ENVIRONMENTAL FACTORS INFLUENCING DROUGHT IN ASPEN, COLORADO

MICHAEL GRYBKO
UNIVERSITY OF COLORADO, BOULDER

FEBRUARY 27, 2025

ABSTRACT

This report evaluates the impact of environmental factors on drought conditions in Aspen, Colorado, using data from mid-2010 to December 2023. Supervised and unsupervised machine learning techniques were applied to assess feature importance and predict drought severity. A neural network classified five drought levels with approximately 95% accuracy. Principal Component Analysis (PCA) identified air temperature, soil moisture, and soil temperature as the most influential features. PCA improved K-means clustering performance but did not enhance neural network predictions. Given the economic and environmental risks of drought, understanding key drivers and improving predictive models could aid mitigation efforts.

TABLE OF CONTENTS

INTRODUCTION

Drought is an environmental phenomenon when there is not enough water available is to meet demand (Bolinger et al., 2024). Four of the worst droughts to impact Colorado have occurred after 2000 (Colorado Climate Action). Since Colorado is a large state with diverse geography, it can be divided into multiple climate divisions, each which may be differently affected by drought (Bolinger et al., 2024). Aspen sits at almost 8,000 feet above sea level and is in the Central Mountains climate division.

Colorado's economy can be significantly impacted by drought conditions through reduced agricultural production, while also impacting tourism and recreation industries reliant on healthy water sources (United States Department of Agriculture). This can lead to potential job losses and decreased revenue across various businesses. Given its mountainous terrain, Aspen draws many tourists to enjoy activities such as skiing, camping, hunting and fishing (United States Department of Agriculture). Aspen is in a headwaters area, meaning it is the initial source of waters for rivers such as the Roaring Fork (City of Aspen). Therefore, water shortages in the Aspen area could have negative economic fallout locally as well as in downstream communities.

Drought in Colorado can significantly impact the environment by causing decreased plant growth, increased wildfire risk, forest loss due to pest infestations, such as the bark beetle, altered wildlife populations, reduced stream flows, impacting aquatic life, and disrupting natural ecological processes (United States Department of Agriculture). These impacts are particularly pronounced in mountainous regions such as Aspen where snowpack is crucial for water supply throughout the year (Colorado Climate Action).

1. METHODOLOGY

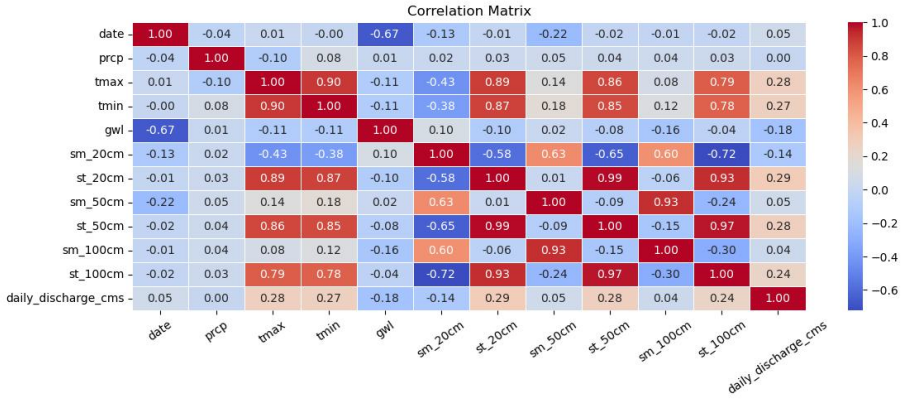
1.1. PREDICTOR DATA IMPORT, CLEANING, AND EXPLORATION

Five datasets containing environmental features were combined to train statistical models related to drought prediction. Weather data was collected from the NOAA website and ground water level and stream flow data was collected from the USGS website, specific URLs are given within the blocks of code and listed in the data references section. The predictors consisted of a weather dataset, a soil moisture and temperature dataset, a groundwater dataset and a streamflow dataset. The target variable was a drought severity dataset. All data were inspected for missing and inconsistent values. Missing values were removed from the datasets. Columns of features that were not necessary were dropped. Individual data features were plotted as time series plots and visually inspected for outliers. Only one outlier was detected in the maximum temperature data and that point was removed. Groundwater and streamflow data were in feet and cubic feet per second respectively. These were converted to centimeters and centimeters per second for consistency as the rest of the data was in metric units. After the data was cleaned the data frames were merged by date creating one dataset, df_aspen. After cleaning all datasets were saved in the Apache Parquet file format to enhance performance and efficiency.

1.2. FEATURE LEGEND

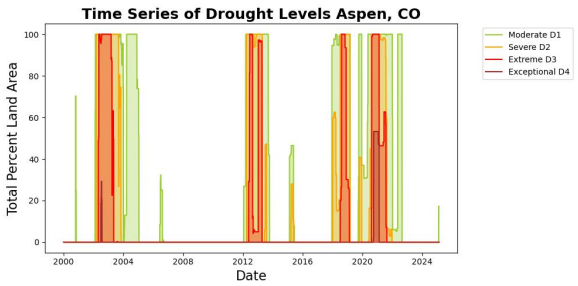
- date: Date of Observations in Format year-month-day from 2010-06-22 to 2023-12-27
- prcp: Total Precipitation in Centimeters
- tmax: Maximum Temperature in Celsius
- tmin: Minimum Temperature in Celsius
- gwl: Level of the Water Table in Centimeters (measures how far the water is from the surface, larger measurements indicating there is less water in the well)
- sm_20cm: Soil Moisture Volumetric Water Content at 20 Centimeters depth
- st_20cm: Soil Temperature at 20 Centimeters depth
- sm_50cm: Soil Moisture Volumetric Water Content at 50 Centimeters depth
- st_50cm: Soil Temperature at 50 Centimeters depth
- sm_100cm: Soil Moisture Volumetric Water Content at 100 Centimeters depth
- st_100cm: Soil Temperature at 100 Centimeters depth
- daily_discharge_cms: Water Flow in Cubic Centimeters per Second

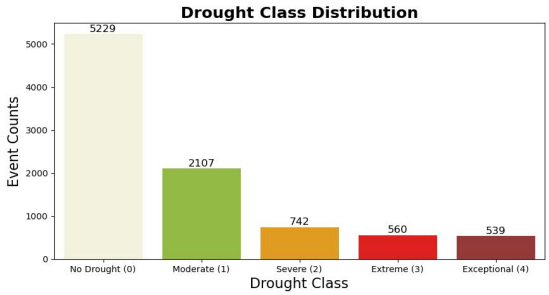
1.3. CORRELATION MATRIX



1.4. DROUGHT DATA IMPORT, CLEANING, AND EXPLORATION

The target drought dataset contains five drought severity levels, None, D1 Moderate, D2 Severe, D3 Extreme, and D4 Exceptional. There is a sixth category, Abnormally Dry (D0), indicating areas that may be going into or are coming out of drought. Each of these categories had its own column. The D0 Abnormally Dry condition was excluded from the classification dataset since it was not a defined drought condition. This dataset was formatted for categorical classification by first using the idxmax function to determine the most severe drought level for each row. Then numerical values were assigned to the drought levels using the map function, creating a new column called drought_level_encoded, which can be used for classification. The original drought levels were dropped leaving the date and drought_level_encoded columns. This dataset was then merged by date with the cleaned df_aspen data frame containing the predictors, creating the df_class_aspen data frame.





2. MODELS

The df_class_aspen data frame was first split by the response variable, drought_level_encoded and the predictors, and the date column was removed. For supervised models the dataset was split into training and test sets (80/20 split), with stratified sampling. To ensure consistency, predictor variables were standardized using StandardScaler. From the correlation matrix there are many correlated features. The soil temperature and moister features are all highly correlated, however none of these features were removed, because differences in soil moisture and temperature at different levels have been shown to be important to droughts at different stages (Xu et al., 2021). Although daily high and low temperatures are also correlated, both were evaluated because they provide important information about the diurnal temperature range.

2.1. NEURAL NETWORK

Neural networks are a supervised machine learning technique. A neural network was built with nine layers, including four dense layers with ReLU activation. Each dense layer was followed by a dropout layer to mitigate overfitting. The output layer contained five neurons, one for each classification category, with a softmax activation function for multi-class classification. This produced a probability vector representing the likelihood of each input belonging to a class (Enslin, 2025). Accuracy and recall scores were recorded to evaluate performance. Various architectures, including different neuron counts and dropout levels, were tested before selecting the final model.

2.2. K-MEANS CLUSTERING

K-means is an unsupervised, iterative clustering algorithm that partitions data into k groups based on similarity. It minimizes the sum of squared Euclidean distances between data points and their cluster centroids. The algorithm iteratively updates centroids as data points are reassigned to the nearest cluster until convergence is reached (Kavlakoglu and Winland, 2024). Inertial and silhouette scores were recoded to determine clustering efficiency. Also, the hyperparameter number of clusters (n_clusters), was optimized using 2, 3, 4, 5, 6 and 10 clusters.

2.3. PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA is a linear dimensionality reduction technique that transforms data into a lower-dimensional space while preserving maximum variance. It achieves this by computing principal components, which are uncorrelated variables derived through singular value decomposition. Here, the number

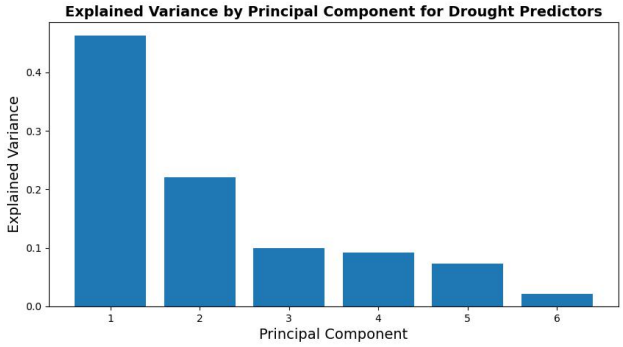
of components (n_components) was optimized to assess the most significant features (Pramoditha, 2025).

3. RESULTS

Classification and clustering results were compared for neural networks and k-means clustering algorithms trained on both the full dataset and a PCA-reduced dataset. PCA guided feature selection helped with unsupervised clustering but did not enhance supervised neural network performance. PCAs impact on clustering was further examined using data visualization.

3.1. PCA

PCA was used for feature selection, leveraging the first two principal components, which together capture approximately 68% of the variance in the dataset. After applying PCA, the absolute values of the variable coefficients were examined to determine feature importance (Singh, 2022). Here, variables with absolute coefficients greater than 0.1 in the first two principal components were selected. Based on this criterion, the most significant features were included, maximum and minimum temperature, soil moisture at 20 cm, 50 cm, and 100 cm depths, soil temperature at 20 cm, 50 cm, and 100 cm depths, and daily river discharge.



Principal Components 1 & 2 Important Features		
Feature	PC1	PC2
daily_discharge_cms	0.148	0.11
sm_100cm	NA	0.602
sm_20cm	0.315	0.366
sm_50cm	NA	0.613
st_100cm	0.428	NA
st_20cm	0.43	0.101
st_50cm	0.435	NA
tmax	0.394	0.199
tmin	0.387	0.226

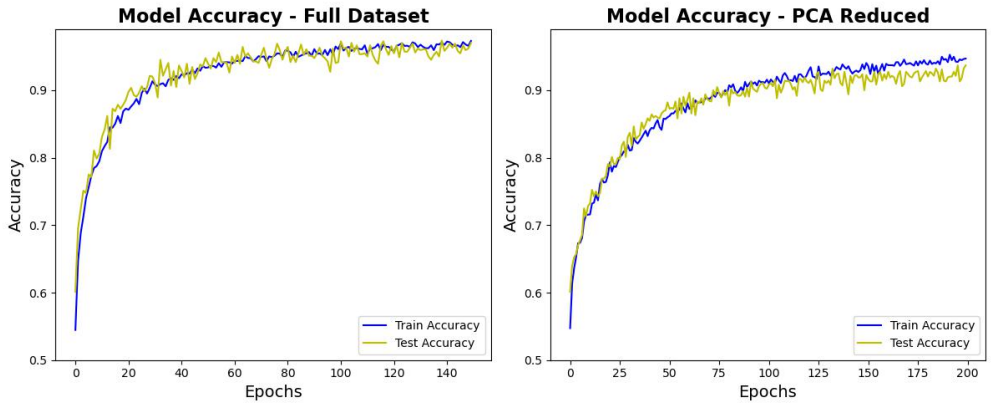
PCA Component Coefficients											
Component	prcp	tmax	tmin	gwl	sm_20cm	sm_50cm	sm_100cm	st_20cm	st_50cm	st_100cm	daily_discharge
PC_1	0.005	0.394	0.387	0.055	0.315	0.071	0.094	0.43	0.435	0.428	0.148
PC_2	0.045	0.199	0.226	0.086	0.366	0.613	0.602	0.101	0.035	0.072	0.11
PC_3	0.241	0.038	0.092	0.794	0.077	0.12	0.03	0.063	0.075	0.089	0.513
PC_4	0.957	0.16	0.027	0.193	0.034	0.046	0.01	0.008	0.005	0.01	0.13
PC_5	0.007	0.047	0.057	0.538	0.08	0.025	0.107	0.031	0.027	0.025	0.827
PC_6	0.027	0.287	0.408	0.096	0.696	0.263	0.406	0.083	0.095	0.081	0.013

3.2. NEURAL NETWORKS

The features prcp and gwl were removed based on PCA results, and the remaining data was processed as previously described. A neural network was then trained on the PCA-reduced training set, with accuracy and recall scores recorded. To improve performance, slight modifications were made to the architecture. The number of neurons in the first dense layer was increased from 16 to 64, while the other dense layers remained unchanged. Dropout levels were slightly increased to mitigate overfitting, which was more pronounced with the reduced dataset. Additionally, the number of training epochs needed to be increased to ensure model convergence.

3.2.1. NEURAL NETWORKS - ACCURACY

The neural network trained on the full dataset outperformed the model trained on the PCA-reduced dataset. The test accuracy for the full dataset model reached approximately 97%, compared to 92% for the PCA-reduced model.

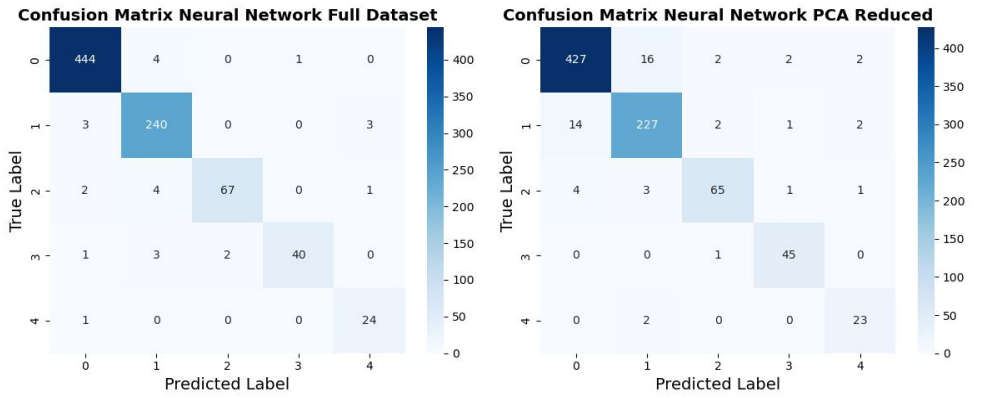


3.2.2. NEURAL NETWORKS - CONFUSION MATRIX & CLASSIFICATION

Since failing to predict a drought, especially an extreme (D4) drought, could have severe environmental and economic consequences, recall was closely examined. The model trained on the full dataset achieved a higher overall recall than the PCA-reduced model (97% vs. 93%). Notably, for the most severe drought level (D4), the full dataset model demonstrated a higher recall (96% vs. 92%).

Neural Network Classification Report - Full Dataset				
Drought Class	Precision	Recall	F1-score	Support
None (0)	0.984	0.989	0.987	449
Moderate (1)	0.956	0.976	0.966	246
Severe (2)	0.971	0.905	0.937	74
Extreme (3)	0.976	0.87	0.92	46
Exceptional (4)	0.857	0.96	0.906	25
Summary Metrics				
Accuracy	0.97	0.97	0.97	0.97
Macro Avg	0.949	0.94	0.943	840
Weighted Avg	0.971	0.97	0.97	840

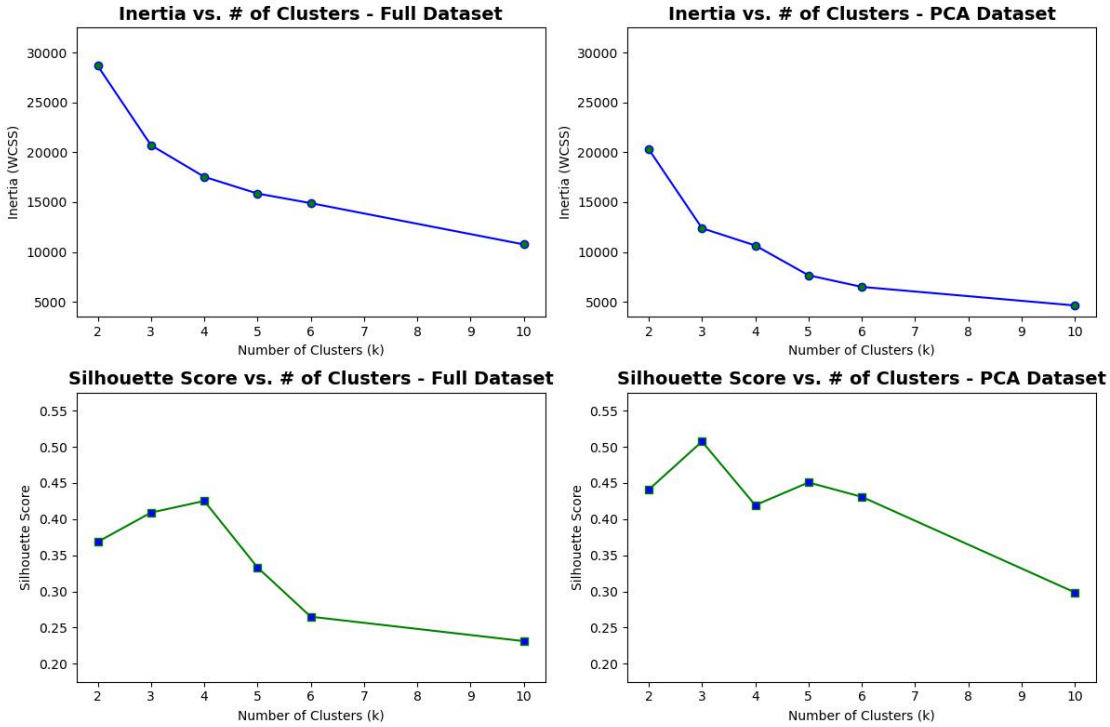
Neural Network Classification Report - Reduced Dataset				
Drought Class	Precision	Recall	F1-score	Support
None (0)	0.96	0.951	0.955	449
Moderate (1)	0.915	0.923	0.919	246
Severe (2)	0.929	0.878	0.903	74
Extreme (3)	0.918	0.978	0.947	46
Exceptional (4)	0.821	0.92	0.868	25
Summary Metrics				
Accuracy	0.937	0.937	0.937	0.937
Macro Avg	0.909	0.93	0.918	840
Weighted Avg	0.938	0.937	0.937	840



3.3. K-MEANS CLUSTERING

For clustering evaluation, the inertia score was used to measure how tightly data points were clustered. This metric quantifies the sum of squared distances between each point and its assigned cluster centroid. Lower inertia scores indicate better-defined clusters. The elbow method was applied to determine the optimal number of clusters, identifying the "elbow point" where inertia begins to decrease more gradually. Additionally, the silhouette score was used to assess cluster separation, ranging from -1 to 1, with higher scores indicating better-defined clusters (Singh, 2022).

Based on these metrics both datasets could be partitioned optimally into four or five clusters. Comparing the inertia and silhouette plots for both datasets, the PCA-reduced dataset exhibited lower inertia scores and higher silhouette scores than the full dataset. This suggests that PCA improved clustering performance by enhancing cluster separation and reducing within-cluster variance.



3.4. DATA VISUALIZATION WITH K-MEANS CLUSTERING AND PCA

PCA can be a powerful tool for data visualization when using two or three principal components (Pramoditha, 2025; Singh, 2022). To further explore how PCA influenced the clustering efficiency of the K-means algorithm, both the full and PCA-reduced datasets were analyzed using K-means clustering with five clusters and two principal components. The clustered data points were visualized on a scatter plot, with individual points color-coded by cluster assignment. Examining the scatter plots, the PCA-reduced dataset exhibits more compact and well-separated clusters, indicating that PCA effectively reduced noise and improved clustering performance.

4. LIMITATIONS

The dataset used in this study only had a record of total precipitation which was not subdivided into snow and rain totals. It has been reported that snowpack is crucial for water supply in Colorado's mountain regions such as Aspen (Colorado Climate Action). Therefore, it may be possible to create more accurate models if data regarding snow totals and snowpack measurements were included. Also, it was surprising that there are so few soil moisture measurement locations throughout the state, and there was no soil moisture location in Aspen. The closest soil moisture location to Aspen was in Montrose Colorado, which is about 70 miles away. Given the importance of this type of data to drought prediction (Xu et al., 2021), and the integral role Aspen has as a water source statewide and beyond (Bolinger et al., 2024), it seems prudent to measure soil moisture in Colorado's mountain regions.

5. CONCLUSION

Drought severity is classified into four stages: D1, with early signs of water shortages and voluntary restrictions; D2, where agricultural losses increase and mandatory restrictions are imposed; D3, with major agricultural losses and widespread water shortages; and D4, the most severe, leading to extreme water shortages and emergencies (NOAA's National Weather Service). Colorado is considered a headwaters state because several major river systems have headwaters within its borders (Bolinger et al., 2024). Mountain regions such as Aspen supply much of the water for these river systems, because 60 to 80% of their annual streamflow comes from snowpack (Bolinger et al., 2024). Given the significant environmental and economic impacts of drought, identifying key contributing factors and developing predictive models is crucial

This study assembled a dataset with variables related to soil moisture and temperature, streamflow, groundwater level, precipitation, and air temperature to analyze drought conditions in Aspen, CO. A neural network achieved over 97% accuracy in classifying drought severity. Since minimizing false negatives, misclassifying drought conditions as non-drought is critical, recall was a primary focus. The model achieved a 97% total recall and 96% recall for the most extreme (D4) drought level, demonstrating strong predictive performance.

Applying PCA improved K-means clustering but did not enhance neural network performance. This result aligns with expectations. The full dataset contained correlated features known to be important for drought prediction (Xu et al., 2021). K-means clustering, which relies on Euclidean distance, struggles with highly correlated variables, leading to poorly defined clusters. PCA, by removing linear dependencies, allowed K-means to capture clearer patterns. However, neural networks, with their ability to model non-linear relationships, retained crucial predictive power even with correlated features. This suggests that while PCA can enhance clustering, reducing the dataset may remove valuable information for accurate drought prediction.

6. REFERENCES

6.2. TEXT REFERENCES

Bolinger, R.A., J.J. Lukas, R.S. Schumacher, and P.E. Goble, 2024: Climate Change in Colorado, 3rd edition. Colorado State University, https://doi.org/10.25675/10217/237323.

City of Aspen. *Water Conservation | Aspen, CO*. www.aspen.gov/592/Water-Conservation.

Colorado Climate Action. *Climate Change in Colorado: Public Health and Environmental Impacts | Climate*. climate.colorado.gov/health-and-environmental-impacts.

Enslin, Shaun. "The Complete Guide to Neural Networks Multinomial Classification." *Towards Data Science*, 21 Jan. 2025, https://towardsdatascience.com/the-complete-guide-to-neural-networks-multinomial-classification-4fe88bde7839/

Kavlakoglu, Eda, and Vanna Winland. "K-Means Clustering." *IBM Think*, 19 Dec. 2024, www.ibm.com/think/topics/k-means-clustering.

NOAA's National Weather Service. *Colorado Drought*. www.weather.gov/bou/co_drought.

Pramoditha, Rukshan. "How to Select the Best Number of Principal Components for the Dataset." *Towards Data Science*, 28 Jan. 2025, https://towardsdatascience.com/how-to-select-the-best-number-of-principal-components-for-the-dataset-287e64b14c6d/.

Saunders, Stephen, et al. "Climate Change in the Headwaters Water and Snow Impacts." *A Report to the Northwest Colorado Council of Governments*, by Northwest Colorado

Council of Governments, 2018, nwccog.org/wp-content/uploads/2018/02/Climate-Change-in-the-Headwaters.pdf.

Singh, Shivangi. “K Means Clustering on High Dimensional Data. - the Startup - Medium.” *Medium*, 10 Apr. 2022, https://medium.com/swlh/k-means-clustering-on-high-dimensional-data-d2151e1a4240

United States Department of Agriculture. *Drought Impacts in the Rocky Mountain Region*. Sept. 2017, www.fs.usda.gov/sites/default/files/r2-droughtfactsheet.pdf.

Xu, Zheng-Guang, et al. “Comparison of Soil Moisture at Different Depths for Drought Monitoring Based on Improved Soil Moisture Anomaly Percentage Index.” *Water Science and Engineering*, vol. 14, no. 3, Aug. 2021, pp. 171–83. https://doi.org/10.1016/j.wse.2021.08.008.

6.3. DATA REFERENCES

Weather data, precipitation, max temperature and min temperature, Aspen Pitkin Co Airport Sardy Field, CO US: https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00093073/detail

Daily discharge data for Lincoln Creek, Near Aspen, CO: https://dashboard.waterdata.usgs.gov/api/gwis/2.1/service/site?agencyCode=USGS&siteNumber=09073005&open=211313

Ground water level data, drought well near Aspen, CO: https://dashboard.waterdata.usgs.gov/api/gwis/2.1/service/site?agencyCode=USGS&siteNumber=395136108210004&open=212313

Soil moisture and temperature data Montrose, CO:

https://www.ncei.noaa.gov/data/us-climate-reference-network/access/derived-products/soil/soilanom/

Drought data, Pitkin County, CO:

https://www.drought.gov/states/colorado/county/pitkin