set from Snoqualmie Pass (921 m), Washington Cascades, US, Univ. Washingt. Res. Work. Arch., Seattle, WA, 1–20, https://doi.org/10.6069/H57P8W91, 2015.

[4] Winstral, A., Marks, D., and Gurney, R.: Simulating wind-affected snow accumulations at catchment to basin scales, Adv. Water Resour., 55, 64–79, https://doi.org/10.1016/j.advwatres.2012.08.011, 2013.

[5] https://vitalflux.com/pandas-impute-missing-values-mean-median-mode/

[6] http://www.sthda.com/english/articles/40-regression-analysis/168-multiple-linear-regression-in-r/

[7] https://www.statology.org/ridge-regression-in-r/

[8] https://www.statology.org/lasso-regression-in-r/

[9] https://www.soa.org/digital-publishing-platform/emerging-topics/principal-component-analysis-using-r/

[10] http://www.sthda.com/english/wiki/ggplot2-density-plot-quick-start-guide-r-software-and-data-visualization

Data Sources

Data from: Eleven years of mountain weather, snow, soil moisture and stream flow data from the rain-snow transition zone - the Johnston Draw catchment, Reynolds Creek Experimental Watershed and Critical Zone Observatory, USA. v1.1 | Ag Data Commons (usda.gov)

Contributions

Please see the attached excel sheet link for all tasks assigned and completed:

https://docs.google.com/spreadsheets/d/15V9oflA0iLn_IG6HBPQFmMJ3pcxsx8Yt/edit?usp=sharing&ouid=116682760471926626069&rtpof=true&sd=true