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

Estimation of the spatial distribution of snow water equivalent (SWE) in mountainous areas is currently one of the most important unsolved problems in snow hydrology. A combined physically-based and machine learning approaches based on different SOTA implementations of Gradient Boosting Machine algorithm was developed to predict SWE at 1 km resolution at over the Western U.S. The solution is using a wide range near real-time data sources (in-situ, remote sensing data, results of general circulation modeling) as well as terrain parameters as predictors. The approach was developed in course of the Snowcast Showdown Drivendata.org competition hosted by Bureau of Reclamation. The dataset based on ground measured SNOTEL, CDEC and airborne ASO LiDAR data used as ground truth for model training and validation was provided by the contest organizers. A dataset of 121 features was created, describing the physics of the process, as well as natural-climatic features of the region under study. The top features in terms of importance include ground snow measure data (SNOTEL, CDEC), and remote sensing of snow cover (MODIS Terra MOD10A1). The top 1-4 features include as well physically-based indirect predictors of SWE: seasonal cumulative sum and average values of solid precipitation, seasonal cumulative sum and average values of air temperature and the mean seasonal value of solar radiation. The robustness of the model was checked by validation and hidden sampling. The optimized RMSE metric values amounted to 3.97 inch SWE on the public leaderboard of the competition. The solution scored 2d among 333 participants (https://www.drivendata.org/competitions/90/competition-reclamation-snow-water-eval/leaderboard/). An end-to-end solution and automated real-time forecast pipeline was developed to reproduce the forecast for each week. Hence, the developed solution is universal and can be used both for operational monitoring and for forecasting in advance.