Snowcast Showdown - Our Entry into the Driven Data Competition

Team No.1: Naive Unsupervised Children AKA "Snowdown Showdown"

Competition Link
Github

Team Members:

Benjamin Hendel: Preprocessing data files, forecast modeling and tuning

Keren Portugais: EDA and preprocessing

Haim Shragai: Spatial interpolation stage, API

Aviv Tahar: Data preparation, forecast baseline modeling

Introduction

The Bureau of Reclamation is a federal agency responsible for water resources management across western USA. Specifically it applies to the oversight and operation of the diversion, delivery and storage projects that it has built throughout western United States for irrigation, water supply, and attendant hydroelectric power generation.

Published by the data competitions and challenges platform Data Driven, the bureau of reclamation has sponsored the following challenge to predict Snow Water Equivalent (SWE) across western United States.

The goal of this challenge is to estimate SWE at a high spatiotemporal resolution over the Western U.S. using near real-time data sources. A total prize of $500,000 (Further information available on the Data Driven data science competition platform under the name **'Snowcast Showdown**', details of the development stage competition are available at the 'Development Stage' link) will be awarded to teams that most accurately estimate the SWE during 2020-21 for ~9,000 cells (1x1 km) scattered over western US.

Getting better SWE estimates for mountain watersheds and headwater catchments is intended to help water resources management professionals to improve runoff and water supply forecasts, which in turn will help reservoir operators manage limited water supplies and respond to extreme weather events such as floods and droughts. For contest submission we created a model to predict snow with minimum error.

Dataset Description & EDA

Several resources are available and approved by the data provider for a range of spatial grid cells and measuring stations with weekly SWE data of the snow seasons over the years 2013-19. The goal of this project is to forecast values for the snow seasons over the years 2020-21 on a weekly basis for locations defined by the competition, which are surrounding existing given data locations scattered around the US.

The bundled dataset includes almost complete data for 700 snow monitoring stations (SNOTEL/CDEC), that are spread but are not geographically identical to the competition's target cells. Figure 1 below shows an example of such a SNOTEL station.

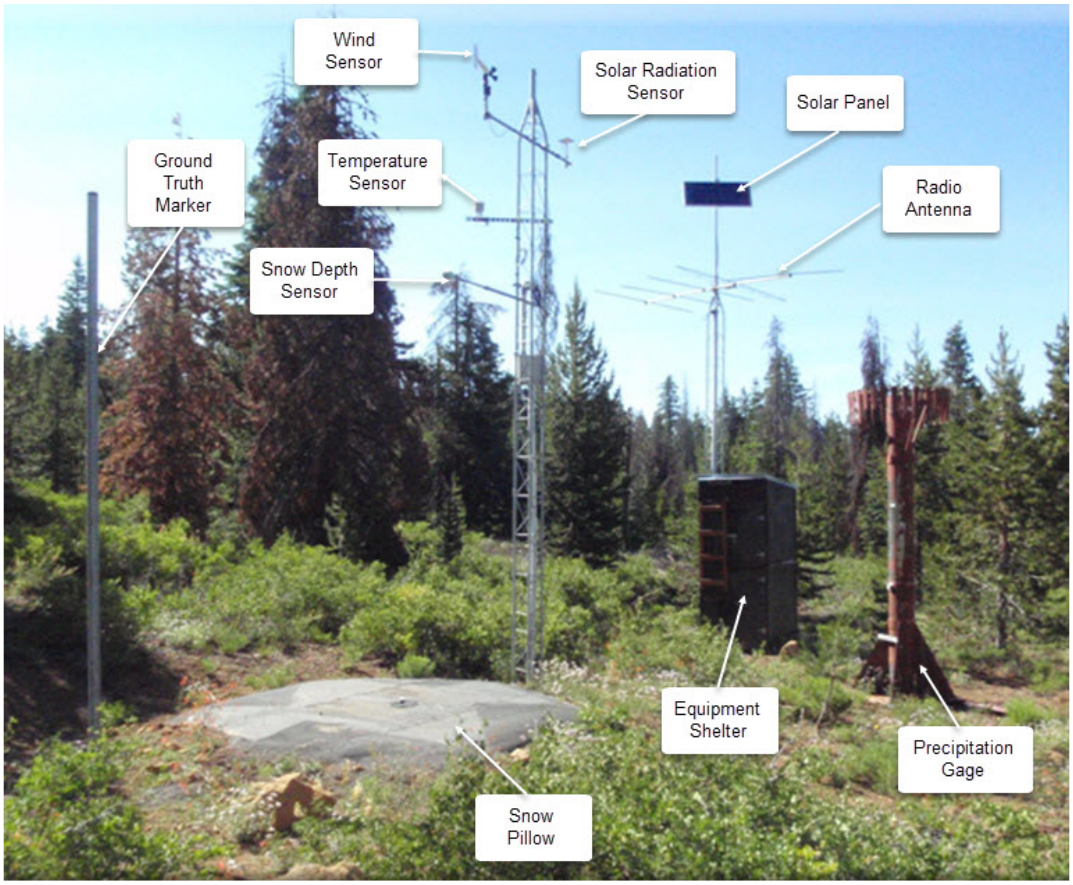


Figure 1: SNOTEL station and its various parts.

The SNOTEL and CDEC stations are spread widely across the western US.
These stations are not necessarily placed where forecasts should be. This amplifies the complexity of this problem, which requires handling both time and spatial components.

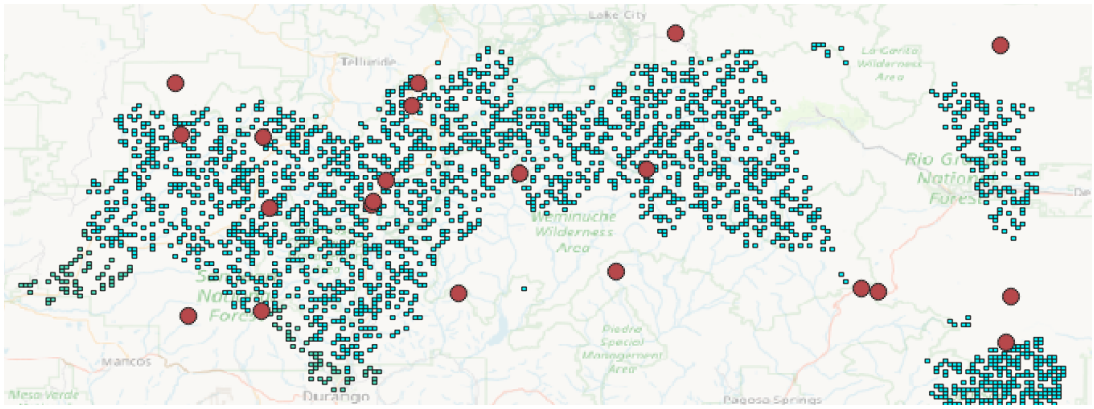


Figure 2: A close up of some 1x1 km target cells in cyan, SNOTEL and CDEC stations in red.

Rather than including SWE readings during the whole year, the data includes only weekly SWE readings over the snow season, which spreads from December to July. The data is scattered through space and time components. Including approximately 4 percent of null values.

Two examples of station training data are shown in figure 3 below. Note that the left plot shows a complete time series, while the right plot shows a time series with missing values which are interpolable.

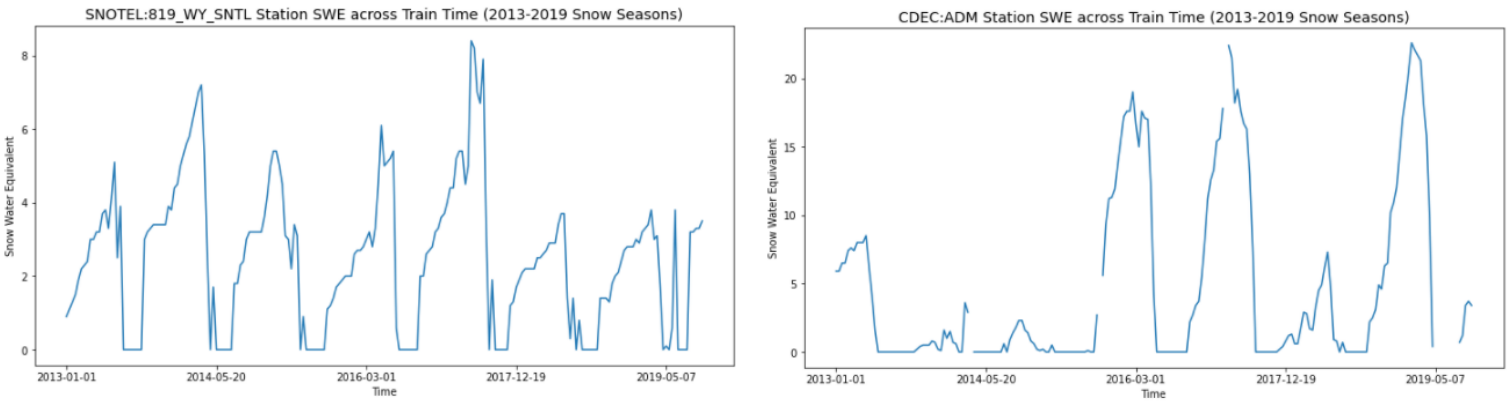


Figure 3: “CDEC:ADM” and “SNOTEL:819_WY_SNTL” time series of SWE during 2013-2019 (snow seasons only).

To deal with null values (~4%) we used linear interpolation. Some stations’ time series were mostly missing, meaning interpolation would be insufficient to estimate the data accurately (extrapolation in this case), so we removed those stations. In addition, we found stations’ time series whose beginning is missing, which led to interpolation failing to recognize entire seasons, as seen in figure 4 below. Those stations were also removed. 45 stations were removed, decreasing the number of stations from 700 to 655.

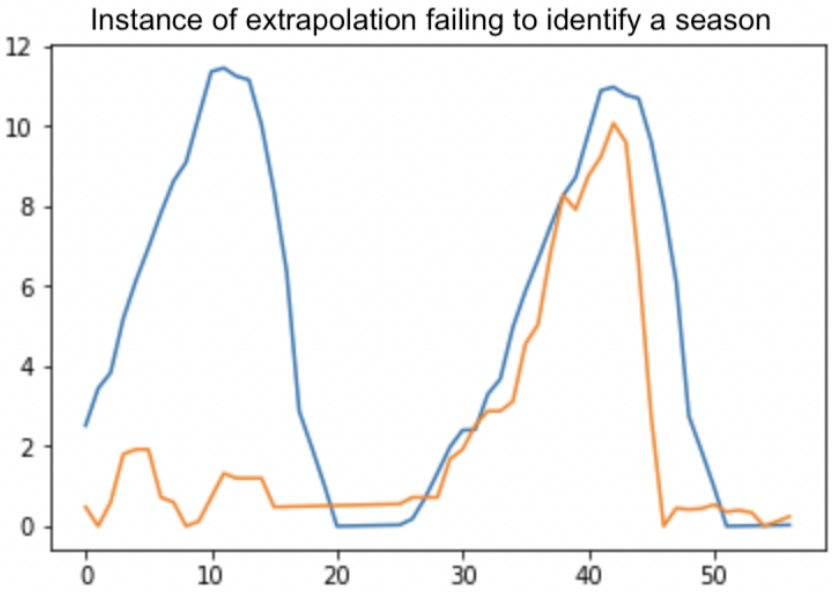


Figure 4: an example of extrapolation failing to accurately describe missing data. Orange line is extrapolated SWE, blue line is true SWE.

These dataset files are ‘ground_measures_train_features.csv’ (SWE over 2013-19) which we use for training our forecasting models and ‘ground_measures_test_features.csv’ (SWE over 2020-21) which are used to test our predictions.

Some other files that we compiled for spatial orientation consist of location ID, longitude and latitude.
“ground_measures_metadata.csv” for the stations in train and test set and “target_metadata.csv” for target cells.

Solution overview

To solve this problem we chose to use the following structure:

1. **Forecasting stage:**
   The first stage forecasts the future SWE of all stations in the train dataset. For each station, using SWE of snow seasons over the years 2013-19, we forecast the values of SWE of snow seasons over the years 2020-21.
   To find the best method for this stage we split the data into train and evaluation sets, forecast SWE for each station separately and evaluate overall performance of all stations.

2. **Spatial interpolation stage:**
   The second stage uses the forecasts of the previous stage as training data and interpolates SWE in target cells surrounding the stations in the train set.
   To find the best method for this stage we split the data into train and evaluation sets, interpolate SWE for each date separately and evaluate overall performance over all dates of snow seasons of the years 2020-21.

Since predicting SWE (forecast-wise and spatial-wise) is a regression problem, we chose RMSE as our metric for evaluating performance for both stages separately and for the entire architecture as a whole.

Forecasting stage

The goal of this stage is to predict SWE for snow seasons during 2020-21 at the 655 preprocessed stations based on previous seasons (2013-19).

We used random walk (predicting the last value for the entire target period) and seasonal (repeating last season) as our baseline models to evaluate the performance of our more sophisticated models, which were SARIMAX and Exponential smoothing.
An example of one station forecast with all stated models can be seen below in figure 5.

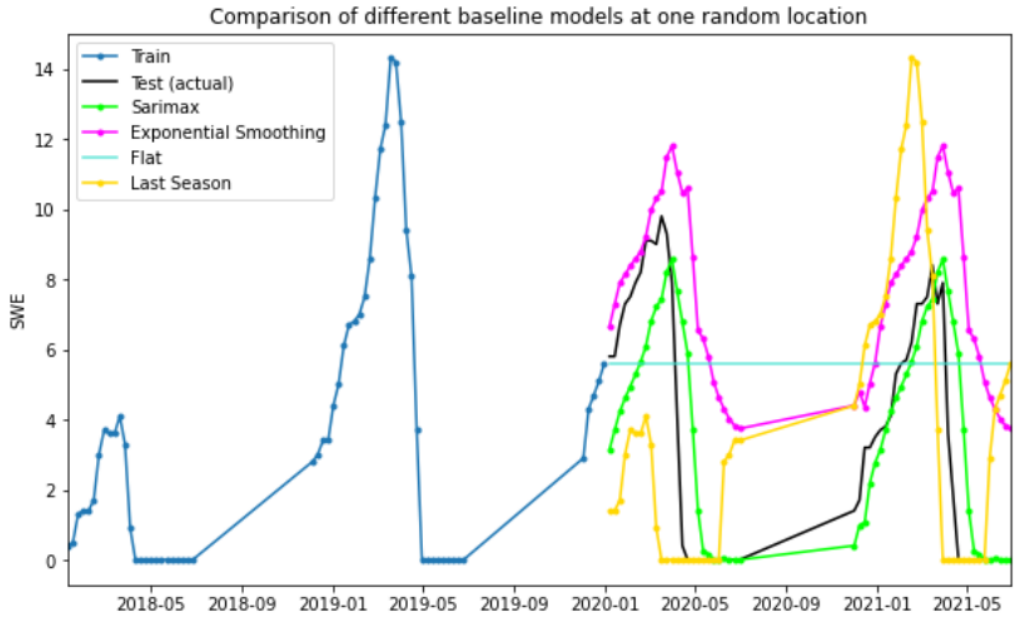


Figure 5: time series forecast comparison.

Performance report of all stated forecasting models, sorted by RMSE:

MODEL	RMSE
Flat/Random Walk	9.59
Previous Season	8.7
Exponential Smoothing	7.26
SARIMAX (0,0,0)(2,0,1,31)	6.13

The random walk model predictably performed the worst, and the previous season model performed slightly better, but both did not perform well enough as 2019 was an unusually stormy year while 2020 was relatively less stormy, and both baselines did not have access to data other than the snow season of 2019. The simple Exponential Smoothing model captured the seasonality component well and showed a great improvement overall.

Our best forecast model is SARIMAX with hyperparameters (0,0,0) (2, 0, 1, 31).
The hyperparameters were tuned with a combination of intuition and randomized search, as a complete randomized search would have been very expensive given having seven hyperparameters, and a different model for each station (655). Overall the seasonality was ~31 weeks in a snow season, and no obvious trend.

The SARIMAX model outperformed the Exponential Smoothing model overall stations in general (and specifically for the time series in figure 5 above).

It is important to note that sometimes, models other than SARIMAX got better forecasts (based on RMSE score), so we tried to build some mixed model, which selects the best forecasting model (out of the models stated above) for each station using a validation set. The SARIMAX model was usually selected since it performed the best, though occasionally it was outperformed by baselines in anomalous locations.
A pie chart of the distribution of the models selected by this method is shown below in figure 6.

The predictions of this model were proven to be worse while evaluating the output of the whole architecture, possibly due to an unusual cycle during the validation set.

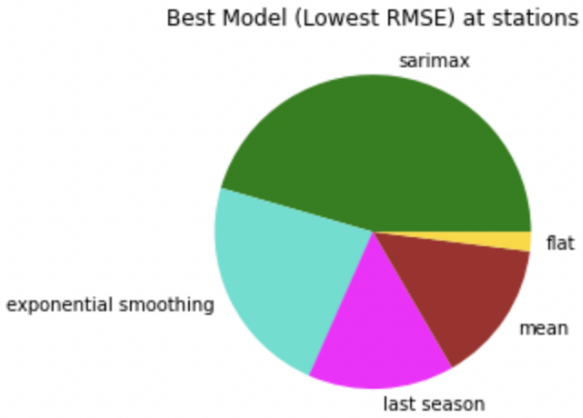


Figure 6: distribution of best forecasting models for all 655 stations.

Spatial interpolation stage

The goal of this stage is to interpolate SWE in ~9000 target cells for snow seasons during 2020-21 using the output of the previous stage as training data.

We used Radial Basis Function (RBF) and Gaussian Process Regression (Kriging) for interpolation. Both RBF and Kriging have an important function that defines how to perform the interpolation, Radial Function for RBF and Variogram for Kriging.

We applied those methods over the whole time frame, interpolating target cells based on all stations forecasted SWE for that day. Such interpolation is shown in figure 7 below.

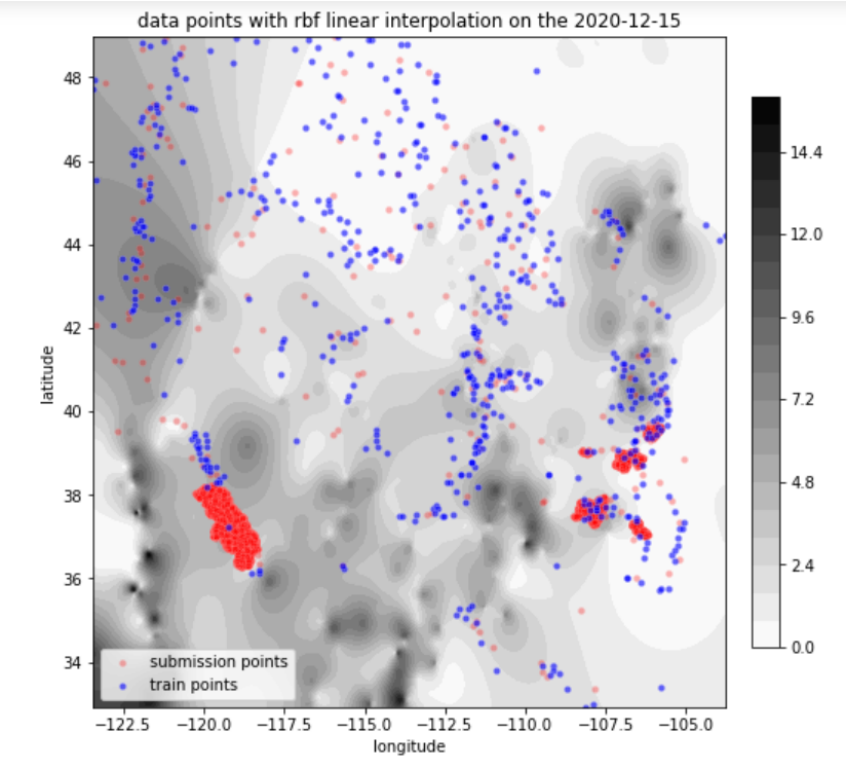


Figure 7: interpolated SWE colored map based on station forecasts (blue) on 2020-12-15. Target cells are in red.

Performance report of interpolation functions, sorted by RMSE:

Method	RMSE
Linear Kriging	7.75
Spherical Kriging	7.78
Exponential Kriging	7.97
Linear RBF	9.21
Thin Plate RBF	18.15

Final performance

Performances of both stages, forecasting stage and spatial interpolation stage, are calculated using an internal validation set derived from the training data.

As for the performance of the whole architecture, it is calculated via a submission system provided by the competition organizers. Our final RMSE score is 11.81. A link to our powerpoint presentation is here.

It is worth mentioning that after submitting output of the spatial stage with test data (2020-21 SWE) as input instead of forecasts from the forecast stage we got an RMSE score of 8.9 which ranked our team at 58 out of 993 competitors at the time.

API

Our API provides two services:

- Plot_station
- Plot_date

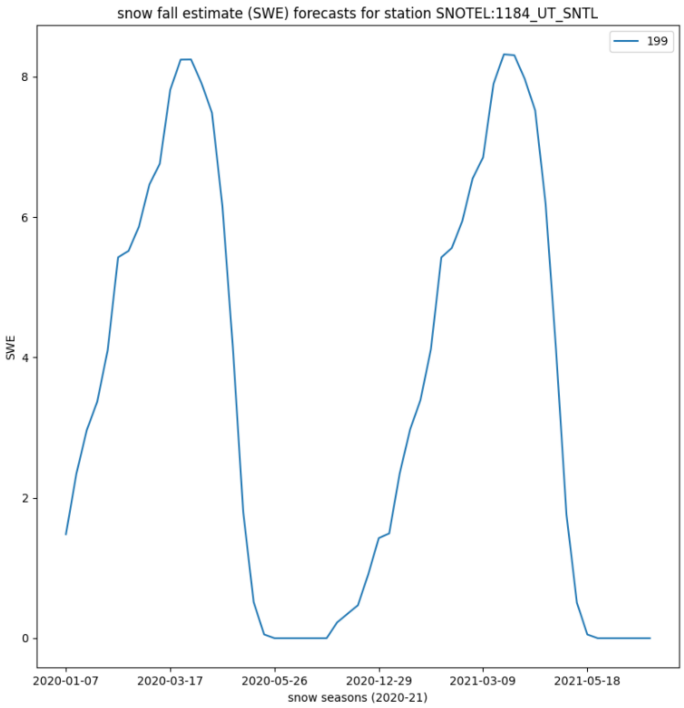
Plot_station expects an argument “station” and is plotting forecasts of 2020-21 snow seasons for the given station.

Plot_date expects an argument “date” and is plotting an interpolated colored map of SWE in western US on that given date.

Both services accept “random” as station/date value to output some random station or date data, since knowing names of specific stations or even the dates of the weekly snow season cycle is a little niche.

With some additional work this API can be used to get local snow season estimates for any location based on a URL call.

Examples of the API output are shown in figure 8 and 9 below.



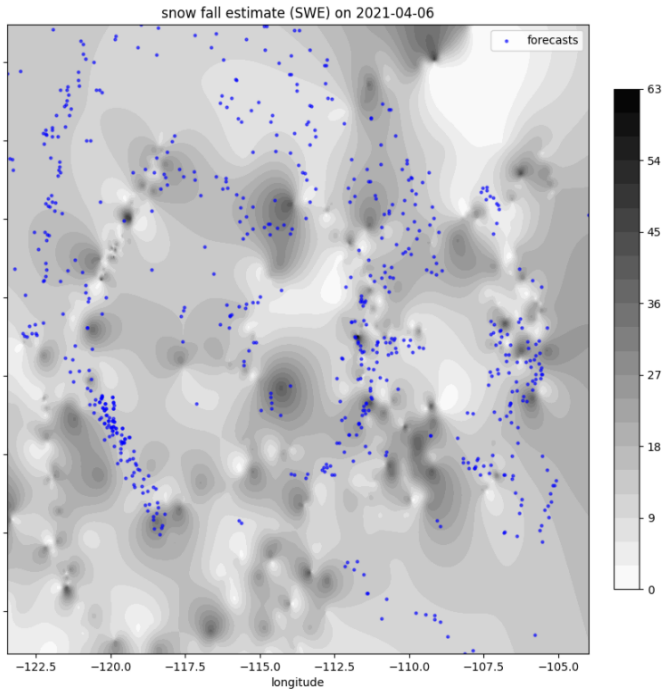


Figure 8: time series for SWE forecasts of a specific station
Figure 9: interpolated colored map of SWE on a specific date

Future Steps

At this stage, our model would benefit from more data. Using only SNOTEL/CDEC stations SWE is not enough to accurately crack this problem, time and spatial components both. There are a lot of other sources of data we can integrate into our models, such as:

Elevation: A high resolution DEM, since snowfall is associated with higher elevations

Temperature: Averages and weekly readings, since snowfall is associated with lower temperatures

Satellite imagery: LANDSAT and clear hi res imagery to detect white color

Reflectance: Remote Sensing data from HRRR (https://microsoft.github.io/AIforEarthDataSets/data/noaa-hrrr.html), snow reflects light and can be estimated using an index

Land cover: Detailed land cover types for terrains such as mountain