Combined physically-based and machine learning approach for operational estimation of snow water equivalent across the Western U.S.: Snowcast Showdown competition

Evgeniy Malygin[1], Ekaterina Rets[2*], Maxim Kharlamov[3], Ivan Malygin[4], Maria Sakirkina[5]

[1] PJSC MegaFon, Department of Big Data Analysis and Machine Learning, Moscow, 127006, Russia
[2] Institute of Geophysics, Polish Academy of Sciences, Warszawa, 01-452, Poland
[3] Water Problems Institute, Russian Academy of Sciences, Moscow, 119333, Russia
[4] Schmidt Institute of Physics of the Earth of Russian Academy of Sciences, Moscow, 123242, Russia
[5] Higher School of Economics National Research University, Faculty of Geography and Geoinformation Technology, Moscow, 109028, Russia

Task statement

In this study we aimed to combine physically-based and machine learning approaches to operational SWE prediction at 1km resolution across the entire Western U.S. using near real-time data sources. The solution was developed for the **Snowcast Showdown** Drivendata.org competition hosted by Bureau of Reclamation. The data sources approved in the competition set up were used in the study for features generation. The inference time of the model was limited to 8 hours. The model with frozen model weights was to be run to each week to generate and submit near real-time predictions throughout the winter and spring season 2022. Predictions were evaluated using RMSE metric against ground-based and lidar airborne measurements as they become available.

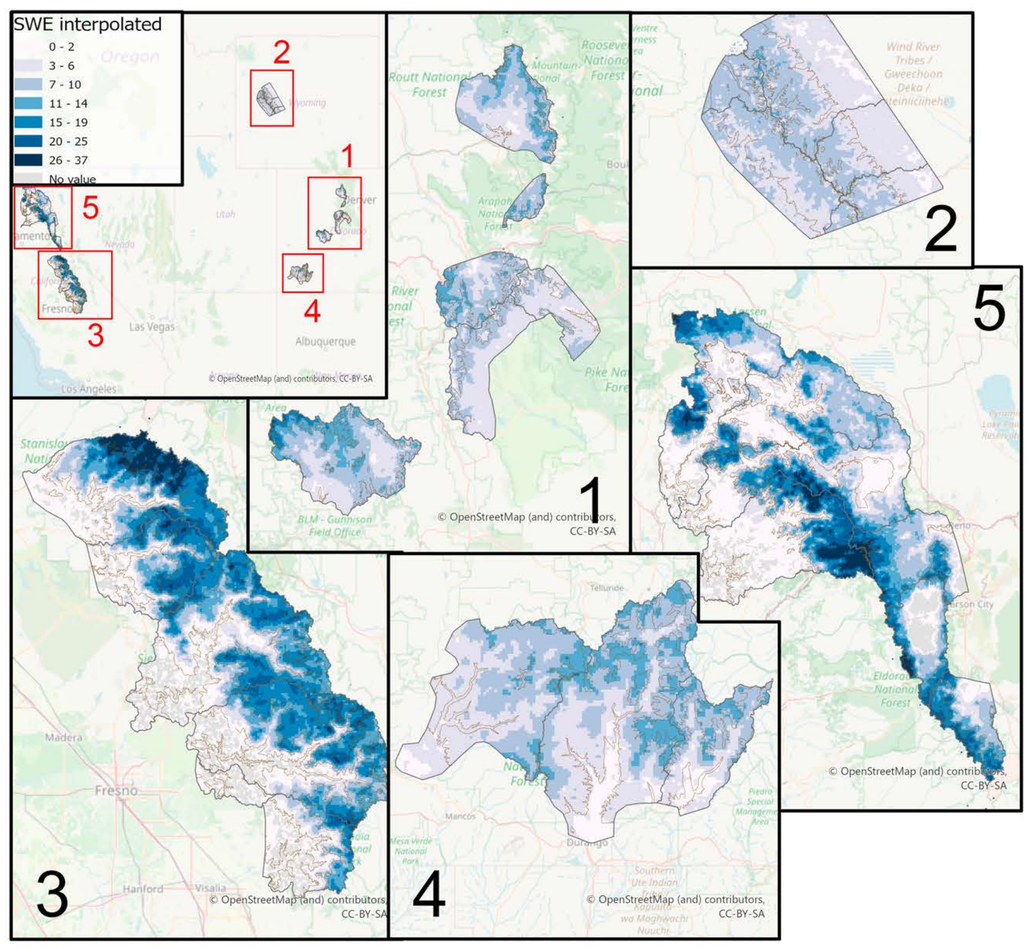
Study area and Data

The study area is represented by several scattered sections covering 32 063 km² on the western slope of the Sierra Nevada with grid cells elevation range from 12 to 4309 m and 17 858 km² in the Rocky Mountains with elevation range from 1800 to 4200 m.

The target dataset consisted of the SWE values given for each of the 11 000 grid cells for 2013-2019 and 9 000 grid cells for 2020-2021 with irregular time step (from 1 to 250 days) derived from a combination of ground-based SNOTEL, CDEC sites and ASO LiDAR measurements of the NSIDC.

Following data sources approved in the competition set up were used to develop **feature set** for the solution:
•SNOTEL
•CDEC
•MODIS Terra MOD10A1 satellite imagery product (Snow Cover Daily L3 Global 500m SIN Grid) .
•The High-Resolution Rapid Refresh (HRRR) NOAA real-time 3-km hourly updated atmospheric model.
•Copernicus DEM (90 m resolution)

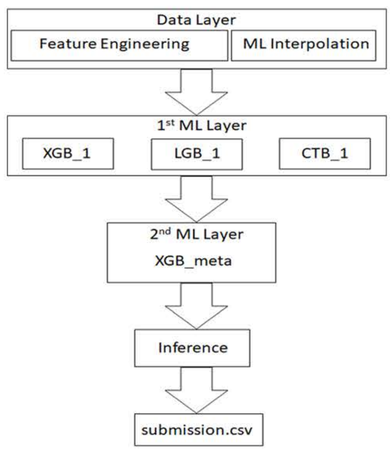
Example of the SWE modelling (in inches of w.e.) on 01.03.2021



Model describtion

The model consists of several layers. The first one includes data preparation and ML Interpolation process. Next two layers are the main stacking model. At the first ML layer we use different SOTA implementations of Gradient Boosting Machine algorithm: XGBoost, LightGBM and CatBoost. Next ML layer is a meta-model which is based on predictions from the previous layer. Meta regressor is another instance of XGBoost.
The final meta-model builds predictions based on the responses of the first-level models.

The output of the model is the predicted SWE values for the given day in a format fitting the contest submission file requirements.



Feature engineering

The designed feature set of 121 features includes the parameters directly characterizing *snow cover properties*:

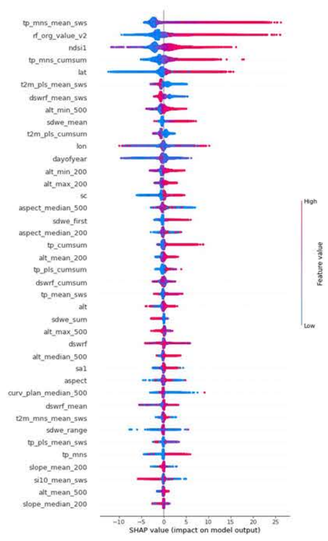
• MODIS parameters (NDSI),
• SNOTEL and CDEC,
• SWE characteristic from HRRR,

as well as a set of features serving as predictors of physical processes underlying snow cover evolution:

1. predictors of snow accumulation
• precipitation sum characteristics
• sum of solid precipitation
• negative temperature sums

2. *predictors of snow redistribution:*
• different characteristics of altitude, slope, aspect, curvature, topographic ruggedness index

3. *predictors of snow ablation:*
• downward radiation
• relief parameters characterizing spatial distribution of solar direct and diffuse radiation in mountainous areas (slope, aspect, elevation)
• snow albedo (MODIS data)
• air temperature, wind speed (characteristics of turbulent heat fluxes)
• liquid precipitation

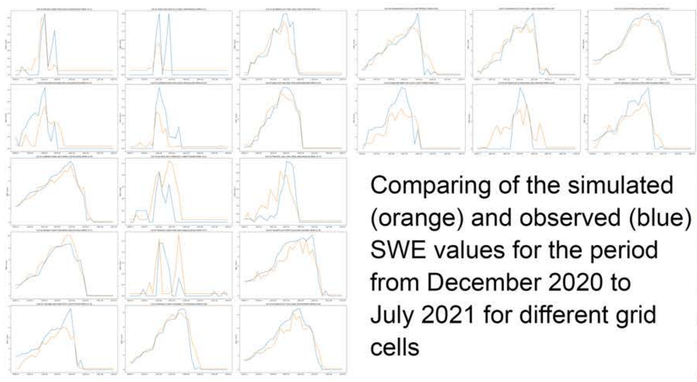


All meteorological parameters were aggregated for 3 time periods: 1 day, 7 days and winter period (from December of the previous year to the day of the forecast)

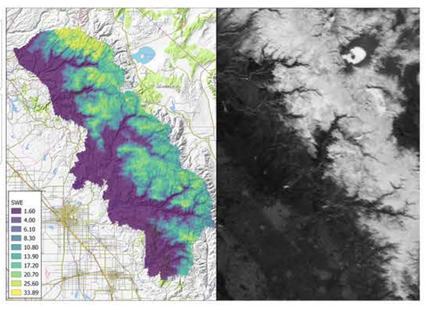
The stability and consistency of the top of the feature importance structure for all three Gradient Boosting models and from one experiment set up to another indicates potential interpretability of the developed solution.
The top features include first of all spatial-temporal interpolation based on SNOTEL and CDEC, ndsi1 and sc (MODIS Terra MOD10A1). First 1-4 places in all cases include as well physically-based indirect predictors of SWE: first of all seasonal cumulative sum, average solid precipitation. Among features characterizing snow ablation the most important are characteristics of air temperature and solar radiation. In all cases cumulative sums and long-term averages of meteorological characteristics give a more important signal than daily values, that reflects the inertness of SWE dynamics.
Among terrain parameters features characterizing spatial differences in incoming solar radiation, especially the characteristics of the surface aspect, display substantial level of importance.
Higher importance of latitude and longitude compared to altitude characteristics outlines prevalence of spatial differences in geographic and climatic conditions over the study area compared to altitude effects.

Results

The developed model is an end-to-end pipeline that allows you to predict SWE for the desired date and grid cell over the western slope of Sierra-Nevada and Rocky Mountains. There are the following steps in the pipeline:
1. Downloading and processing the meteorological data for the previous week and winter period
2. Downloading and processing of MODIS space images released during the week
3. Data collection from updating (meteorological data, space images) and stable (relief, etc.) resources
4. Loading model weights from the pretrained model
5. Making predictions and preparing a weekly submission to the contest.

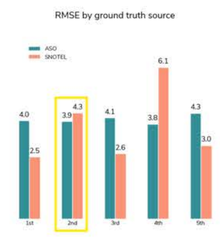


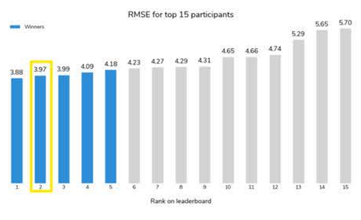
Comparing of the simulated (orange) and observed (blue) SWE values for the period from December 2020 to July 2021 for different grid cells



A comparison of the simulated SWE values and snow coverage according to MODIS satellite imagery data

Analysis of the results and model's metrics shows that it has a fine spatiotemporal generalizing ability. The solution scored 2d among 1063 participants with RMSE= 3.97 inch of w.e. and R^2= 0.69





Malygin, E., Rets, E., Kharlamov, M., Malygin, I., Sakirkina, M. Application of combined physically-based and machine learning approaches for operational estimation of snow water equivalent across the Western U.S. Water Resources Research *(Under review)*